Detecting YouTube Clickbait with Transformer Models: A Comparative Study

Abstrak

Clickbait masih menjadi strategi umum di YouTube, di mana judul video sering kali dibuat untuk memaksimalkan keterlibatan penonton. Meskipun teknologi machine learning berbasis Transformer telah berkembang pesat, studi yang secara khusus meneliti clickbait pada judul video YouTube masih jarang ditemukan, padahal judul-judul tersebut memiliki karakteristik bahasa yang unik, yaitu lebih pendek, informal, dan ambigu dibandingkan teks berita atau media sosial lainnya. Penelitian ini membandingkan tiga model berbasis Transformer, yaitu BERT, RoBERTa, dan XLNet, untuk tugas deteksi clickbait menggunakan dua dataset acuan. Setiap model di-fine-tune dan dievaluasi dengan metrik klasifikasi standar, disertai analisis efisiensi pelatihan dan inferensi. Hasil menunjukkan bahwa ketiga model mencapai akurasi di atas 95 persen. RoBERTa memberikan kinerja terbaik pada dataset Chaudhary (99,84 persen), sedangkan BERT-cased paling efektif pada dataset Vierti (96,91 persen). Sebaliknya, XLNet tertinggal dalam akurasi dan efisiensi komputasi, dengan waktu inferensi melebihi enam detik per batch. Penelitian ini menunjukkan peningkatan akurasi sebesar 1,31 persen dibandingkan metode SVM sebelumnya dan memberikan evaluasi komprehensif terhadap tiga arsitektur Transformer dalam konteks YouTube, menghasilkan panduan empiris untuk deteksi clickbait yang lebih efektif.

Kata kunci*—Clickbait, Youtube, Transformer, RoBERTa, Text Classification*

Abstract

Clickbait remains a common strategy on YouTube, where video titles are often crafted to maximize viewer engagement. Although transformer-based machine learning technologies have advanced rapidly, studies that specifically investigate clickbait in YouTube video titles are still rare, even though such titles have unique linguistic characteristics that are shorter, more informal, and more ambiguous than news headlines or other social media texts. This study compares three Transformer models, namely BERT, RoBERTa, and XLNet, for the task of clickbait detection using two benchmark datasets. Each model was fine-tuned and evaluated using standard classification metrics, with additional analyses on training and inference efficiency. The results show that all three models achieved accuracy above 95 percent. RoBERTa achieved the best performance on the Chaudhary dataset (99.84 percent), while BERT cased performed best on the Vierti dataset (96.91 percent). In contrast, XLNet lagged in both accuracy and computational efficiency, with inference times exceeding six seconds per batch. This study demonstrates a 1.31 percent improvement in accuracy compared to previous SVM-based methods and provides a comprehensive evaluation of three Transformer architectures in the YouTube context, offering empirical guidance for more effective clickbait detection.

Keywords—*Clickbait, Youtube, Transformer, RoBERTa, Text Classification*

1. INTRODUCTION

YouTube has grown into one of the most influential digital platforms, where people not only watch but also interact with an immense variety of video content. Every day, millions of new videos are uploaded, creating a vast pool of information and entertainment. To stand out in the highly competitive environment, many creators adopt strategies aimed at boosting engagement and visibility. One common tactic is the use of clickbait, which involves creating provocative, exaggerated, or misleading video titles that are designed to entice users to click [1].

Although clickbait can temporarily increase view counts and advertising revenue, it often diminishes the overall user experience. Misleading titles tend to frustrate audiences, erode their trust in the creator, and contribute to the spread of misinformation. In the long run, this practice not only affects user satisfaction but also harms the credibility and trust of the platform media itself[2]. As manual detection of clickbait is inefficient and subjective, there is a growing need for automated methods to accurately identify and mitigate such content [1].

Developments in Natural Language Processing (NLP) recently have led to Transformer-based models that are superior to traditional approaches in text classification tasks[3]. Among these models are BERT (Bidirectional Encoder Representations from Transformers), XLNet, and RoBERTa (Robustly Optimized BERT Pretraining Approach). BERT applies a bidirectional auto-encoding method to understand context more effectively[4], while XLNet integrates autoregressive modeling with permutation of word sequences to capture deeper semantic dependencies[5]. RoBERTa, an enhanced version of BERT, undergoes more extensive pretraining [5]. These models have also shown strong performance in detecting clickbait, particularly in the context of online news and social media. A study by [4] applied BERT to English news headlines and achieved 98.86% accuracy, outperforming traditional machine learning methods. Study by [6] utilized RoBERTa and IndoBERT on Indonesian news data, attaining accuracy above 92%. Finaly, a study by [5] compared BERT, XLNet, and RoBERTa for clickbait detection on Twitter posts and found that RoBERTa performed best overall, while XLNet showed notable strength in handling out-of-domain data.

