



Benchmarking Transformer Models for Sentiment Analysis on YouTube User Comments

Li-Hua Li¹, Agus Cahyo Nugroho^{1,2*}

¹Department of Information Management, Chaoyang University of Technology, Taichung, Taiwan

²Department of Information System, Universitas Katolik Soegijapranata, Semarang, Indonesia

*agus.nugroho@unika.ac.id

Abstract

Social media has become one of the most important platforms for public expression due to its ability to provide users with an unprecedented scale for sharing opinions, sentiments, and experiences. Consequently, social media platforms provide valuable insights into public perception and user sentiment across various topics. This study presents a comparative evaluation of five transformer-based models, namely DistilBERT, BERT, BERTweet, RoBERTa, and Twitter-RoBERTa, on a cleaned and preprocessed YouTube Comments Dataset for sentiment analysis tasks. All models were evaluated under identical experimental settings and preprocessing procedures to reduce textual noise and improve data quality. Each transformer model was fine-tuned under common training conditions to assess its effectiveness in handling informal and user-generated social media content. Experimental results show that Twitter-RoBERTa outperforms other approaches and achieves the best overall accuracy: 90.13% (F1-score: 90.01%), demonstrating its ability to learn from contextual nuances and social media-specific language patterns. In addition, McNemar statistical analysis further supports the statistical significance of many performance differences between the models evaluated including those between Twitter-RoBERTa and smaller general-purpose models such as DistilBERT and BERT. These results demonstrate the value of using domain-specific pretraining on noisy social media text to improve sentiment classification performance and that transformer architectures built with fine-tuning in mind can effectively capture platform specific aspects of linguistics. In summary, this work presents a strong benchmark for transformer-based-sentiment-analysis on YouTube comments and valuable information for future research in social media sentiment analysis.

Keywords: bert; distilbert; roberta; sentiment analysis; twitter-roberta

How to Cite: L.-H. Li and A. C. Nugroho, "Benchmarking Transformer Models for Sentiment Analysis on YouTube User Comments", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 5, pp. 1277–1286, Oct. 2026, doi: [10.29207/resti.v10i5.7752](https://doi.org/10.29207/resti.v10i5.7752).
Permalink/DOI: <https://doi.org/10.29207/resti.v10i5.7752>

Received: April 25, 2026

Accepted: September 14, 2026

Available Online: October 2, 2026

*This is an open-access article under the [CC BY 4.0 License](https://creativecommons.org/licenses/by/4.0/)
Published by [Ikatan Ahli Informatika Indonesia](https://www.iaii.or.id/)*

1. Introduction

Sentiment analysis has become one of the most important tasks in Natural Language Processing (NLP) because social media platforms continuously generate massive amounts of user opinions that can support decision-making in business, healthcare, public policy, and marketing. Among various social media platforms, YouTube has emerged as one of the largest sources of user-generated content, where millions of users express their opinions through comments on videos covering diverse topics [1]. Although, YouTube comments contain informal writing styles and written in slang, abbreviated form and using some emojis, spelling variations, conversational speech patterns etc. These types of features add a great deal of noise to the data,

making sentiment classification much more difficult than in standard text classification tasks.

Since their release, improvements in contextual representation and semantic understanding of the text prompted the use of new transformer-based language models for many NLP tasks. BERT [2] introduced contextualized encodings in a bidirectional manner leading to significant improvement across many language understanding tasks. Later transformer models improved either compute efficiency or the way context is represented. DistilBERT [3] compressed the model with some loss in language understanding performance, while RoBERTa [4] optimized the pretraining strategy to maximize downstream task performance. Moreover, there are domain-specific transformer models including BERTweet and Twitter-

RoBERTa [5] which were pretrained on rich social media corpus to capture informal language, abbreviations, hashtags, emojis and other specific linguistic features of user-generated content.

Though much progress has been made, several research gaps still exist. First, existing studies primarily assess transformer models on benchmark sentiment datasets or data from Twitter, and there are few comprehensive explorations on Youtube comments. One, prior works usually analyze one transformer architecture or only a few models to evaluate which transformer is better when considering the same experimental settings combining lightweight, general-purpose and domain-specific transformers. Finally, few studies utilize hypothesis testing to significantly validate the observed performance differences among transformer models, limiting confidence in whether previously reported improvements are statistically meaningful.

This study fills this gap by providing a comparative analysis of five transformer-based models- DistilBERT, BERT, BERTweet, RoBERTa and Twitter-RoBERTa on the sentiment classification task using cleaned and preprocessed version of YouTube Comments Dataset. All models fine-tuned under the same preprocessing, training configuration and evaluation metrics for a fair comparison. In addition, McNemar's statistical test is applied to see if the apparent performance bias exists.

The main contributions of this study include a comprehensive comparative evaluation of DistilBERT, BERT, BERTweet, RoBERTa, and Twitter-RoBERTa for sentiment analysis on YouTube comments. The study investigates the robustness of general-purpose and domain-specific transformer models in handling the noisy, conversational, and informal language commonly found in social media content. It also examines the influence of domain-specific pretraining on sentiment classification performance for user-generated content. Furthermore, McNemar's test is employed to determine whether the performance differences among the evaluated models are statistically significant. Finally, this study establishes a benchmark for transformer-based sentiment analysis on YouTube comments, providing a useful reference for future research on social media sentiment analysis.

