

Comparative IndoBERT, Logistic Regression, SVM, Random Forest, and Naïve Bayes for TikTok Deepfake Misinformation

Tri Pidiyanto*, Iwan Setiawan Wibisono

Fakultas Komputer dan Pendidikan, Program Studi Teknik Informatika, Universitas Ngudi Waluyo, Semarang, Indonesia

Email: ^{1,*}trippiidiantoo@gmail.com, ²iwansetiawan@unw.ac.id

Correspondence Author Email: triippiidiantoo@gmail.com

Received: 19/07/2026, Revised: 21/08/2026, Accepted: 04/09/2026, Published: 08/09/2026

Abstract—Generative AI has made deepfake content, especially face-swap videos and cloned voices, easier to produce and spread, and short-video platforms such as TikTok have become a major channel for this material. Most existing work still focuses on detecting the manipulated audio or video itself, leaving audience reactions in the comment section relatively unexplored. The problem this study addresses is that, without reading these comments, moderation systems have no way to know whether audiences are being misled by AI-generated content; the objective is to determine whether a fine-tuned IndoBERT model can flag such misinformation-leaning comments more reliably than conventional machine-learning classifiers. This paper takes a comparative approach to that gap, testing whether a fine-tuned IndoBERT model can flag potential misinformation in Indonesian-language TikTok comments more reliably than four TF-IDF-based classifiers: Logistic Regression, Random Forest, Multinomial Naive Bayes, and Linear SVM. Starting from 1,345 comments scraped across 15 TikTok videos, a multi-stage cleaning process left 1,071 usable comments, manually sorted into three labels: Misinformation, Skeptical, and Other. IndoBERT was fine-tuned with a weighted cross-entropy loss to offset class imbalance and evaluated through a train-validation-test split and Stratified 5-Fold Cross Validation. The fine-tuned model reached 82.41% accuracy and 81.58% F1-Macro, beating every machine-learning baseline by at least 13.84 points in F1-Macro, with five-fold results holding steady at an average F1-Macro of 80.17%. These results suggest IndoBERT is a stronger option than conventional machine learning for flagging misinformation-leaning comments on AI-generated content, offering a practical foundation for automated moderation on social platforms.

Keywords: IndoBERT; Machine Learning; Misinformation Detection; TikTok; Deepfake Comments

1. INTRODUCTION

Artificial intelligence has changed not just how digital content gets made, but how it moves through the information ecosystem in the first place. It is no longer only a productivity tool; algorithms now decide, largely on their own, what reaches which users and when. That shift has produced a social media environment that is fast, personalized, and, unfortunately, easy to exploit [1]. One of the clearest examples of this is the deepfake: media manipulated with deep learning to the point that images, audio, or video look and sound authentic even though none of it actually happened [2].

Two techniques dominate how deepfakes get made in practice: face-swapping and voice cloning. Face-swapping replaces a person's facial identity in a photo or video using AI models, while voice cloning synthesizes speech that sounds almost indistinguishable from a real person's voice. Facial manipulation alone spans several subtypes synthesis, attribute editing, identity swapping, and expression transfer [3] and detecting them typically relies on deep architectures such as CNNs or Vision Transformers, though detection keeps getting harder as generation quality improves and as the computational cost of running these detectors climbs [4]. Voice cloning raises a parallel set of concerns, since it can be repurposed for disinformation, fraud, or copyright violations, which is exactly why building reliable fake-audio detectors has become an active research problem in its own right [5], with recent work on self-supervised speech models showing promising gains in this direction [6].

TikTok's scale more than 1.6 billion active users makes it a particularly attractive vector for shaping public opinion, and not always for benign reasons [7]. What makes TikTok interesting from a research standpoint is that AI-manipulated content does not travel through the video and audio alone; it travels through the comment section too, and analyzing that text is a useful lens for spotting misinformation [8]. BERT-style language models are well suited to this because they read text with genuine contextual awareness rather than treating words as isolated tokens [9], and prior work has repeatedly found that this contextual understanding outperforms traditional machine learning on social media text classification tasks [10], including for Indonesian-language misinformation detection specifically [11].

Detection research in this space has been moving from single-modality methods toward multimodal ones that combine visual, audio, and textual signals [12]. Visual pipelines typically target facial manipulation, audio pipelines target synthetic voices, and NLP methods such as BERT are used to make sense of the surrounding text. The results so far are encouraging, but most studies still isolate one manipulation type or one data source at a time [13].

That last point is where the gap in the literature sits: there is not yet much work on systems that can recognize AI-manipulated content such as face-swaps and voice clones while also reading the surrounding conversation for context. Comments, when they appear in existing studies at all, tend to be used only to gauge engagement rather than as a direct signal feeding into the detection model [14]. Audience-focused studies have begun to close part of this gap from the other direction: topic modelling and sentiment analysis have been used to map public reaction to deepfake-related posts and comments [15], and controlled experiments show that the comments attached to a deepfake actively shape how viewers judge its authenticity rather than merely reflecting engagement [16], neither line of work, however, performs automated misinformation classification of the comments themselves. Because most research still treats

visual manipulation, audio manipulation, and text analysis as separate problems, the picture painted does not really match how information actually spreads on a platform like TikTok, even though BERT-based classifiers have already shown strong results for rumor and malicious-comment detection on other platforms [17], and hybrid BERT architectures have pushed misinformation-detection accuracy even further in other social contexts [18].

Prior research can be grouped into at least four relevant directions. First, visual deepfake studies have developed CNN- and Transformer-based detectors for manipulated faces [4]. Second, voice-deepfake studies have explored self-supervised speech models for synthetic-audio detection [6]. Third, audience-oriented research has used topic modelling and sentiment analysis to examine public perceptions of deepfakes [15]. Fourth, misinformation studies have applied BERT-based or hybrid BERT architectures to social-media text classification [11], [17], [18]. Although these studies establish strong foundations, they generally examine media manipulation, public perception, or misinformation text as separate problems; in particular, limited attention has been given to classifying Indonesian TikTok comments themselves according to whether they reinforce misinformation, express skepticism or clarification, or fall outside the research topic.

This study responds to that gap by treating user comments, rather than the audio or video itself, as the primary signal for spotting potential misinformation tied to AI-generated content specifically content showing signs of voice-cloning or face-swapping on TikTok. Instead of trying to detect manipulation in the media directly, the analysis focuses on how users respond to it in writing. BERT is used here precisely because it captures semantic relationships and context well, and it is tested directly against conventional machine learning classifiers to see which approach actually performs better on this task. The broader aim is to contribute something useful toward content-moderation systems for a media ecosystem that is increasingly shaped by generative AI.