Studies that directly compare the three models, namely BERT, XLNet, and RoBERTa, specifically in the context of YouTube video titles, remain scarce, despite the strong performance of Transformer models in other domains. Compared to formal news headlines, YouTube titles are typically shorter, less structured, and highly informal, making them difficult to classify [7]. One notable study [8] attempted to detect clickbait on YouTube using traditional machine learning techniques such as Naïve Bayes, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM), trained on a dataset of 31,987 video titles. The highest performance was achieved by an SVM model using a kernel-based TF-IDF representation, reaching 98.53% accuracy, precision, recall, and F1-score. However, no Transformer-based models were explored in that work, leaving open the question of whether modern NLP architecture could provide further improvements. Moreover, the influence of dataset variations on model performance has not been widely explored.

To address these gaps, this study implements and compares the three models on YouTube clickbait detection using two English-language datasets with different characteristics. In addition to evaluating classification metrics such as accuracy, precision, recall, and F1-score, this study also examines training and inference time efficiency, supported by statistical testing using Cochran's Q test and McNemar’s post hoc analysis. This study contributes new empirical evidence on the comparative effectiveness of Transformer models for clickbait detection in informal and unstructured digital content, particularly within real-world YouTube scenarios, thus addressing a gap in the existing body of research.

2. METHODS

The architecture of the proposed system is structured into four key stages: data collection, preprocessing, model generation, and model evaluation. A visual representation of this architecture is presented in Figure 1. The process begins with data collection, where two annotated datasets, one from Chaudhary [9] and another from Vierti [10], are gathered to ensure diversity and robustness in model training. These datasets then undergo a preprocessing phase consisting of tokenization and data splitting, with duplicated versions going through emoji removal. Following preprocessing, BERT, RoBERTa, and XLNet are fine-tuned to perform binary classification on the YouTube title data. Finally, the models are evaluated using a combination of standard classification metrics (accuracy, precision, recall, and F1-score), time-based metrics (training and inference time), and statistical significance testing through Cochran's Q Test and McNemar's Post Hoc analysis to determine the best-performing and statistically significant model variations.

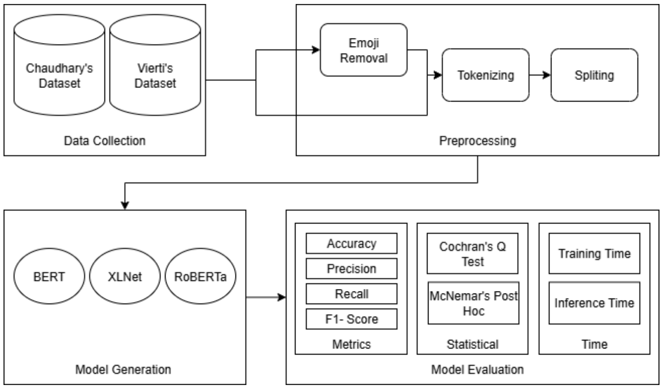

Figure 1. Proposed Method

2.1 Data Collection

The dataset used for clickbait classification in this study is sourced from two primary collections. The first dataset is the "Dataset of Clickbait and Non-clickbait Titles" obtained from Chaudhary's GitHub [9]. It contains 31,986 rows of labeled YouTube video titles, with two main columns: title, which represents the video title, and label, where 1 indicates clickbait and 0 indicates non-clickbait. The dataset is relatively balanced, comprising 16,000 clickbait entries and 15,986 non-clickbait entries. A few representative samples from the dataset are shown in Table 1.

Table 1 Example Entries from Chaudhary Dataset[9]

title	label
15 Highly Important Questions About Adulthood, Answered By Michael Ian Black	1
250 Nuns Just Cycled All The Way From Kathmandu To New Delhi	1
"Australian comedians ""could have been shot"" during APEC prank"	0
Lycos launches screensaver to increase spammers' bills	0
Fußball-Bundesliga 2008–09: Goalkeeper Butt signs with Bayern Munich	0

The second dataset is sourced from Vierti's "youtube-clickbait-detector", also obtained via GitHub [10]. The dataset was originally provided in pickle format and split into x_train, x_test, y_train, and y_test. For this study, training and testing splits were merged into a single CSV file. Only two main attributes were used in this study: video_title and label. The dataset contains 32,000 entries, evenly distributed between clickbait (16,000) and non-clickbait (16,000). Example entries from this dataset are shown in Table 2.