The results yield theoretical and practical implications regarding the effectiveness of transformer-based architectures for processing user generated content with noise in applications like sentiment analysis. It even provides advice on choosing the right transformer models for real-world NLP tasks like social media monitoring and opinion mining that involve informal typed online text, among others.

Sentiment analysis or opinion mining is one of the key tasks in Natural Language Processing. Sentiment analysis is supervised machine learning which is specifically trained to find and analyze the emotion or attitude of a given text. This is beneficial since it may offer knowledge into consumer behaviour, social trends

and public opinion due to the immense amount of data contributed by users [6].

Traditional Machine Learning techniques such as Naïve bayes, Support Vector Machines (SVM), and Logistic Regression have also been dated back to sentiment analysis. They are good with structured datasets and need manual feature engineering [6]; but for unstructured text user-generated content is generally ambiguous and colloquial which cannot be captured easily with such models [7].

But everything changed in the game with deep learning. Neural networks, Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and Long Short-Term Memory networks (LSTM) [8], [9] can automatically extract features from complex datasets and allow obtaining extraction concerning sequential and contextual characteristics. Context Vector Transformers-based models, for example BERT and RoBERTa [9] that can acquire context vector of words and long-distance dependencies have recently achieved SOTA (static-to-art) on sentiment classification tasks [10].

Domain-specific models and larger language models have been shown to outperform general-purpose models when analyzing YouTube comments and other social media texts, particularly because domain-specific pretraining enables them to better capture the informal and context-dependent nature of user-generated content [7], [11]. Sentiment analysis has been applied across various domains, including business intelligence for monitoring customer feedback and brand sentiment; culture and public opinion for tracking attitudes toward important social issues; healthcare for improving services based on patient experiences; and finance for predicting market movements based on investor sentiment [12].

Since the introduction of Transformer architecture by Vaswani et al. Natural Language Processing (NLP) has undergone a drastic shift and gave up recursion and convolutions in exchange for self-attention techniques [13] in 2017. As a result, this invention led to an increase in parallelization available within the model allowing for better long-distance correlations when compared with traditional recursive models. It led transformer models to quickly become the backbone of modern NLP applications.

In parallel, numerous variations of the standard Transformer model have been developed to increase performance and effectiveness. Devlin et al. introduced one such innovation called BERT (Bidirectional Encoder Representations from Transformers). using a bidirectional attention mechanism for contextual word understanding within phrases [14]. RoBERTa (Robustly Optimized BERT Approach) is a fine-tuned version of BERT that achieves improved performance in downstream tasks by optimizing pre-training strategy by increasing the data size and training time, as well as removing next sentence prediction task [15].

Another major improvement is the distillation of knowledge implemented by DistilBERT that learns a lighter model like BERT. This makes it suitable for use on edge devices since it reduces the model parametric value without sacrificing great loss in language interpret capability [16]. Generative models autoregressively apply the Transformer architecture, such as in language generation tasks like conversational AI [16] and content creation [17].

Some of the widespread applications are named entity recognition, text classification (text categorization), question answering, sentiment analysis and translation from one language to another. By modeling contextual information in an extremely parallelizable and scalable way, they have significantly advanced the state of the art in this domain. Still, it faces such challenges in computational cost management, long sequence processing and resource-poor devices utilization. And these restrictions push researchers to investigate the methods of domain adaptation, efficient attention mechanism and model compression.

BERT [18] (Bidirectional Encoder Representations from Transformers), a language representation model proposed by Devlin et al. stronger performance across different Natural Language Processing (NLP) tasks [14]. Feature extraction is done with a basic architecture model; BERT uses a Transformer encoder structure because it has 12 layers as shown in Figure 1, while most models handle this problem unidirectionally (left to right – or vice versa), then perform contrast. What makes BERT [19] unique is the fact that its training is bidirectional, meaning it looks at the left and right context of each word simultaneously to better capture the full context of a word at once.

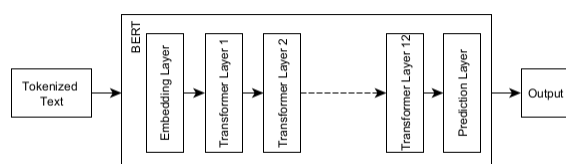


Figure 1. BERT Architecture

Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) are two unsupervised tasks used to pre-train the model on potentially large datasets [14]. MLM randomly replaces tokens in the input sequence with an [MASK] token and trains the model to predict these entries to allow the model to learn about word-level relationships. NSP, on the other hand, improves BERT's comprehension of relationships between sentences by predicting if a particular pair of sentences appear consecutively in the original text.

Multiple studies highlighted BERT's adaptability and performance in downstream NLP tasks. These are just a few examples of how BERT has been fine-tuned for downstream tasks such as sentiment analysis, question answering, named entity recognition, and text summarization where it provides state-of-the-art performance on many [15], [20] of the datasets. Hence,

two domain-specific versions were created, BioBERT and ClinicalBERT that was pre-trained on biomedical corpus lead by improved performance in domain-specific task [21], [22].