Four contributions come out of this work: (1) an annotated dataset of Indonesian-language TikTok comments, sorted into misinformation, skepticism/clarification, and other categories, tied to deepfake content, (2) a head-to-head comparison between a fine-tuned IndoBERT model and four conventional machine learning baselines Logistic Regression, Random Forest, Multinomial Naive Bayes, and Linear SVM on this classification task, (3) a Stratified 5-Fold Cross Validation of IndoBERT's performance, run to make sure the reported results are not an artifact of a single lucky data split, and (4) an early-stage analytical framework for gauging how much misinformation risk a video carries based on its comment patterns and engagement levels. Taken together, these contributions advance scientific knowledge by showing that comment-level language modeling, rather than media-level manipulation detection alone, can serve as a viable signal for misinformation moderation on AI-generated video content, extending BERT-based misinformation detection research into the previously underexplored context of Indonesian-language, deepfake-adjacent social media discourse.

2. RESEARCH METHODOLOGY

2.1 Research Stages

The research follows a fairly linear path, starting with comment collection and ending with model evaluation and comparison. The main stages are data cleaning, exploratory data analysis (EDA), manual labeling, label encoding, splitting the dataset into training, validation, and test sets, fine-tuning IndoBERT alongside training the machine learning baselines, evaluating both with a confusion matrix and standard classification metrics, and validating the results with 5-Fold Cross-Validation. The overall flow is shown in Figure 1.



Figure 1. Research Stages

2.2 Data Acquisition

Data Source Comments were pulled from Indonesian TikTok users on videos suspected of containing manipulated content, deepfakes, face-swaps, or voice clones. Fifteen videos were purposively chosen based on viral reach,

comment activity, and hashtags such as #aiinfluencer, #aitrend, #aiviral, #deepfake, or #aikonten, as well as altered faces, bodies, or public-figure voices, such as fake song covers. Content Selection each candidate video was manually checked before comments were pulled to confirm it genuinely showed signs of deepfake or generative-AI manipulation. Comments had to be in Indonesian, public (not from private accounts), and tied to face-swap or voice-clone content to make the cut. Data Collection Procedure comments were extracted with two Chrome extensions, TTGET-TikTok Comment Exporter and Comment Exporter For TikTok, then exported to CSV and merged into one dataset. Nothing in the comment text was altered during extraction, so the data still reflects real user interaction on the platform. Extracted Data Each record carried the comment text, comment language, timestamp, reply count, Digg (like) count, video URL, and video ID. Anything tied to user identity was stripped out after the analysis stage and was never used in modeling only briefly, during a data-quality check. In total, 1,345 comments were collected. Research Ethics Only publicly available TikTok data was used, with no user intervention and no redistribution of content, which keeps this an observational study on public data. All identifying information was anonymized, and the dataset was used strictly for academic purposes in line with standard ethical principles for AI research and digital data protection.

2.3 Data Preprocessing

Attribute Selection Irrelevant attributes were filtered out, keeping only those needed for the data-processing stage. Case Folding Every letter in each comment was converted to lowercase to keep the text format consistent for processing. Data Cleaning Unnecessary text was removed at this stage, including non-Indonesian comments, emoji-only comments, comments with unnatural character lengths, duplicates, missing values, URLs, hashtags, mentions, special characters, irrelevant numbers, and repeated symbols carrying no linguistic meaning. Text Normalization Slang, abbreviations, and colloquial TikTok language were normalized using an Indonesian slang dictionary built from the Colloquial Indonesian Lexicon, extended with platform-specific terms found during data review. Emojis were converted into text so their emotional meaning would not be lost. This kind of lexical normalization follows standard practice for informal social media text [19]. Tokenization Tokenization followed the IndoBERT Base (WordPiece) approach, adding a [CLS] token at the start of each sentence and a [SEP] token at the end, with a maximum sequence length of 128 tokens. Text Representation Tokenized text was converted into numeric input suitable for the Transformer model. Stemming and stopword removal were deliberately skipped, since BERT relies on complete sentence structure to interpret linguistic context.

2.4 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was carried out to understand the dataset's characteristics before labeling and model training, covering comment-language distribution, comment length, number of comments per video, word-count distribution, emoji distribution, and the most frequent words and bigrams. The EDA results show a dataset of 1,071 comments spanning 15 TikTok videos, with an average comment length of 37.49 characters and 6.54 words. Word-frequency analysis surfaced terms such as *AI*, *yg*, *nya*, *ya*, and *gw* as the most common, while bigram analysis picked up phrases like *jackie chan*, *ya allah*, *terima kasih*, *mr bert*, and *rafi ahmad*. All of which reflect how users typically react to deepfake content. EDA was also used to examine Digg Count as a measure of user interaction, revealing a heavily skewed distribution in which only a small share of comments attracted a disproportionately high number of likes. This was used solely to describe the data's characteristics and was not used as a feature in the classification model.

2.5 Data Labeling

MIS (Misinformation) Class this class covers comments that take AI-generated content at face value, believing or reinforcing it as though it were genuine information. SKP (Skeptical/Clarification) Class this class covers comments expressing doubt, seeking clarification, or otherwise showing awareness that the content results from AI manipulation. OTH Class (Other): This class covers comments unrelated to the research focus, including spam, jokes, and other off-topic remarks. Annotation Process this labeling was carried out by the researchers following annotation guidelines adapted from prior studies to keep labels consistent, and the supervising lecturer validated a portion of the annotated data before it was used for model training.

2.6 Modeling Using IndoBERT and Machine Learning Baselines

Model Selection IndoBERT Base (indobenchmark/indobert-base-p1) was chosen as the primary classification model, given that it was trained on the Indo4B corpus over three billion words of Indonesian text giving it a strong semantic representation [20], [21]. Four conventional TF-IDF-based classifiers served as comparison baselines: Logistic Regression, Random Forest, Multinomial Naive Bayes, and Linear SVM, all of which have been widely used in recent Indonesian text classification and sentiment analysis research. These four algorithms were chosen because they represent complementary families of sparse, TF-IDF-based classification linear (Logistic Regression, Linear SVM), probabilistic (Multinomial Naive Bayes), and ensemble tree-based (Random Forest) so that IndoBERT's contextual representation is benchmarked against the range of approaches most commonly reported in prior BERT-versus-traditional-ML studies on social media text [9], [10]. This design follows the same rationale used in earlier evaluations of contextual language models against lexical baselines for misinformation-style classification tasks [11]. Model

Architecture the IndoBERT architecture consists of an input layer, a Transformer encoder, a [CLS]-token pooling layer, a fully connected layer, and a Softmax activation function that produces classification probabilities [22]. The machine learning baselines, by contrast, rely on TF-IDF vectorization as their input representation. Hyperparameter Optimization were tuned through a series of experiments in Google Colab, testing different combinations of learning rate, batch size, dropout rate, and weight decay. The best combination was chosen based on Macro F1-Score on the validation set, aiming for a configuration that delivered optimal classification performance. Model Fine-Tuning full fine-tuning was performed in Google Colab on a Tesla T4 GPU, updating all parameters of the IndoBERT Base model on labeled training data for up to five epochs, with early stopping guided by validation F1-Macro to prevent overfitting. Handling Data Imbalance because the class distribution was not perfectly balanced, IndoBERT was trained with Weighted Cross-Entropy Loss so that the minority class carried a heavier error weight than the majority class.