Table 2 Example Entries from Vierti Dataset[10]

title	label
5 TIMES BIGGEST WAVES SURFED CAUGHT ON CAMERA & SPOTTED IN REAL LIFE!	1
The Ugliest Wedding Dresses Ever Pt. 4	1
The Awesome And Inspiring Evolution Of Barbie Doll	1
Is the European Union Worth It Or Should We End It?	0
‘Father Figure’ 💼 The Patrick Star ‘Sitcom’ Show Episode 5 | SpongeBob SquarePants | Nick	0

In both datasets, the label 1 represents clickbait titles, while the label 0 denotes non-clickbait titles. The two datasets differ slightly in terms of linguistic tone and content style. A summary comparison is presented in Table 3 to highlight key differences.

Table 3 Comparison of the Datasets Used

Source (Year)	Dataset Name	Entries	Class Balance	Language Style	Attributes Used
Chaudhary (2024)[9]	Dataset of Clickbait and Non-clickbait Titles	31,986	16,000 clickbait / 15,986 non-clickbait	Formal, informative, neutral	title, label
Vierti (2023)[10]	youtube-clickbait-detector	32,000	16,000 clickbait / 16,000 non-clickbait	Provocative, hyperbolic, emotional	video_title, label

2.2 Preprocessing

From the dataset collected, we conducted preprocessing to ensure the textual data was ready to acquire the best result for classification. Emojis were removed in one approach, but were left them remain in the second approach to investigate the effect of minimal cleaning on model performance[11]. After that, all datasets will be tokenized using the native tokenizer of each Transformer model (BERT, RoBERTa, and XLNet). However, no stopword removal, punctuation stripping, or lowercasing was applied, as such elements may carry discriminative signals that differentiate clickbait from non-clickbait [5], [12]. Each dataset was divided into training, validation, and test sets with a ratio of 80:10:10 to ensure consistent evaluation [8]. The split was stratified to maintain the original class distribution across all subsets.

2.3 Model Generation

Transformer-based models have become the foundation of modern natural language processing (NLP) tasks due to their ability to model long-range dependencies through self-attention mechanisms [13]. This study adopts BERT, XLNet, and RoBERTa as the core text classification models, each distinguished by its unique pretraining approach influencing performance on downstream tasks.

All models were trained using the Hugging Face Trainer API with early stopping to prevent overfitting. Training was conducted on Quadro RTX 5000 16GB GPU with automatic memory optimizations such as mixed-precision training (fp16) and gradient checkpointing, except for XLNet, which does not support it. Evaluation was performed at the end of each epoch using the validation F1-score as the primary metric. Only the best checkpoint was retained to save storage, and unnecessary checkpoints were removed after training.

An initial hyperparameter search was conducted using the Trainer.hyperparameter_search function with 10 trials per model-dataset pair. The search space was customized per model to account for GPU memory constraints and specific architecture requirements. Key parameters included learning rate, batch size, number of epochs, and weight decay. For example, XLNet used smaller batch sizes due to the lack of gradient checkpointing support. This process identified a set of optimal base hyperparameters for each setting.

To further refine model performance, Optuna was employed for fine-tuning using the Tree-structured Parzen Estimator (TPE) sampler and Median Pruner. The search focused on ranges around the previously discovered optimal values. Each trial used early stopping, a consistent evaluation strategy, and memory-efficient configurations. The best F1-score from Optuna was selected, and the corresponding hyperparameters were merged with the initial configuration for final model training.

2.3.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) utilizes a multi-layer bidirectional Transformer encoder to learn deep contextual representations of language by using Masked Language Modelling (MLM) and Next Sentence Prediction (NSP)[4]. In this study, two variants of BERT will be utilized. BERT-base-uncased lowercases all input and does not preserve casing information, while BERT-base-cased retains case sensitivity and distinguishes between uppercase and lowercase letters. This selection allows analysis of the impact of case information on clickbait detection performance [7].

2.3.2 XLNet

XLNet is an autoregressive (AR) model that uses a Permutation Language Modeling (PLM) mechanism to capture bidirectional context. It maximizes the expected log-likelihood over all possible permutations of the token sequence. XLNet also employs two-stream self-attention to distinguish content and query representations[5]. We used XLNet's only version, named xlnet-base-cased, which retains text casing.