While BERT is very effective in NER, it has some drawbacks, namely that large models can be difficult to run and can be a barrier to implementation in resource-constrained environments. On top of this, its fixed-length input can make it somewhat limited when working with longer texts. In response, several distilled versions of these models have been proposed [23], as well as a few model compression approaches reducing inference time and memory consumption while minimally impacting performance [24].

DistilBERT is a smaller and faster version of BERT that was developed in the goal of being able to reduce computational requirements without losing too much language comprehension ability of the source model [23]. This method is called knowledge distillation, which was implemented in the process of creating DistilBERT introduced by Sanh et al., where a larger proposed “teacher” model trains a smaller “student” model to replicate behavior done by BERT [23]. Removing the token type embedding and pooling layers from BERT and reducing the number of layers from 12 to 6 (see Figure 2), yields a model that is 40% smaller, runs 60% faster while preserving ~97% of BERT performance when tested across a diverse range of NLP benchmarks [23], [25]. This results in good accuracy with better resources through triple loss functions (cosine distance loss + distillation loss + masked language modeling).

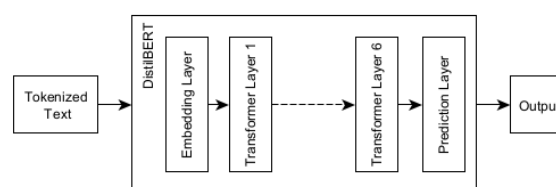


Figure 2. DistilBERT Architecture

Research demonstrated that DistilBERT was able to achieve state-of-the-art performance on other tasks such machine reading comprehension, named entity recognition and sentiment analysis, as well as tasks in a more real-world environment like the extraction of patient feedback about medical care [26], [27]. DistilBERT has also shown promise in forensic and low-resource use cases, where it achieves a good compromise between speed and accuracy [28]. Its smaller size and computational efficiency make it suitable for application in environments with limited processing, such as mobile devices or embedded systems but still achieves state-of-the-art NLP performance [23], [29].

RoBERTa (Robustly optimized BERT approach) is a state-of-the-art language representation model based on the architecture of BERT and aims to improve performance from changes in pretraining [15]. And it

specializes only in the Masked Language Modeling (MLM) tasks and removes Next Sentence Prediction (NSP), which was one of BERT's original training objectives proposed by Liu et al [15]. Also, bigger batch sizes and extended training periods entail a much larger dataset for training resulting in better contextual information of the language. In addition, RoBERTa improves the generalization of this model by using a dynamic masking mechanism for the input text during training rather than applying it statically [15].

Using these modifications, RoBERTa beats BERT on most general benchmark NLP tasks (GLUE, RACE and SQuAD) and showcases the model ability to perform well across multiple linguistic facilities [15], [30]. In addition, in domain-specific sentiment analysis [31], clinical natural language processing (NLP) [32] and legal document classification [33], it has been reported that the use RoBERTa outperformed previous models because it could better capture intricate wording relationships. Due to its scalability and enhanced performance, it has become the desired model in high-performance NLP tasks for researchers and practitioners like [15], [30].

Twitter-RoBERTa is a variant of the RoBERTa model that has been tuned to Twitter dataset linguistic and structural properties, such as hashtags, emoticons, informal phrases and abbreviations [34]. BERT, RoBERTa and other multi-purpose language models often underperform on tweet-based sentiment analysis tasks due to divergent vocabulary used in social media [15], [34]. To overcome this limitation, researchers demonstrated domain-specific models such as BERTweet that employs the pretraining framework of RoBERTa and is trained on a large corpus of 850 million English tweets [34]. Proven to achieve better performance than BERT and RoBERTa in various tweet tasks, including sentiment classification and emotion detection [15], [34], BERTweet has become one of the state-of-the-art models for Twitter data analysis.

Furthermore, Twitter-RoBERTa has been adapted to specific domains [29] such as the healthcare domain [30], where it was successfully used to detect sentiments in tweets regarding insomnia, thereby providing valuable insights into public health perspectives. These results illustrate the degree to which Twitter-specific fine-tuning can enhance model performance for real-time user-generated text input [34–37].

Existing studies, as summarized in Table 1, have demonstrated the effectiveness of transformer-based models for sentiment analysis. However, comprehensive comparative evaluations involving lightweight, general-purpose, and domain-specific transformer architectures on YouTube user comments remain limited. Most previous studies focus on benchmark datasets or Twitter data, evaluate only one or two transformer models, and rarely assess model performance under identical preprocessing, fine-tuning,

and evaluation settings. Furthermore, relatively few studies statistically validate performance differences using hypothesis testing such as McNemar's test. To address these gaps, this study presents a comprehensive comparison of DistilBERT, BERT, BERTweet, RoBERTa, and Twitter-RoBERTa for sentiment analysis on noisy YouTube user-generated content under a unified experimental framework.