2.7 Data Sharing and Validation

Data Distribution the dataset was split using Stratified 5-Fold Cross Validation to keep the proportion of each class (MIS, SKP, OTH) consistent across folds, an approach chosen to keep the training and testing process balanced and the evaluation results more representative. Training and Test Data in each validation round, four folds about 80% of the data served as training data while the remaining fold about 20% served as test data. This was repeated across five iterations so that every data point was used as test data exactly once. Model Validation performance was calculated at each validation iteration and then averaged into a final score, reducing the bias that a single data split can introduce and producing a more stable, reliable evaluation.

2.8 Model Evaluation

Accuracy is used to measure the proportion of predictions that are correctly classified against all test data. Precision captures how accurately the model identifies comments belonging to each class in other words, how many of its positive predictions were correct. Recall is used to measure the model's ability to find all comments that actually belong to a class. F1-Score balances Precision and Recall to give a more complete picture of classification performance. Macro F1-Score served as the primary performance indicator in this study because it weighs each class equally regardless of imbalance, making it a fairer way to judge how well the model identifies MIS, SKP, and OTH comments, and it was also the main basis for comparing IndoBERT against the machine learning baselines. Confusion Matrix a confusion matrix was used to examine the model's misclassification patterns for each class, showing the number of correct and incorrect predictions and making it easier to pinpoint where the model struggles to distinguish misinformation, skepticism, and other comments.

3. RESULT AND DISCUSSION

3.1 Data Pre-processing

Cleaning the dataset was a necessary step before it went into analysis and model training. The initial dataset from the comment-crawling process held 1,345 comments from 15 deepfake-themed TikTok videos. From there, several cleaning steps followed: removing unused attributes, missing values, duplicate comments, spam comments, promotional comments, sticker-only comments, overly short comments, and comments in languages other than Indonesian. Table 1 summarizes the progressive data-cleaning stages, showing how the initial crawling result was reduced to the final research dataset. The dataset was progressively reduced to 1,071 comments that satisfied the research criteria. This reduction should be viewed as a process of enhancing data quality rather than information loss, as it results in a dataset that is more representative for model training. This step is essential because Transformer models depend on accurate semantic representations, and the inclusion of spam, stickers, or irrelevant comments would introduce noise and negatively affect classification performance.

Table 1. Summary of Data Cleaning Stages

Name	Number
Crawling result dataset	1345
Deleting missing values	1319
Deleting spam comments	1310
Remove account promotion/tag	1309
Deleting sticker comments	1187
Deleting very short comments	1172
Language selection	1093
Final verification and removal of duplicates	1071

3.2 Exploratory Data Analysis (EDA) Results

Dataset Statistics: A quick statistical pass was run over the comments used in the study, which confirmed a final dataset of 1,071 comments from 15 TikTok videos. Table 2 summarizes the main descriptive statistics of the final

dataset used for classification. The comments in this dataset are characteristically brief, averaging approximately six words, which aligns with established patterns of social media communication that prioritize concise exchanges. These statistics were calculated based on the 1,093 comments remaining after language filtering, prior to the final duplicate removal. An additional 22 duplicates were subsequently identified and excluded, resulting in the final dataset of 1,071 comments as presented in Table 1. This final adjustment represented only 2.01% of the 1,093 comments and did not substantially alter the overall statistical characteristics of the dataset.

Table 2. Dataset Statistics

Name	Number
Number of comments	1071
Number of videos	15
Average comment length	37.49 characters
Median comment length	28 characters
Average number of words	6.54 words
Average Digg Count	248.39
Median Digg Count	3

Language Distribution Indonesian dominated the dataset. A handful of comments were flagged as Malay or unrecognized by the language detector, usually because of heavy abbreviation, emoji use, slang, or mixing several languages in one sentence; non-Indonesian comments were removed to keep the dataset consistent, as shown in Figure 2.

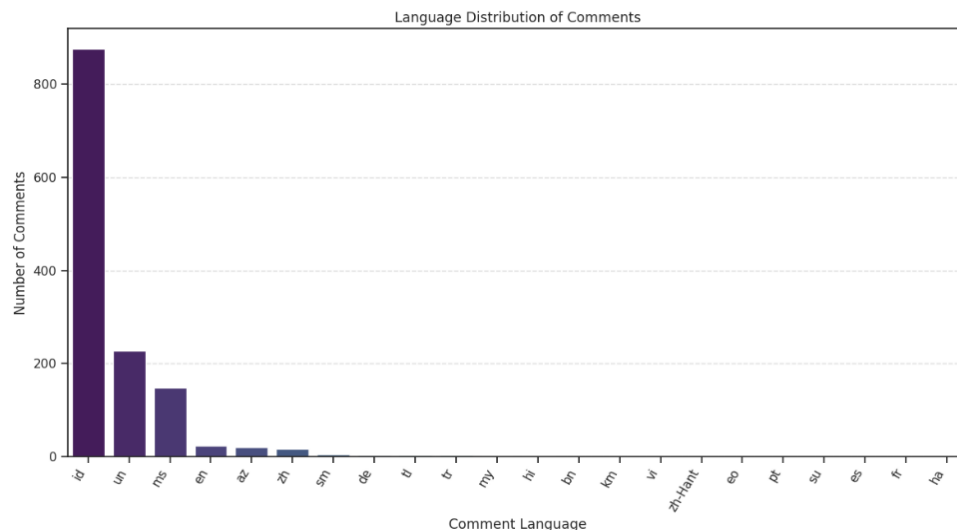


Figure 2. Distribution of comment languages

Comment Length Most comments run under 50 characters, suggesting TikTok users tend to respond briefly to deepfake content, as Figure 3 shows.

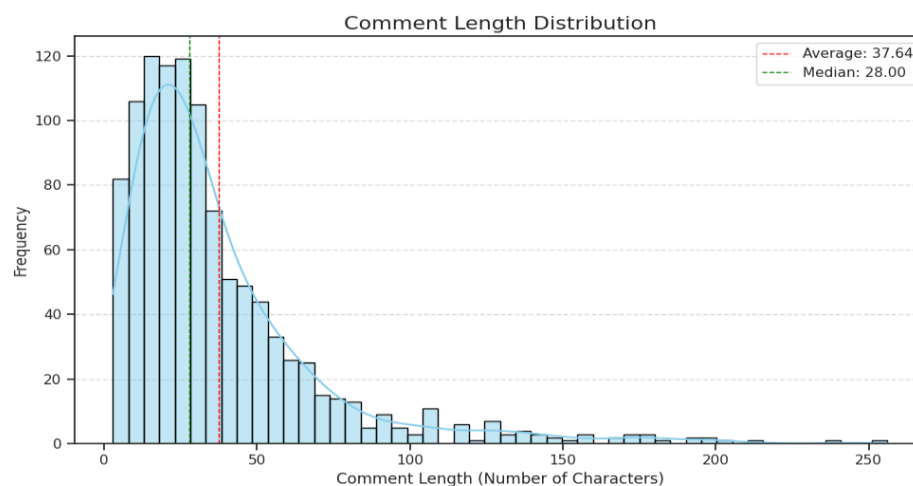


Figure 3. Distribution of comment lengths

Word Count Distribution Nearly two-thirds of comments top out at six words, which means the classification model has to extract meaning from very little textual information, as illustrated in Figure 4.