2.3.3 RoBERTa

RoBERTa (A Robustly Optimized BERT Pretraining Approach) is an improved version of BERT that modifies the pretraining procedure to enhance performance. It removes the Next Sentence Prediction (NSP) objective and applies dynamic masking that changes with each training epoch. Moreover, RoBERTa is pretrained on a much larger corpus (160 GB compared to BERT's 16 GB), with larger mini-batches, higher learning rates, and longer training duration without sentence segmentation constraints, resulting in more robust language representations and demonstrating superior performance across various NLP benchmarks [5]. Roberta-base was used in this study.

2.4 Model Evaluation

The model's performance was evaluated using accuracy, precision, recall, and F1-score. These metrics are widely used in text classification tasks to assess the balance between correctly predicted instances and errors in the classification problem[14]. The values were computed on the test set after training using the best-performing hyperparameters and fine-tuned. Additionally, both training time and inference time were recorded to assess computational efficiency across different models, following best practices in benchmarking machine learning models [4].

For statistical significance testing, the best prediction result from each model was selected and compared using Cochran's Q test, which is appropriate for comparing multiple classifiers over the same dataset. If the Q test indicated significant differences, pairwise comparisons were conducted using the McNemar post-hoc test to identify which models differed significantly in their predictions.

This combination of statistical tests ensures robust validation of model performance differences beyond the chance level [15].

3. RESULTS AND DISCUSSION

This study employs two preprocessing approaches. In the first approach, the text data was directly tokenized using the native tokenizer of each Transformer model (BERT, RoBERTa, and XLNet) without additional modification. However, in the second approach, emojis were removed before tokenization using the respective native tokenizers. The second approach will be referred to as the pre-processed dataset from now on. Table 4 presents the detailed preprocessing steps applied in the experiments.

Table 4 Preprocessing Steps

<table>
<tr><th>Model</th><th>Original Text</th><th>Emoji Removal</th><th>Tokenized</th></tr>
<tr><td rowspan="2">BERT-Cased</td><td rowspan="7">‘Father Figure’ 💼 The Patrick Star ‘Sitcom’ Show Episode 5 | SpongeBob SquarePants | Nick</td><td>-</td><td>['“', 'Father', 'Figure', '”', '[UNK]', 'The', 'Patrick', 'Star', '“', 'Sit', '##com', '”', 'Show', 'Episode', '5', '|', 'S', '##po', '##nge', '##B', '##ob', 'Square', '##P', '##ants', '|', 'Nick']</td></tr>
<tr><td>The Patrick Star ‘Sitcom’ Show Episode 5 | SpongeBob SquarePants | Nick</td><td>['“', 'Father', 'Figure', '”', 'The', 'Patrick', 'Star', '“', 'Sit', '##com', '”', 'Show', 'Episode', '5', '|', 'S', '##po', '##nge', '##B', '##ob', 'Square', '##P', '##ants', '|', 'Nick']</td></tr>
<tr><td rowspan="2">BERT-Uncased</td><td>-</td><td>['“', 'father', 'figure', '”', '[UNK]', 'the', 'patrick', 'star', '“', 'sitcom', '”', 'show', 'episode', '5', '|', 'sponge', '##bo', '##b', 'square', '##pants', '|', 'nick']</td></tr>
<tr><td>The Patrick Star ‘Sitcom’ Show Episode 5 | SpongeBob SquarePants | Nick</td><td>['“', 'father', 'figure', '”', 'the', 'patrick', 'star', '“', 'sitcom', '”', 'show', 'episode', '5', '|', 'sponge', '##bo', '##b', 'square', '##pants', '|', 'nick']</td></tr>
<tr><td rowspan="2">XLNet</td><td>-</td><td>['▁“', 'Father', '▁Figure', '”', '▁', '💼', '▁The', '▁Patrick', '▁Star', '▁“', 'S', 'it', 'com', '”', '▁Show', '▁Episode', '▁5', '▁', '|', '▁Spo', 'nge', 'Bob', '▁Square', 'P', 'ants', '▁', '|', '▁Nick']</td></tr>
<tr><td>The Patrick Star ‘Sitcom’ Show Episode 5 | SpongeBob SquarePants | Nick</td><td>['▁“', 'Father', '▁Figure', '”', '▁The', '▁Patrick', '▁Star', '▁“', 'S', 'it', 'com', '”', '▁Show', '▁Episode', '▁5', '▁', '|', '▁Spo', 'nge', 'Bob', '▁Square', 'P', 'ants', '▁', '|', '▁Nick']</td></tr>
<tr><td>RoBERTa</td><td>-</td><td>['âĢ', 'Ī', 'Father', 'ĠFigure', 'âĢ', 'Ļ', 'ĠðŁ', 'Ĵ', '¼', 'ĠThe', 'ĠPatrick', 'ĠStar', 'ĠâĢ', 'Ī', 'Sit', 'com', 'âĢ', 'Ļ', 'ĠShow', 'ĠEpisode', 'Ġ5', 'Ġ|', 'ĠSponge', 'Bob', 'ĠSquare', 'P', 'ants', 'Ġ|', 'ĠNick']</td></tr>
</table>