Table 1. Summary of Related Works

Source	Model	Domain	Findings	Limitation
[14]	BERT	General NLP benchmarks	Introduced bi-directional transformer pretraining and achieved state-of-the-art results across multiple NLP tasks	Not specifically optimized for social media text
[23]	Distil BERT	General NLP tasks	Reduced model size and inference time while retaining most BERT performance	Slight reduction in contextual representation capability
[15]	RoBERTa	Large-scale corpora	Improved BERT training strategy and achieved superior performance on benchmark datasets	Requires large computational resources
[38]	BERTweet	Twitter social media corpus	Demonstrated strong effectiveness on social media NLP tasks due to Twitter-specific pretraining	Mainly optimized for Twitter-style language
[39]	Twitter-RoBERTa	Twitter sentiment corpus	Achieved high performance in sentiment analysis through domain-adapted pretraining	Limited evaluation on non-Twitter platforms
[40]	Multiple Transformer Models	Benchmark sentiment datasets	Demonstrated superiority of transformer-based models over traditional methods in sentiment analysis	Mostly benchmark datasets rather than platform-specific comments

2. Methods

This section describes the process of building a sentiment analysis model based on YouTube user reviews using various BERT-based models. Figure 3 shows how this study is structured into several phases for data collection, preprocessing, training with various BERT-based models, and evaluation.

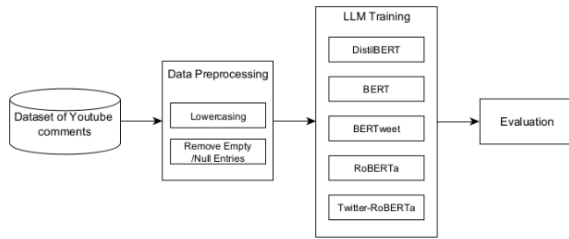


Figure 3. Research Workflow

2.1 Data Collection

The YouTube Comments Dataset available on Kaggle by atifaliak was selected because it provides a large collection of real-world user-generated comments collected from multiple YouTube videos, making it a suitable benchmark for sentiment analysis. The dataset contains information such as the author, comment text, publication time, like count, and sentiment scores (1–5), enabling sentiment classification as well as potential analyses of user engagement and temporal trends. More importantly, YouTube comments exhibit diverse linguistic characteristics, including informal expressions, slang, abbreviations, emojis, spelling variations, and conversational writing styles. These characteristics introduce substantial textual noise and contextual ambiguity, making sentiment classification significantly more challenging than conventional text classification tasks.

Unlike curated benchmark datasets, the comments reflect authentic user opinions generated in natural online interactions, providing a realistic environment for evaluating the robustness and generalization capability of transformer-based language models. The dataset also contains sufficient samples to support reliable fine-tuning and comparative evaluation under identical experimental settings. Although preprocessing is required to remove noisy entries and convert the original sentiment scores into sentiment classes, only minimal text cleaning was performed to preserve the contextual information that transformer models are designed to exploit. Consequently, this dataset provides an appropriate and scientifically meaningful benchmark for comparing the performance of general-purpose and domain-specific transformer models on real-world social media sentiment analysis.

2.2 Data Preprocessing

Proper data preparation is crucial for accurately following the positive and negative qualities of sound comments on YouTube. Dataset cleaning is process of cleaning the data as minimally as possible with basic pre-processing (mainly converting letters to lowercase),

removing empty/null entries, etc. This is very helpful in cleaning the dataset thereby improving the quality of the data and ensuring the model that needs to be trained runs smoothly.

Word2Vec-BERT: BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained NLP model that got rid of stopword removal, lemmatization, stemming and punctuation or special character removal. Because such aspects are preserved, the BERT can be performed with unprocessed text and remain contextually and semantically relevant. Over-cleaning the data can diminish BERT's understanding of nuance and structure. Consequently, minimal and thought-out pre-processing are recommended to maintain the correct direction of input and impact classification.

2.3 Fine Tuning Process

The selected hyperparameters for fine-tuning include 500 warm-up steps, three training epochs, a training batch size of 16, an evaluation batch size of 64, a weight decay of 0.01, and a learning rate of 5×10^{-5} . These settings were chosen to achieve a balance between predictive performance and robustness to variations in input patterns. A dynamic hyperparameter-tuning method was also employed to further optimize model performance.

All models were evaluated using accuracy, precision, recall, and F1-score. The results demonstrate that appropriate hyperparameter tuning improves the models' effectiveness and robustness, thereby enhancing their ability to perform sentiment classification on user-generated YouTube comments.

2.4 Evaluation Metrics

The evaluation of the performance output comes next, following the completion of data training and the production of performance output. Equations 1 – 4 show the accuracy, precision, recall, and f1-scores of the tested models.