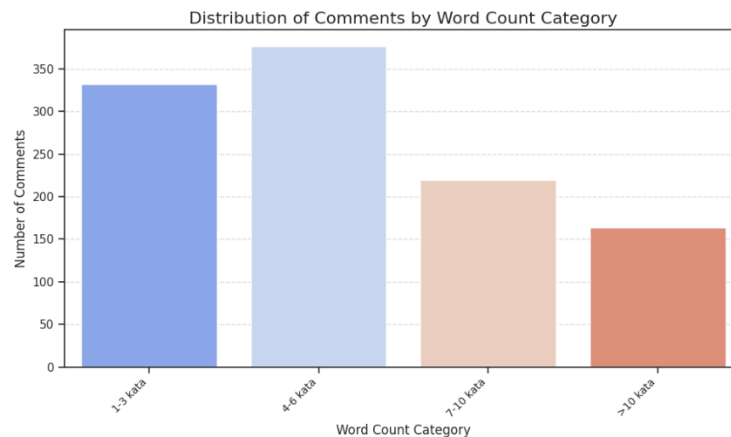


Figure 4. Distribution of comments by word count category

Digg Count Distribution Engagement is highly uneven the median Digg Count is just 3, while the maximum reaches 106,141, and an Interquartile Range check flagged 151 outliers. In short, a small slice of comments pulls in outsized engagement while most get little response, as shown in Figure 5.

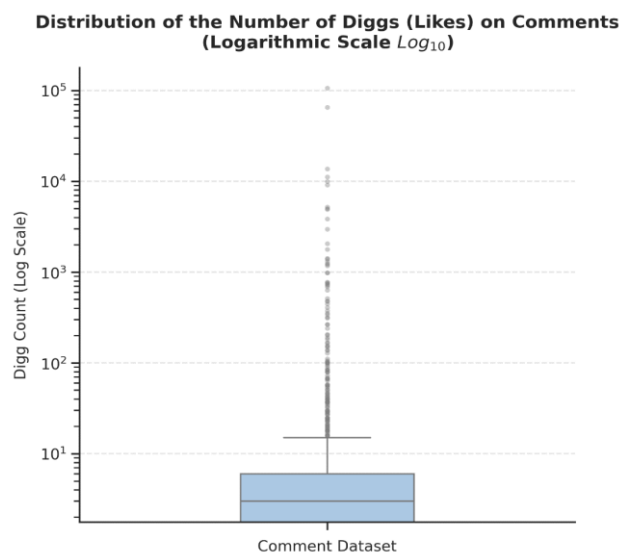


Figure 5. Digg distribution of comments

Comment Distribution Video Comment counts per video are fairly balanced, generally sitting around 90–100 comments each, which lowers the risk of any single video dominating the dataset, as shown in Figure 6.

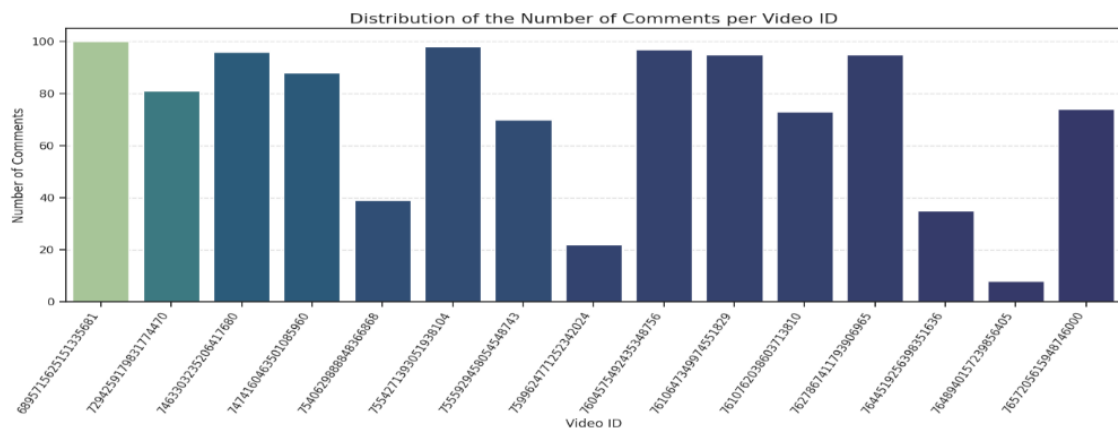


Figure 6. Distribution of the number of comments per video

Word Analysis the most frequently occurring words were "AI," "yg," "nya," "ya," and "gw." These terms indicate three primary themes within the dataset: references to artificial intelligence, responses indicating deception by deepfake content, and general emotional reactions to the videos. The prevalence of "AI" demonstrates that viewers often explicitly associated the content with artificial intelligence, while the occurrence of "ketipu" (deceived) suggests a persistent perception that deepfake content is capable of misleading audiences, as illustrated in Figure 7.

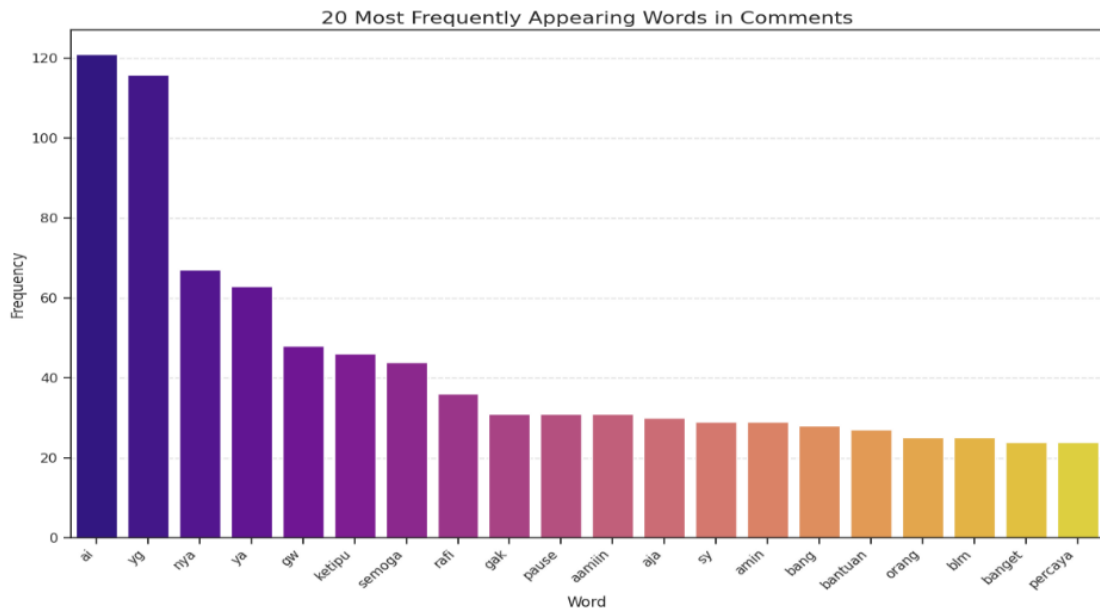


Figure 7. 20 most frequently appearing words in comments.