Model	Original Text	Emoji Removal	Tokenized
		The Patrick Star 'Sitcom' Show Episode 5 | SpongeBob SquarePants | Nick	['âĢ', 'ĺ', 'Father', 'ĠFigure', 'âĢ', 'Ļ', 'Ġ', 'ĠThe', 'ĠPatrick', 'ĠStar', 'ĠâĢ', 'ĺ', 'Sit', 'com', 'âĢ', 'Ļ', 'ĠShow', 'ĠEpisode', 'Ġ5', 'Ġ|', 'ĠSponge', 'Bob', 'ĠSquare', 'P', 'ants', 'Ġ|', 'ĠNick']

The evaluation of BERT, RoBERTa, and XLNet on the Chaudhary[9] and Vierti[10] datasets reveals consistently strong performance across all Transformer-based models, with accuracy, precision, recall, and F1-scores above 95% in nearly all settings (Tables 5 and 6). This confirms the suitability of Transformer architectures for the task of clickbait detection in YouTube video titles. On the Chaudhary dataset[9], RoBERTa achieved the highest overall performance with an accuracy of 99.84%, precision of 99.87%, recall of 99.81%, and F1-score of 99.84%. On the Vierti dataset[10], performance was slightly lower than in Chaudhary[9]. The best-performing model, BERT-cased, achieved an accuracy of 96.91% with balanced precision (97.67%) and recall (96.05%). RoBERTa and XLNet also performed competitively, reaching accuracies of 96.87% and 96.71%, respectively. These results indicate that while all three Transformer models excel, their relative rankings differ depending on dataset characteristics.

Table 5 Evaluation metrics for the BERT, XLNet and RoBERTa models on Chaudhary Dataset

Model	Dataset	Accuracy	Precision	Recall	F1-Score
roberta-base	chaudhary	0.998437	0.998748	0.998124	0.998436
bert-cased	chaudhary-pre-processed	0.996874	0.997495	0.996248	0.996871
roberta-base	chaudhary-pre-processed	0.996874	0.997495	0.996248	0.996871
xlnet-base-cased	chaudhary	0.996561	0.994396	0.998749	0.996568
bert-cased	chaudhary	0.995311	0.993766	0.996873	0.995317
xlnet	chaudhary-pre-processed	0.994998	0.993762	0.996248	0.995003
bert-uncased	chaudhary-pre-processed	0.992185	0.993108	0.991245	0.992175
bert-uncased	chaudhary	0.990309	0.994949	0.985616	0.990261

Table 6 Evaluation results for the BERT, XLNet and RoBERTa models on Vierti Dataset

Model	Dataset	Accuracy	Precision	Recall	F1-Score
bert-cased	vierti	0.969069	0.976717	0.960504	0.968543
roberta-base	vierti-pre-processed	0.968785	0.961127	0.976531	0.968768
xlnet-base-cased	vierti	0.967083	0.963089	0.970807	0.966933
bert-cased	vierti-pre-processed	0.965664	0.969400	0.961076	0.965220
roberta-base	vierti	0.965380	0.960340	0.970235	0.965262
xlnet-base-cased	vierti-pre-processed	0.964813	0.967723	0.961076	0.964388
bert-uncased	vierti	0.963961	0.953020	0.975386	0.964074

Model	Dataset	Accuracy	Precision	Recall	F1-Score
bert-uncased	vierti-pre-processed	0.958286	0.976190	0.938752	0.957105

The performance gap between Chaudhary[9] and Vierti[10] highlights the impact of dataset composition on model effectiveness. The Chaudhary[9] dataset contains more formal and neutral titles, structurally resembling traditional news headlines. In contrast, the Vierti dataset comprises highly informal, emotional, and hyperbolic titles more typical of YouTube content. This stylistic variability likely explains why absolute performance was lower on Vierti[10], suggesting that informal linguistic cues such as exaggeration, slang, and emojis pose greater challenges for Transformer models. The confusion matrices (Figures 2–4) further illustrate this difference, showing that models produced more false positives and false negatives on the Vierti[10] dataset compared to Chaudhary[9]. Interestingly, the effect of light preprocessing (emoji removal) was inconsistent. On the Chaudhary[9] dataset, preprocessing had a negligible impact, whereas on Vierti[10], it slightly improved performance for RoBERTa (from 96.53% to 96.88%) but reduced performance for BERT-uncased (from 96.40% to 95.71%). These results imply that emojis and other stylistic markers can serve as discriminative signals for clickbait detection, and their removal may strip away useful information in certain contexts.