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

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

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

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

The proposed evaluation procedure takes a YouTube comments dataset as input and produces classification metrics for five transformer models: BERT (M1), DistilBERT (M2), RoBERTa (M3), Twitter-RoBERTa (M4), and BERTweet (M5). For each model, the dataset is loaded from a CSV file, converted to lowercase, and cleaned by removing empty or null entries. It is then divided into training, validation, and test sets using a ratio of 70%, 15%, and 15%, respectively. The comments are tokenized and transformed into numerical representations using the tokenizer and

embedding mechanism associated with each pretrained model. The tokenized data are subsequently passed to the corresponding model for supervised fine-tuning. During this process, the model parameters are optimized using the AdamW optimizer with a learning rate of 0.00005, while cross-entropy is employed as the classification loss function. After training for a predefined number of epochs, the model is evaluated on the test set. Its performance is measured using accuracy, precision, recall, and F1-score, and the resulting metrics are reported separately for each model. The results and discussion sections will go into more specifics about this.

2.5 Experimental Design

Comments that are neutral, negative, or positive refer to the preprocessed dataset. It shows if the comments are classified as neutral, negative, or positive sentiment. Tables 2 and 3 provide an overview of these specifics.

Table 2. Dataset Description

Dataset	Positive	Negative	Neutral	Total
Youtube Comments	11,397	2,338	4,630	18,408

Table 3. Dataset Youtube Comments

Example	Score	Content
Ex1	1	"the fact he didnt buy the nicholas cage spoon is outrageous"
Ex2	3	"for now i need both apple pay and the physical credit card"
Ex3	5	"apple pay is so convenient secure and easy to use i used it while at the korean and japanese airports no need for physical credit cards"

Table 4. Training Parameters

Model Parameters	M1-M5
Epoch	3
Learning Rate	0.0005
Warmup_Steps	500
Weight Decay	0.01

The training parameters of the LLMs used were explicitly tuned for efficiency as seen in Table 4, and three epochs were used on the data to ensure fast convergence. A 500-step warm-up was employed to gradually adjust the learning rate, which is set to the initial value of 0.00005. Weight decay of 0.01 is used to prevent overfitting. The dataset was split for testing, validation, and training to improve model building steps.

Table 5. Ratio of Split Dataset

Dataset	Train (70%)	Test (15%)	Validation (15%)	Total
Youtube Comments	12,886	2,761	2,761	18,408

To validate and reproducibility of our experiments this research used 70:15:15 ratio to divide dataset into train, validation and test sets as shown in Table 5. Dataset is split before the model training process happened, post data acquisition and preprocessing stages. Training set was used to learn the model while validation and testing sets were kept aside for performance evaluation and

generalization assessment. Next, the training dataset was used to fine-tune each Large Language Model (LLM) individually. Thirdly, for each LLM-based sentiment analysis model, the performance will be compared and judged based on fine-tuning results.

Performance is assessed by evaluating LLM. Preprocessing includes lowercasing and removing empty or null entries. Anaconda software is used to run experiments on a computer with hardware specifications of Intel i9 14900K CPU, NVIDIA GeForce RTX 4060 GPU, and 32GB RAM.

3. Results and Discussions

3.1 Result

Table 6 compares the five language models—DistilBERT, BERT, BERTweet, RoBERTa, and Twitter-RoBERTa—using four essential performance metrics: accuracy, precision (P), recall (R), and F1 score. Twitter-RoBERTa achieved the best performance with an accuracy of 90.13%, precision of 89.96%, recall of 90.13% and F1 score of 90.01%. This shows that Twitter-RoBERTa is the best model for sentiment classification tasks primarily because it has been trained in social media data, which makes it greatly suitable to analyze informal and context-rich user-generated content from YouTube comments. Both BERTweet and RoBERTa performed very strongly, with such that RoBERTa slightly outperformed BERTweet on all categories. For example, BERTweet which was built directly from Twitter data, has shown improvement over general BERT because of the advantages gained from domain-specific pre-training. The important thing to note is that DistilBERT, although the lowest of the five, nevertheless has a high score for situations where resources are very limited. In summary, the results show that pretrained transformer models on social media platforms perform well for sentiment classification of online user comments.

Table 6. LLM Performance Comparison

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
DistilBERT	86.48	84.72	84.68	84.7
BERT	86.28	86.13	86.28	86.2
BERTweet	88.09	87.98	88.09	88
RoBERTa	88.68	88.51	88.68	88.56
Twitter-RoBERTa	90.13	89.96	90.13	90.01

The McNemar test results in Figure 4 indicate that several transformer-based models exhibit statistically significant differences in classification behavior when evaluated on the same YouTube sentiment dataset. Twitter-RoBERTa also always outperformed DistilBERT and BERT across all models. The presence of statistically significant P values, such as 1.95×10^{-10} and 1.48×10^{-9} for these variables, respectively, indicates that these values are indeed much lower than our set significance level (0.05). The results show that Twitter-RoBERTa made significantly fewer classification mistakes, indicating that domain-specific

pretraining on social media text allows the model to better capture the informal linguistic cues typical of YouTube comments.