Bigram Analysis the most common two-word phrases were "jackie chan," "ya allah," "terima kasih," "mr bert," and "rafi ahmad." This distribution indicates that comments addressed both technological aspects and the public figures featured in the videos, as well as included expressions of emotion, as shown in Figure 8.

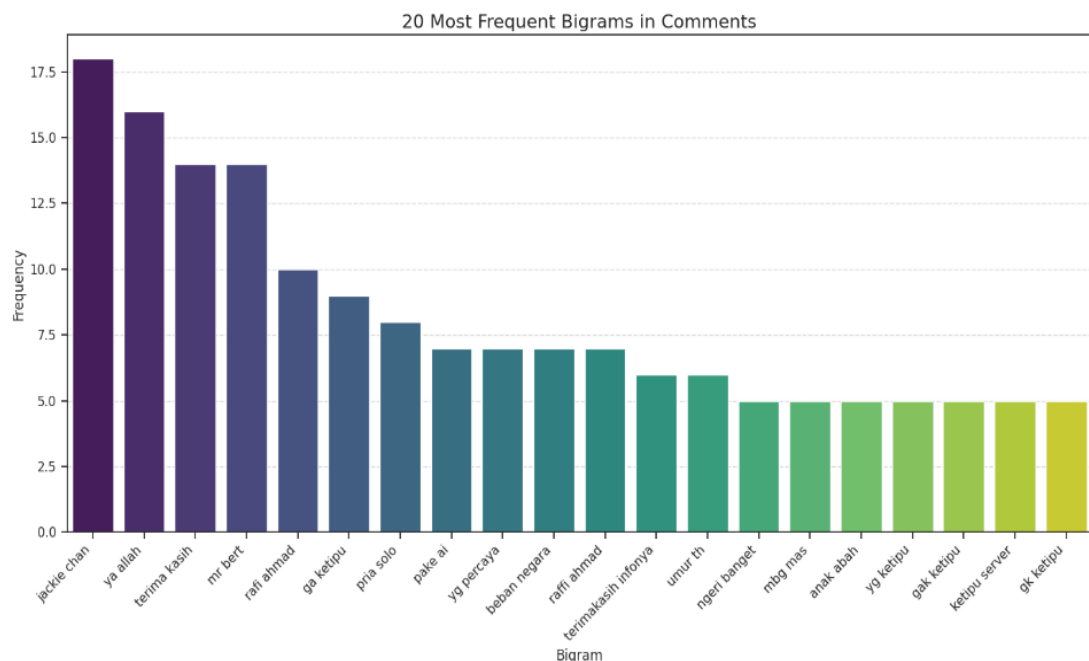


Figure 8. 20 Bigrams that appear most frequently in comments

Emoji Analysis the most frequently used emoji was the "Face With Tears Of Joy," followed by "Person With Folded Hands" and "Loudly Crying Face." The predominance of the laughing emoji indicates that many viewers perceived deepfake videos as entertaining. The folded-hands emoji is associated with comments expressing prayers or hopes, while the crying-face emoji likely reflects surprise at the convincing quality of the videos. These patterns suggest that emoji usage could be informative for sentiment analysis in subsequent research, as shown in Figure 9.

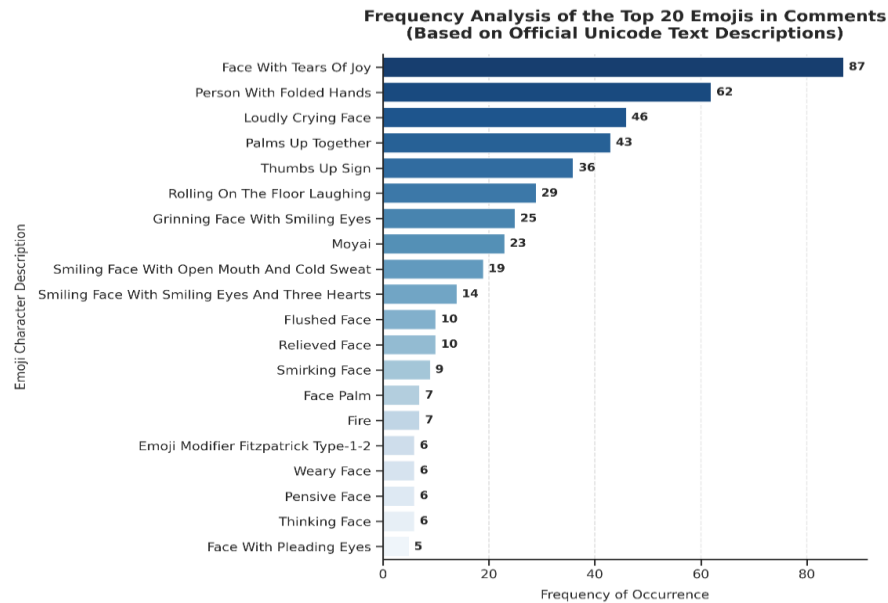


Figure 9. Frequency analysis of the top 20 emojis in comments

3.3 Data Labeling Results

Manually annotating all 1,071 comments produced three labels MIS, OTH, and SKP whose distribution is visualized as a bar chart in Figure 10 and a pie chart in Figure 11. Table 3 presents the frequency and percentage of each annotation label used in the classification task. As Table 3 shows, MIS constitutes the largest class, indicating that the majority of annotated comments are directly related to misinformation or deepfake content. Although the class distribution is not perfectly balanced, the degree of imbalance remains within a range that can be effectively addressed through weighted loss during model fine-tuning.