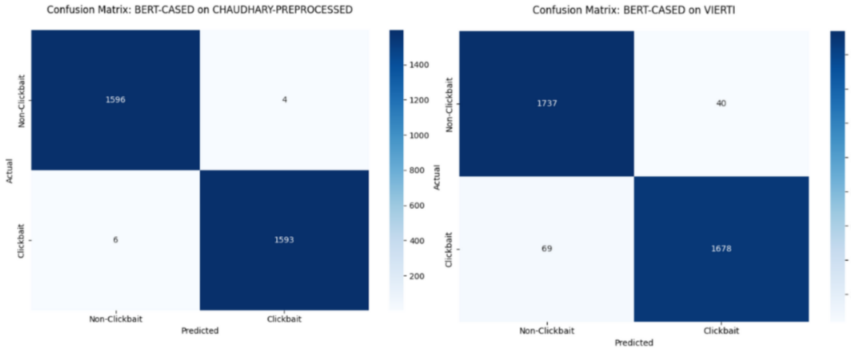

Figure 2 Confussion Matrix Best BERT Model Comparison

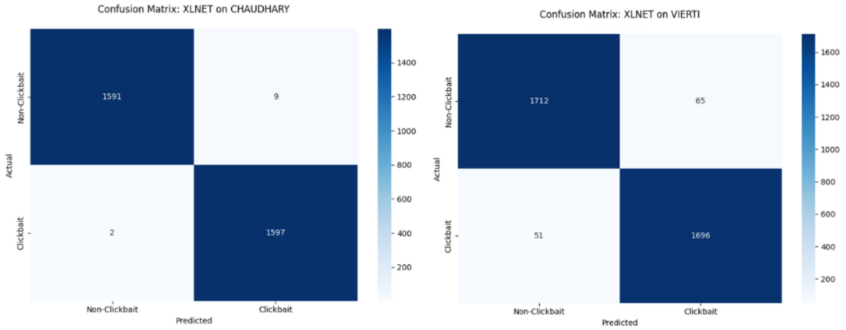

Figure 3 Confussion Matrix Best XLNet Model Comparison

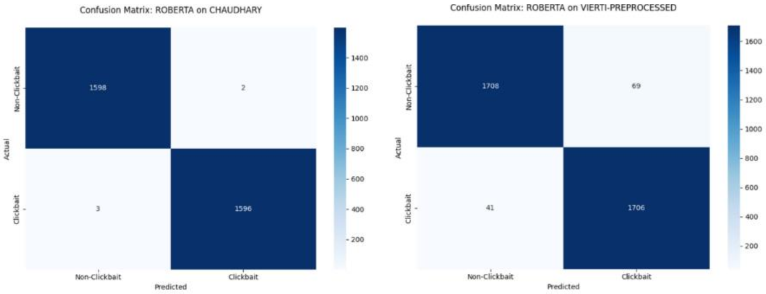

Figure 4 Confussion Matrix Best RoBERTa Comparison

The results of Cochran's Q test (Table 7) confirm that the observed performance differences among best version of the models are statistically significant on both datasets. On both datasets, the null hypothesis of equal classifier performance was rejected ($p < 0.05$). Consequently, McNemar's post hoc test was conducted to identify pairwise differences. Post hoc McNemar analysis further provides more granular insights. On the Chaudhary dataset[9], BERT and RoBERTa, as well as RoBERTa XLNet did not differ significantly, while BERT-XLNet had a significant difference ($p < 0.05$). On the Vierti dataset[10], significant differences were observed between BERT and RoBERTa, as well as between BERT and XLNet, while RoBERTa and XLNet remained statistically indistinguishable. These findings, as presented in Table 8, validate that model superiority is dataset-dependent and emphasize the importance of comparative evaluation rather than relying on single-model benchmarks.