PAIRWISE MCNEMAR TEST RESULTS				
	model_a	model_b	n_both_correct	a_correct_b_wrong \
2	DistilBERT	Twitter-RoBERTa	2231	123
5	BERT	Twitter-RoBERTa	2259	108
1	DistilBERT	RoBERTa	2233	121
4	BERT	RoBERTa	2260	107
3	DistilBERT	BERTweet	2219	135
6	BERT	BERTweet	2252	115
9	Twitter-RoBERTa	BERTweet	2325	151
7	RoBERTa	Twitter-RoBERTa	2359	90
8	RoBERTa	BERTweet	2303	146
0	DistilBERT	BERT	2230	124

	a_wrong_b_correct	n_both_wrong	statistic	p_value \
2	245	156	123.0	1.949741e-10
5	217	171	108.0	1.479158e-09
1	216	185	121.0	2.558409e-07
4	189	199	107.0	2.170414e-06
3	208	193	135.0	9.572421e-05
6	175	213	115.0	5.094835e-04
9	102	177	102.0	2.478969e-03
7	117	189	90.0	7.048535e-02
8	124	182	124.0	2.011550e-01
0	137	264	124.0	4.576792e-01

	method	significant_at_0.05
2	statsmodels exact McNemar	True
5	statsmodels exact McNemar	True
1	statsmodels exact McNemar	True
4	statsmodels exact McNemar	True
3	statsmodels exact McNemar	True
6	statsmodels exact McNemar	True
9	statsmodels exact McNemar	True
7	statsmodels exact McNemar	False
8	statsmodels exact McNemar	False
0	statsmodels exact McNemar	False

Figure 4. McNemar Statistical Test Result

In the same way, RoBERTa performed best among both DistilBERT and BERT with p-values of 2.56×10^{-7} and 2.17×10^{-6} respectively. These results show improvements in transformer training strategies and optimizers that benefit sentiment classification, particularly within a noisy user-generated context. Additionally, BERTweet also outperformed DistilBERT and BERT by a statistically significant margin as well, emphasizing the need for social-media specialized pretraining over general-purpose pretraining for sentiment classification tasks.

By contrast, comparisons of the top performing models showed no statistically significant differences. As an example, the comparison of RoBERTa and Twitter-RoBERTa gave a p-value = 0.070 and when this research compare RoBERTa to BERTweet, the calculated p-value is equal to 0.201 Because they are greater than 0.05, it cannot be concluded that the prediction errors between these models significantly correlate with each other and therefore the null hypothesis is accepted. Although Twitter-RoBERTa produced the highest average performance overall, RoBERTa and BERTweet were competitive with one another on the two sections of resources from which they were trained and exhibited relatively similar classification behavior on this dataset.

Similarly, the comparison between DistilBERT and BERT yielded a p-value of 0.458 (no significant difference between models). Even though BERT is larger and more computationally intensive, it did not perform significantly better than the lightweight DistilBERT model on that task. This result means that this research can provide another option instead of

BERT for projects which are related to sentiment analysis and computational efficiency is important.

In summary, the McNemar analysis shows that domain-adapted transformer models yield significant benefits for noisy text sentiment detection in social media (in particular, Twitter-RoBERTa and BERTweet) over generic pre-trained models. The results further show that the effectiveness of more sophisticated transformer training strategies and social media-specific pretraining outweighs model size.

3.2 Discussion

Of all three applications, RoBERTa and BERTweet performed well but Twitter-RoBERTa achieved highest across all evaluation metrics. This behavior was anticipated and is appropriately aligned to results previously reported showing that training transformer models on domain-specific datasets, such as Twitter corpora in this case, boosts sentiment analysis performance dramatically over informal social media text data. This explains the excellent performance this research sees with BERTweet, which is simply a version of BERT pretrained on Twitter rather than on general-domain corpora like the original BERT variants. On the other hand, while DistilBERT offered lightweight architecture and thus computational efficiency and faster inference time, its performance was comparatively lower, exemplifying the trade-off between computational efficiency (DistilBERT's efficiency) and contextual representation capability (BERT).

These findings are reinforced by the statistical analysis based on the McNemar test. The difference in predictions performance between Twitter-RoBERTa and DistilBERT yielded a very small p-value ($p < 0.05$). Similar significant differences were also found between Twitter-RoBERTa and BERT, as well as for RoBERTa and BERTweet compared to DistilBERT and BERT. The classification errors made on this set further confirm that the domain-adapted transformer models consistently give less errors over noisy YouTube comments. When this research compared the top models (RoBERTa, BERTweet and Twitter-RoBERTa), none of the differences were significant, indicating that there is a tendency for these models to behave similarly in classification even though they had slightly different overall accuracy and F1-score.

Additionally, hybrid architecture is also noteworthy due to their improved generalization capabilities, as demonstrated by the success of Twitter-RoBERTa-based approaches. The authors of TRABSA [31] proposed a hybrid framework that combines Twitter-RoBERTa with an attention-based BiLSTM architecture for sentiment analysis. The model achieved high classification accuracy while also incorporating explainability techniques such as SHAP and LIME to improve interpretability. The attention mechanism enables the model to focus on the most relevant parts of

the input text, thereby improving both prediction accuracy and model transparency.

Additionally, the Dual Joint Classifier (DJC) [32] presented a joint architecture based on a combination of Twitter-RoBERTa and BERTweet for Twitter sentiment classification. Experiments on Twitter US Airline Sentiment and Apple Twitter Sentiment benchmark datasets showed a significant accuracy and F1-score improvement of the ensemble-based system trained with transformer models specialized for each domain area.