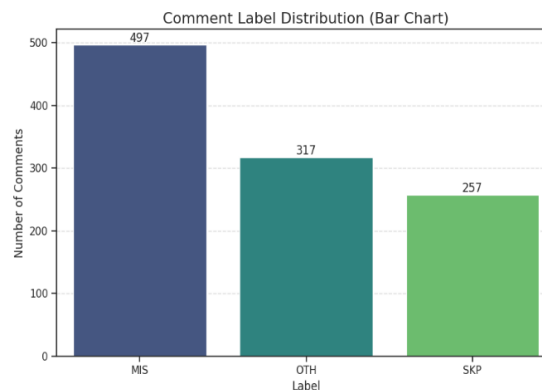


Figure 10. Bar chart of comment label distribution

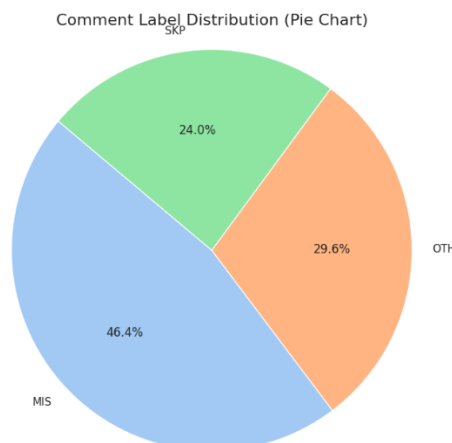


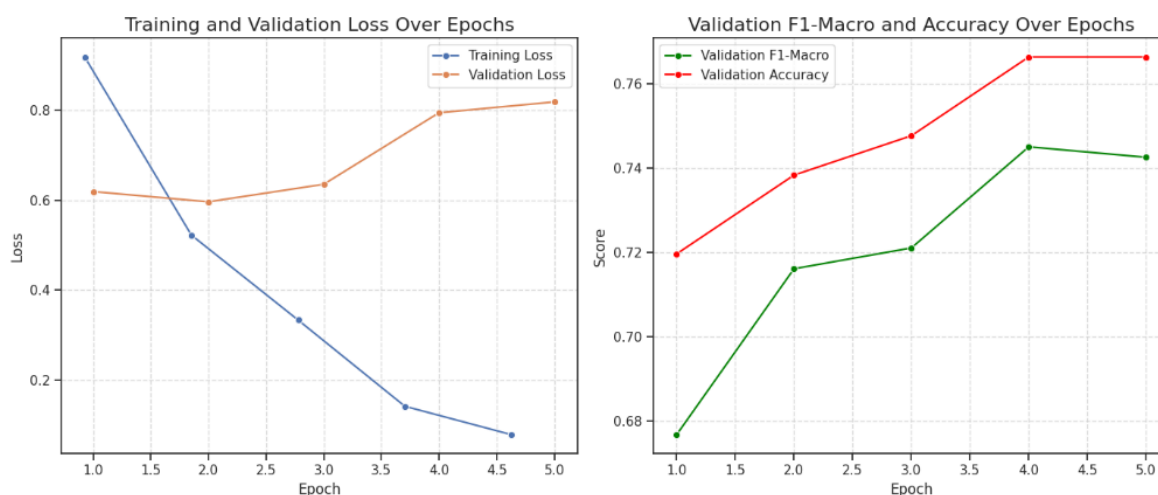
Figure 11. Pie chart of comment label distribution

Table 3. Data Label Results

Label	Amount	Percentage
MIS	497	46.41
OTH	317	29.60
SKP	257	24.00

3.4 IndoBERT Fine-tuning Results

Figure 12 plots the training and validation loss curves recorded across the five fine-tuning epochs, while Table 4 lists the corresponding accuracy and F1-Macro at each epoch. As shown in Figure 12, the training loss decreased consistently over five epochs, from 0.9135 initially to 0.0881 at the conclusion, indicating effective convergence. In contrast, the validation loss decreased until the second epoch 0.6262 but subsequently increased to 0.9062 by the fifth epoch. This divergence indicates the onset of overfitting after the second epoch, as the model begins to fit idiosyncrasies in the training data rather than generalizing. The validation F1-Macro followed a similar trend, increasing from 62.91% in epoch 1 to a maximum of 72.50% in epoch 3, then declining to 69.75% in epoch 4 and slightly increasing to 70.82% in epoch 5. Based on this pattern, the early-stopping procedure, which monitored validation F1-Macro, selected epoch 3 as the optimal checkpoint, corresponding to the peak in validation performance prior to the onset of overfitting.

**Figure 12.** IndoBERT learning curve**Table 4.** Training Development

Epoch	Accuracy	F1-Macro
1	68.22	62.91
2	71.96	69.47
3	74.77	72.50
4	72.90	69.75
5	73.83	70.82

3.5 Model Evaluation

Figure 13 presents the confusion matrix produced on the held-out test set, and Table 5 reports the corresponding precision, recall, and F1-score for each label. The confusion matrix in Figure 13 lays out correct versus incorrect predictions for each class and anchors the error-pattern discussion that follows.

On the 108-comment test set, IndoBERT achieved 82.41% accuracy, 81.14% precision, 82.25% recall, and 81.58% F1-Macro, indicating balanced performance across all three classes. F1-Macro was selected as the primary evaluation metric due to its reduced sensitivity to class imbalance, providing a more reliable assessment of model performance than accuracy alone in this context.

Within the 108-comment test set, class distribution was as follows: 50 MIS, 32 OTH, and 26 SKP comments. The macro-average precision, recall, and F1-score were 0.81, 0.82, and 0.82, respectively, while the weighted averages were 0.83 for precision and 0.82 for both recall and F1-score. The similarity between macro and weighted averages indicates that model performance remains robust despite class imbalance. The MIS class achieved the highest F1-score 0.88, whereas the OTH class recorded a lower F1-score 0.76, likely due to greater linguistic variability within the OTH category.

This heterogeneity contributes to the model's difficulty in consistently classifying OTH comments, in contrast to MIS comments, which typically exhibit more explicit and topic-specific language. Analysis of the confusion matrix

indicates that most misclassifications occur between OTH and SKP, likely because implicitly skeptical comments and off-topic comments may appear similar in their surface linguistic features, complicating their differentiation based solely on textual content.

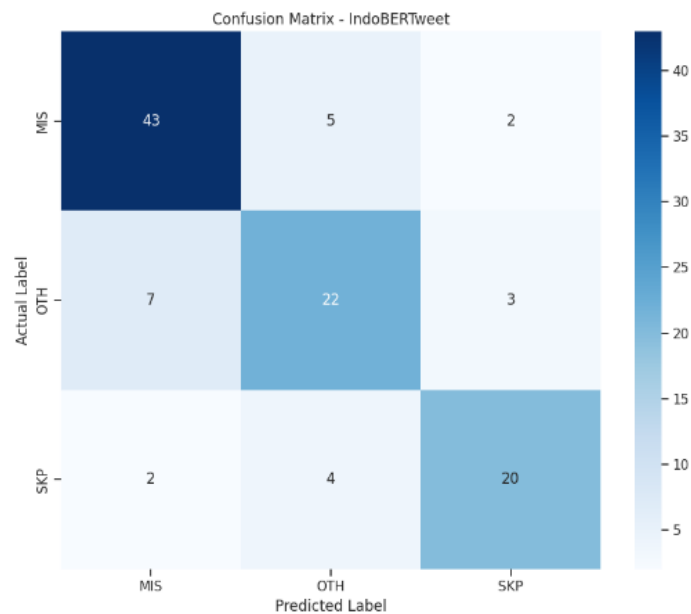


Figure 13. Confusion Matrix

Table 5. Classification Report

Label	Precision	Recall	F1-score
MIS	0.91	0.84	0.88
OTH	0.74	0.78	0.76
SKP	0.79	0.85	0.81

3.6 5-Fold Cross Validation Results

To check that the single-split result was not a lucky partition, IndoBERT was also validated with Stratified 5-Fold Cross-Validation; the averaged scores across the five folds are summarized in Table 6. The F1-Macro standard deviation of 0.0289 across folds indicates stable model performance, closely aligning with the single-split result of 81.58%. This consistency suggests that the model generalizes effectively and is not reliant on a specific data partition.