Table 7 Result of Cochran's Q Test

Dataset (Year)	Cochran's Q statistic	P-value	Conclusion
Chaudhary (2024)[8]	8.111	0.017	Significant differences
Vierti (2023)[10]	40.106	0.000	Significant differences

Table 8 Result of McNemar's Post hoc Test

Dataset (Year)	Comparison	McNemar Method	McNemar's statistic	P-value	Conclusion
Chaudhary (2024)[8]	BERT vs RoBERTa	Binomial	5.0	1.0000	No significant differences
	BERT vs XLNet		1.0	0.0117	Significant differences
	RoBERTa vs XLNet		3.0	0.0574	No significant differences
Vierti (2023)[10]	BERT vs RoBERTa	Chi-square	35.7032	0.0000	Significant differences
	BERT vs XLNet		21.2528	0.0000	Significant differences
	RoBERTa vs XLNet		2.2790	0.1311	No significant differences

Across datasets, RoBERTa consistently achieved strong or near-best results, particularly on the Chaudhary dataset[9], likely due to its robust pretraining and optimization strategies. At the same time, BERT-cased, after fine-tuning, was able to rival RoBERTa in terms of classification metrics. Although the confusion matrices (Figures 2–4) and statistical tests (Table 7) indicate broadly similar error distributions and comparable significance patterns between BERT and RoBERTa, their computational efficiency varied by dataset, as seen in Tables 9 and 10. Specifically, RoBERTa was faster to train and infer on the Chaudhary dataset[9], while BERT-cased showed shorter training and inference times on the Vierti dataset[10]. By contrast, XLNet exhibited the slowest training and inference times (over 6 seconds per inference batch in some cases) while only achieving marginally lower accuracy than BERT and RoBERTa. Given its higher computational demands and lack of consistent performance advantage, XLNet appears less efficient for real-world deployment in clickbait detection systems.

Table 9 Training and Inference Time of The Models on Chaudhary Dataset

Model	Dataset	Train Time	Inference Time
roberta-base	chaudhary	147.012148	1.988522
bert-cased	chaudhary-pre-processed	433.904581	2.066190
roberta-base	chaudhary-pre-processed	213.094364	2.152620
xlnet-base-cased	chaudhary	574.133470	6.047956
bert-cased	chaudhary	251.715735	2.041615
xlnet-base-cased	chaudhary-pre-processed	452.743262	6.167669
bert-uncased	chaudhary-pre-processed	450.093621	2.110424
bert-uncased	chaudhary	191.597009	2.098935

Table 10 Training and Inference Time of The Models on Vierti Dataset

Model	Dataset	Train Time	Inference Time
bert-cased	vierti	370.929373	2.164251
roberta-base	vierti-pre-processed	376.969751	3.578848
xlnet-base-cased	vierti	964.197576	6.690678
bert-cased	vierti-pre-processed	237.648406	2.351731
roberta-base	vierti	498.524578	3.404936
xlnet-base-cased	vierti-pre-processed	501.598217	6.620288
bert-uncased	vierti	369.582200	2.258519
bert-uncased	vierti-pre-processed	370.718779	2.367368

Overall, we conclude RoBERTa with an accuracy of 99.84% in Chaudhary's dataset[9] is the best performing model on classifying Youtube video clickbait tittle as seen in table 11. This marks a notable improvement over the strongest traditional baseline reported in prior work, namely SVM with TF–IDF features, which achieved 98.53% across all metrics[8]. While the absolute accuracy gain of RoBERTa over SVM is relatively modest (+1.31%), the improvement is statistically meaningful given the large dataset size.

Table 11 Research Comparison on Chaudhary's Dataset

Year	Algorithm	Accuracy	Precision	Recall	F1-Score
2023	SVM	98.53%	98.53%	98.53%	98.53%
2025	**RoBERTa (Our Research)**	99.84%	99.87%	99.81%	99.84%

4. CONCLUSIONS

This study set out to evaluate the effectiveness of Transformer-based models: BERT, RoBERTa, and XLNet in detecting clickbait on YouTube video titles. The experiments revealed that although all three models delivered strong results, their performance varied depending on dataset characteristics. RoBERTa consistently demonstrated superior performance on Chaudhary's dataset, achieving 99.84% accuracy, while BERT-cased outperformed others on Vierti's dataset with 96.91%. This study clarified that no single model is perfect for all data scenarios. From this study, we learn that removing emojis might subtly change results, indicating that stylistic signals in titles may have predictive value in addition to raw accuracy. Overall, the models reach more than 95% accuracy in all scenarios; however, statistical testing reveals that there is a significant difference in the model's results. From an efficiency standpoint, while RoBERTa generally achieved the highest accuracy, fine-tuned BERT can compete with a tolerable level of predictive performance while having faster training and inference, making it a strong candidate in scenarios with limited computational resources. In practical terms, this study shows that Transformer models, particularly RoBERTa and BERT, are highly effective for detecting clickbait in informal online environments such as YouTube. While this study is limited to a single modality, future work should explore multimodal approaches by integrating textual, visual, and contextual features to better reflect real-world clickbait detection through all senses.