In conclusion, the results from our experiment and the previous studies suggest that transformer models pre-trained on social media corporation deliver good performance for sentiment analysis tasks of user-generated content. As a result, Twitter-RoBERTa consistently outperformed the other models, likely due to its ability to understand social media-specific linguistic patterns and contextual intricacies. Finally, the McNemar statistical analysis demonstrates that improvements obtained by domain-adapted models are quantitatively greater and statistically significant. Such results support the conclusion: that domain-specialised pretraining, with hybrid architecture and attention-based mechanisms, provides greater robustness, interpretability and even superior solutions in many practical NLP-based sentiment analysis tasks.

4. Conclusions

Five transformer-based models were investigated in this paper in the context of sentiment analysis on user-generated YouTube comments, such as DistilBERT, BERT, BERTweet, RoBERTa and Twitter-RoBERTa. Twitter-RoBERTa works best overall of all models considered, with Accuracy = 90.13% and Precision = 89.96%, Recall = 90.13% and F1 Score = 90.01%. Experiments show that domain-specific pretraining markedly boosts performance on sentiment classification, especially for noisily written and informal social media text where contextual embeddings and conversational language structure matter.

These results further highlight the need to distinguish between general-purpose and domain-specific transformer architectures for sentiment analysis problems. Models that were pretrained on social media corpora, such as Twitter-RoBERTa and BERTweet, consistently outperformed more general transformer models, as their capacity for adapting to the linguistic characteristics of user-generated content facilitated performance. On the other hand, whereas DistilBERT provides computational efficiency and a smaller model size, it demonstrates lower classification performance indicating a trade-off between efficiency versus contextual representation ability.

In additional analyses, this research used McNemar' test to demonstrate that differences in performance were statistically significant between some of the evaluated

models. The predictions of Twitter-RoBERTa outperformed those of DistilBERT and BERT in a statistically significant manner, suggesting that the improvements reached by domain-adapted transformer models are both numerically stronger and statistically sound. Nonetheless, models are very closely comparable for the top three performing models (RoBERTa, BERTweet and Twitter-RoBERTa) with only marginal evaluation metric differences; statistical sensitivity analysis showed no treatment effect difference but similar classification behavior.

The research also opens avenues for future work, especially in real-time monitoring systems incorporating transformer-based sentiment analysis models and their applications in multilingual and cross-platform social media environments. Moreover, upcoming research could investigate hybrid architectures and explanations (e.g., SHAP, LIME) to provide aspects of interpretability, transparency, and trust in such models to end users across applications with high-stakes NLP components.

Declarations

AI Usage Statement

The authors used Quilbot solely to improve the readability and language of the manuscript; all scientific content, analysis, and conclusions are the authors' own.

Author Contributions

Agus Cahyo Nugroho: conceptualization, methodology, writing – original draft, data curation, formal analysis; Li-Hua Li: writing – review and editing.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest Statement

The authors declare no conflict of interest.

Ethical Clearance

This study did not involve human participants or animal subjects. It used a publicly available dataset of YouTube comments; therefore, formal ethical approval was not required.

Data and Code Availability

The dataset used in this study is publicly available through <https://www.kaggle.com/datasets/atifaliak/youtube-comments-dataset>. The source code used in this study is not publicly available but may be provided by the corresponding author upon reasonable request.

References

- [1] D. Drašković and S. Milanović, "Aspect-based sentiment analysis of user-generated content from a microblogging