Table 6. Classification Model Performance Results

Metric	Mark
Accuracy	81.11
Precision	80.15
Recall	81.25
F1-Macro	80.17

3.7 Comparison with Baseline Model

Table 7 places IndoBERT's fine-tuned performance side by side with the four TF-IDF-based baselines under identical data splits. To evaluate the added value of a Transformer-based model in this context, IndoBERT was compared with four TF-IDF-based classifiers (using unigrams and bigrams) under identical training, validation, and test splits to ensure a fair comparison.

IndoBERT outperformed all four baseline models across each evaluation metric. The F1-Macro difference between IndoBERT (81.58%) and the strongest baseline, Logistic Regression (67.74%), was 13.84 percentage points. The other baselines achieved F1-Macro scores between 65.83% and 67.04%, indicating that TF-IDF-based representations may have an inherent performance limitation in this task, irrespective of the classifier used.

This performance gap is consistent with the limitations of TF-IDF, which relies on lexical overlap and sparse representations and is therefore less effective at capturing semantic relationships between words. These limitations are exacerbated in this dataset due to the brevity of comments (average 6.54 words) and the prevalence of informal language, abbreviations, slang, and code-mixing, which increase feature sparsity and hinder the recognition of

semantically similar expressions. In contrast, IndoBERT leverages contextual representations pretrained on a large Indonesian corpus, enabling it to interpret sentence meaning beyond simple word occurrence.

This advantage is consistent with earlier observations that OTH and SKP classes are frequently confused due to surface-level linguistic similarity, a scenario in which TF-IDF, lacking contextual awareness, is particularly limited. Collectively, these results support the conclusion that IndoBERT's superior performance is attributable to its contextual representation capabilities rather than random variation.

It should be noted that the baseline models were evaluated using a single data split and were not subjected to 5-Fold Cross Validation as IndoBERT was. Therefore, a comprehensive comparison of model stability between the two approaches remains an area for future investigation.

Table 7. Comparison of IndoBERT Performance with Baseline Model

Model	Accuracy	Precision	Recall	F1-Macro
IndoBERT (Fine-tuned)	82.41	81.14	82.25	81.58
Logistic Regression	70.37	68.35	67.40	67.74
Random Forest	69.44	70.98	65.85	67.04
Multinomial Naive Bayes	70.37	69.74	65.79	66.87
Linear SVM	68.52	67.21	65.31	65.83

3.8 Discussion

Placed alongside prior work, IndoBERT's 81.58% F1-Macro is lower than the accuracy figures reported for hybrid architectures evaluated on larger, non-Indonesian corpora for example, a BERT-LSTM model reached 93.51% accuracy and 92.73% F1 on general misinformation text [17], and a BERT-BiLSTM model reached 0.94 accuracy on toxic-comment data [18] which suggests that combining IndoBERT with a sequence layer, or training on a larger labeled set, is a plausible way to close that gap in future work. At the same time, IndoBERT's margin over Logistic Regression here 13.84 points in F1-Macro is comparable in direction to earlier findings that contextual BERT-style models outperform lexical baselines on Indonesian misinformation-style classification, such as the two-stage BERT classifier used for Indonesian COVID-19 misinformation tweets [11], reinforcing that the advantage of contextual representations over TF-IDF-based features is not specific to this dataset but consistent with prior Indonesian-language findings. This addresses the research problem directly: because IndoBERT can read short, informal TikTok comments and still separate misinformation-leaning reactions from skeptical or off-topic ones with F1-Macro well above the strongest lexical baseline, comment text is a workable, deployable signal for flagging misinformation-adjacent risk around AI-manipulated video content, even without analyzing the underlying audio or video.

Several limitations should be acknowledged. First, the labeling process was conducted by a single researcher, precluding the calculation of inter-annotator agreement metrics such as Cohen's Kappa, and thus the objectivity of the labels requires further validation. Second, the dataset, comprising 1,071 comments from 15 purposively selected videos, is limited in both size and diversity, and the generalizability of the model to broader Indonesian-language TikTok content remains to be established with larger and more varied data. Third, the class distribution (MIS 46.41%, OTH 29.60%, SKP 24.00%) is imbalanced, which may still impact the model's ability to accurately classify minority classes, despite the use of weighted cross-entropy loss.

Future work could reasonably expand the data volume, bring in multiple annotators for validation, and compare against other social-media-tuned BERT variants such as IndoBERTweet to strengthen these findings.

4. CONCLUSION

Data cleaning was a critical step, reducing the dataset from 1,345 to 1,071 comments suitable for classification. Exploratory analysis indicated that deepfake-related TikTok comments are typically brief, averaging 6.54 words, and are primarily focused on issues of AI authenticity. Fine-tuning IndoBERT resulted in 82.41% accuracy and 81.58% F1-Macro, with 5-fold cross-validation yielding an average F1-Macro of 80.17% and a low standard deviation, indicating stable model performance. Compared to four conventional TF-IDF-based baselines (Logistic Regression, Random Forest, Multinomial Naive Bayes, and Linear SVM), IndoBERT consistently outperformed the best baseline by up to 13.84 points in F1-Macro, providing strong empirical support for the use of Transformer-based models in this context. These findings further suggest that contextual representations are particularly effective for short, informal, and code-mixed text, where TF-IDF-based features are limited. However, the study is constrained by the relatively small, single-language dataset and the exclusion of visual and audio modalities. Future research should consider expanding to multilingual and multimodal datasets, increasing data volume and diversity, and evaluating lighter Transformer variants for real-time moderation applications. A further limitation is that all comments were labeled by a single researcher without a second annotator to compute inter-annotator agreement, so the objectivity of the MIS, SKP, and OTH labels has not been independently confirmed. The main scientific contribution of this work is showing, with a labeled Indonesian TikTok dataset, an IndoBERT-versus-baseline comparison, and 5-fold validation, that comment text alone can carry a usable misinformation-risk signal for AI-generated video content, a

signal that current detection pipelines built around the media itself do not capture. Overall, IndoBERT demonstrates potential as a foundation for automated misinformation detection on social media platforms.