REFERENCES

[1] D. Varshney and D. K. Vishwakarma, "A unified approach for detection of Clickbait videos on YouTube using cognitive evidences," *Appl. Intell. Dordr. Neth.*, vol. 51, no. 7, pp. 4214–4235, 2021, doi: 10.1007/s10489-020-02057-9.

[2] D. Fayvishenko and I. Shudrak, "Clickbait and Its Impact on Media Trust: Analytical Review," *State Reg. Ser. Soc. Commun.*, no. 1(61), pp. 26–32, June 2025, doi: 10.32840/cpu2219-8741/2025.1(61).4.

[3] H. Lu, L. Ehwerhemuepha, and C. Rakovski, "A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance," *BMC Med. Res. Methodol.*, vol. 22, p. 181, July 2022, doi: 10.1186/s12874-022-01665-y.

[4] A. Chowanda, N. Nadia, and L. M. M. Kolbe, "Identifying clickbait in online news using deep learning," *Bull. Electr. Eng. Inform.*, vol. 12, no. 3, Art. no. 3, June 2023, doi: 10.11591/eei.v12i3.4444.

[5] P. Rajapaksha, R. Farahbakhsh, and N. Crespi, "BERT, XLNet or RoBERTa: The Best Transfer Learning Model to Detect Clickbaits," *IEEE Access*, vol. 9, pp. 154704–154716, 2021, doi: 10.1109/ACCESS.2021.3128742.

[6] J. Sirusstara, N. Alexander, A. Alfarisy, S. Achmad, and R. Sutoyo, "Clickbait Headline Detection in Indonesian News Sites using Robustly Optimized BERT Pre-training Approach (RoBERTa)," in *2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, Sept. 2022, pp. 1–6. doi: 10.1109/AiDAS56890.2022.9918678.

[7] R. Kemm, "The Linguistic and Typological Features of Clickbait in Youtube Video Titles," *Soc. Commun.*, vol. 23, no. 1, Art. no. 1, Jan. 2022, doi: 10.2478/sc-2022-0007.

[8] T. S. Y. Winarto, K. Wijaya, M. A. Faqih, S. Y. Prasetyo, and Y. Muliono, "Tackling Clickbait with Machine Learning: A Comparative Study of Binary Classification Models for YouTube Title," *Procedia Comput. Sci.*, vol. 227, pp. 282–290, Jan. 2023, doi: 10.1016/j.procs.2023.10.526.

[9] A. Chaudhary, "Dataset of clickbait and non-clickbait titles." Accessed: May 21, 2025. [Online]. Available: https://gist.github.com/amitness/0a2ddbcb61c34eab04bad5a17fd8c86b

[10] A. Vierti, *alessiovierti/youtube-clickbait-detector*. (Aug. 20, 2023). Jupyter Notebook. Accessed: May 21, 2025. [Online]. Available: https://github.com/alessiovierti/youtube-clickbait-detector

[11] H. Alawneh, A. Hasasneh, and M. Maree, “On the Utilization of Emoji Encoding and Data Preprocessing with a Combined CNN-LSTM Framework for Arabic Sentiment Analysis,” *Modelling*, vol. 5, no. 4, pp. 1469–1489, Dec. 2024, doi: 10.3390/modelling5040076.

[12] S. Kurniawan, A. S. Pramayoga, and Y. F. Ashari, “An Ensemble-Based Approach for Detecting Clickbait in Indonesian Online Media,” *J. Masy. Inform.*, vol. 16, no. 1, pp. 104–118, May 2025, doi: 10.14710/jmasif.16.1.73115.

[13] S. Islam *et al.*, “A comprehensive survey on applications of transformers for deep learning tasks,” *Expert Syst. Appl.*, vol. 241, p. 122666, May 2024, doi: 10.1016/j.eswa.2023.122666.

[14] F. S. Amalia and Y. Suyanto, “Offensive Language and Hate Speech Detection using BERT Model,” *IJCCS Indones. J. Comput. Cybern. Syst.*, vol. 18, no. 4, Art. no. 4, Oct. 2024, doi: 10.22146/ijccs.99841.

[15] I. J. David, M. U. Adehi, and P. O. Ikwuoche, “Cochran’s Q-Test on Soil Helminth Prevalence,” *Biom. Lett.*, vol. 58, no. 2, pp. 169–185, Dec. 2021, doi: 10.2478/bile-2021-0013.