5. DECLARATIONS

5.1 Author Contributions

Conceptualization: T.P., and I.S.W.;

Methodology: T.P.;

Software: T.P.;

Validation: T.P., I.S.W.;

Formal Analysis: T.P.;

Investigation: T.P.;

Resources: T.P.;

Data Curation: T.P.;

Writing Original Draft Preparation: T.P.;

Writing Review and Editing: T.P., and I.S.W.;

Visualization: T.P.;

All authors have read and agreed to the published version of the manuscript.

5.2 Data Availability

The data presented in this study are available on request from the corresponding author.

5.3 Funding

This research and the publication of this manuscript were fully funded by the author's personal funds. No external funding was received for the research, preparation, or publication of this manuscript.

5.4 Institutional Review Board Statement

Not applicable.

5.5 Informed Consent Statement

Not applicable.

5.6 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

5.7 Generative AI and AI-assisted Technologies in The Writing Process

The authors used Generative AI to improve the writing clarity of this paper. They reviewed and edited the AI-assisted content and took full responsibility for the final publication.

REFERENCES

- [1] F. N. Fa'izah and R. J. Kurniasari, "Deep fake crime on social media: A juridical analysis of the spread of hoaxes and the manipulation of public opinion in state security," *Yurijaya: Jurnal Ilmiah Hukum*, vol. 7, no. 3, pp. 522–550, 2025, doi: 10.51213/yurijaya.v7i3.238.
- [2] S. Alanazi and S. Asif, "Exploring deepfake technology: creation, consequences and countermeasures," *Human-Intelligent Systems Integration*, vol. 6, no. 1, pp. 49–60, Dec. 2024, doi: 10.1007/s42454-024-00054-8.
- [3] Á. F. Gambín, A. Yazidi, A. Vasilakos, H. Haugerud, and Y. Djenouri, "Deepfakes: current and future trends," *Artif. Intell. Rev.*, vol. 57, no. 3, Art. no. 64, Mar. 2024, doi: 10.1007/s10462-023-10679-x.
- [4] A. H. Soudy *et al.*, "Deepfake detection using convolutional vision transformers and convolutional neural networks," *Neural Comput. Appl.*, vol. 36, no. 31, pp. 19759–19775, Nov. 2024, doi: 10.1007/s00521-024-10181-7.
- [5] H. Liu, Y. Chen, A. Narayanan, A. Balachandran, P. J. Moreno, and L. Wang, "Can DeepFake Speech be Reliably Detected?," arxiv, Oct. 2024, [Online]. Available: <http://arxiv.org/abs/2410.06572>
- [6] L. Li, T. Lu, X. Ma, M. Yuan, and D. Wan, "Voice Deepfake Detection Using the Self-Supervised Pre-Training Model HuBERT," *Applied Sciences*, vol. 13, no. 14, p. 8488, 2023, doi: 10.3390/app13148488.
- [7] L. Luceri, T. V. Salkar, A. Balasubramanian, G. Pinto, C. Sun, and E. Ferrara, "Coordinated Inauthentic Behavior on TikTok: Challenges and Opportunities for Detection in a Video-First Ecosystem," Oct. 2025, [Online]. Available: <http://arxiv.org/abs/2505.10867>
- [8] R. H. Nugroho, I. R. Kusumasari, V. Febrianto, M. A. Farhan N. H, and M. R. Mahardika, "Strategi Teknologi Artificial Intelligence (AI) dalam Pengambilan Keputusan Bisnis di Era Digital," *Jurnal Bisnis dan Komunikasi Digital*, vol. 2, no. 2, p. 7, Dec. 2024, doi: 10.47134/jbk.d.v2i2.3476.

- [9] R. Anggrainingsih, G. M. Hassan, and A. Datta, "Evaluating BERT-based language models for detecting misinformation," *Neural Comput. Appl.*, vol. 37, no. 16, pp. 9937–9968, Jun. 2025, doi: 10.1007/s00521-025-11101-z.
- [10] M. L. Jamil, S. Pais, J. Cordeiro, and G. Dias, "Detection of extreme sentiments on social networks with BERT," *Soc. Netw. Anal. Min.*, vol. 12, no. 1, Dec. 2022, doi: 10.1007/s13278-022-00882-z.
- [11] D. R. Faisal and R. Mahendra, "Two-Stage Classifier for COVID-19 Misinformation Detection Using BERT: a Study on Indonesian Tweets," arxiv, Jun. 2022, [Online]. Available: <http://arxiv.org/abs/2206.15359>
- [12] F. Abbas and A. Taeihagh, "Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence," *Expert Syst. Appl.*, vol. 252, p. 124260, Oct. 2024, doi: 10.1016/j.eswa.2024.124260.
- [13] M. Masood, M. Nawaz, K. M. Malik, A. Javed, and A. Irtaza, "Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward," *Appl. Intell.*, vol. 53, pp. 3974–4026, 2023, doi: 10.1007/s10489-022-03766-z.
- [14] J. J. Bird and A. Lotfi, "Real-time Detection of AI-Generated Speech for DeepFake Voice Conversion," Aug. 2023, [Online]. Available: <http://arxiv.org/abs/2308.12734>
- [15] Z. Xu, X. Wen, G. Zhong, X. Pan, and C. Zhang, "Public perception towards deepfake through topic modelling and sentiment analysis of social media data," *Soc. Netw. Anal. Min.*, vol. 15, no. 16, pp. 1–13, 2025, doi: 10.1007/s13278-025-01445-8.
- [16] M. Wack and S. Prochaska, "Making sense of AI-generated disinformation: How audience interpretations influence the impact of deepfakes in Kenya," *Soc. Media Soc.*, vol. 12, no. 3, Jul. 2026, doi: 10.1177/20563051261462092.
- [17] J. Wang, X. Wang, and A. Yu, "Tackling misinformation in mobile social networks: a BERT-LSTM approach for enhancing digital literacy," *Sci. Rep.*, vol. 15, no. 1, Art. no. 1118, Jan. 2025, doi: 10.1038/s41598-025-85308-4.
- [18] Z. Fang, H. Zhang, J. He, Z. Qi, and H. Zheng, "Semantic and contextual modeling for malicious comment detection with BERT-BiLSTM," arXiv:2503.11084, Mar. 2025.
- [19] A. Bustamin, A. A. Prayogi, D. Siswanto, M. Rafrin, and A. Nurdin, "Text normalization for Indonesian slang words in sentiment analysis development," *ICIC Express Letters, Part B: Applications*, vol. 16, no. 2, pp. 121–129, Feb. 2025, doi: 10.24507/icicelb.16.02.121.
- [20] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
- [21] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd ed. draft, 2025. [Online]. Available: <https://web.stanford.edu/~jurafsky/slp3/>
- [22] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, *Dive into Deep Learning*. 2023.