- platform,” *Journal of Big Data*, vol. 12, no. 1, p. 186, 2025, <https://doi.org/10.1186/s40537-025-01244-0>.
- [2] N. M. Gardazi, A. Daud, M. K. Malik, A. Bukhari, T. Alsahfi, and B. Alshemaimri, “BERT applications in natural language processing: a review,” *Artif. Intell. Rev.*, vol. 58, no. 6, pp. 1–49, 2025, <https://doi.org/10.1007/s10462-025-11162-5>.
 - [3] C. Liu, A. Arulappan, R. Naha, A. Mahanti, J. Kamruzzaman, and I. H. Ra, “Large language models and sentiment analysis in financial markets: A review, datasets, and case study,” *IEEE Access*, vol. 12, pp. 134041–134061, 2024, <https://doi.org/10.1109/ACCESS.2024.3445413>.
 - [4] A. Alamsyah and Y. Sagama, “Empowering Indonesian internet users: An approach to counter online toxicity and enhance digital well-being,” *Intell. Syst. Appl.*, vol. 22, p. 200394, 2024, <https://doi.org/10.1016/j.iswa.2024.200394>.
 - [5] G. Fenza, V. Loia, C. Stanzione, and M. Di Gisi, “Robustness of models addressing Information Disorder: A comprehensive review and benchmarking study,” *Neurocomputing*, p. 127951, 2024, <https://doi.org/10.1016/j.neucom.2024.127951>.
 - [6] B. Liu, “Sentiment analysis: A fascinating problem,” in *Sentiment Analysis and Opinion Mining*, Cham, Switzerland: Springer International Publishing, 2022, pp. 1–8, https://doi.org/10.1007/978-3-031-02145-9_1.
 - [7] M. Kastrati, Z. Kastrati, A. Shariq Imran, and M. Biba, “Leveraging distant supervision and deep learning for Twitter sentiment and emotion classification,” *Journal of Intelligent Information Systems*, vol. 62, no. 4, pp. 1045–1070, 2024, <https://doi.org/10.1007/s10844-024-00845-0>.
 - [8] S. Yuan, Z. Zou, H. Zhou, Y. Ren, J. Wu, and K. Yan, “SECNN: Squeeze-and-excitation convolutional neural network for sentence classification,” *IEEE Access*, vol. 13, pp. 42858–42865, 2025, <https://ieeexplore.ieee.org/document/11070808>.
 - [9] Q. Guo, Z. He, and Z. Wang, “Assessing the effectiveness of long short-term memory and artificial neural network in predicting daily ozone concentrations in Liao Cheng City,” *Scientific Reports*, vol. 15, no. 1, p. 6798, 2025, <https://doi.org/10.1038/s41598-025-91329-w>.
 - [10] M. Kowsher, A. A. Sami, N. J. Prottasha, M. S. Arefin, P. K. Dhar, and T. Koshiba, “Bangla-BERT: Transformer-based efficient model for transfer learning and language understanding,” *IEEE Access*, vol. 10, pp. 91855–91870, 2022, <https://doi.org/10.1109/ACCESS.2022.3197662>.
 - [11] A. Benayas, M. A. Sicilia, and M. Mora-Cantallos, “A comparative analysis of encoder-only and decoder-only models in intent classification and sentiment analysis: Navigating the trade-offs in model size and performance,” *Language Resources and Evaluation*, vol. 59, no. 3, pp. 2007–2030, 2025, <https://doi.org/10.1007/s10579-024-09796-y>.
 - [12] K. Du, F. Xing, R. Mao, and E. Cambria, “Financial sentiment analysis: Techniques and applications,” *ACM Computing Surveys*, vol. 56, no. 9, pp. 1–42, 2024, <https://doi.org/10.1145/3649451>.
 - [13] J. Liang, L. Cao, S. Madden, Z. Ives, and G. Li, “RITA: Group attention is all you need for timeseries analytics,” *Proceedings of the ACM on Management of Data*, vol. 2, no. 1, pp. 1–28, 2024, <https://doi.org/10.1145/3639317>.
 - [14] B. Tahayna and R. K. Ayyasamy, “Applying English idiomatic expressions to classify deep sentiments in COVID-19 tweets,” *Comput. Syst. Sci. Eng.*, vol. 47, no. 1, 2023, <https://doi.org/10.32604/csse.2023.036648>.
 - [15] 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 *Proc. 2022 3rd Int. Conf. Artificial Intelligence and Data Sciences (AiDAS)*, 2022, pp. 1–6, IEEE, <https://doi.org/10.1109/AiDAS56890.2022.9918678>.
 - [16] W. M. N. W. B. Manike, R. A. Wijesinghe, and A. M. T. N. Adasuriya, “Optimizing DistilBERT for low-resource sentiment analysis: A study on global unstructured L1-norm pruning and accuracy-efficiency trade-offs,” in *Proc. 2026 6th Int. Conf. Advanced Research in Computing (ICARC)*, 2026, pp. 1–6, IEEE, <https://doi.org/10.1109/ICARC68737.2026.11453647>.
 - [17] H. Su, J. Kasai, C. H. Wu, W. Shi, T. Wang, J. Xin, et al., “Selective annotation makes language models better few-shot learners,” *arXiv preprint arXiv:2209.01975*, 2022, <https://doi.org/10.48550/arXiv.2209.01975>.
 - [18] L. Gorenstein, E. Konen, M. Green, and E. Klang, “Bidirectional encoder representations from transformers in radiology: A systematic review of natural language processing applications,” *Journal of the American College of Radiology*, vol. 21, no. 6, pp. 914–941, 2024, <https://doi.org/10.1016/j.jacr.2024.01.012>.
 - [19] M. S. Islam, L. Xiangdong, and J. Ahmed, “BERT: Advancements in language understanding for different NLP tasks: Challenges and future perspectives,” *Journal of Electrical Systems and Information Technology*, vol. 13, no. 1, Art. no. 49, 2026, <https://doi.org/10.1186/s43067-026-00341-1>.
 - [20] B. Li, C. Li, J. Sun, X. Zeng, X. Chen, and J. Zheng, “ERNIE-UIE: Advancing information extraction in Chinese medical knowledge graph,” *PLOS ONE*, vol. 20, no. 5, p. e0325082, 2025, <https://doi.org/10.1371/journal.pone.0325082>.
 - [21] B. Wang, Q. Xie, J. Pei, Z. Chen, P. Tiwari, Z. Li, and J. Fu, “Pre-trained language models in biomedical domain: A systematic survey,” *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–52, 2023, <https://doi.org/10.1145/3611651>.
 - [22] D. Vithanage, P. Yu, L. Wang, and C. Deng, “Contextual word embedding for biomedical knowledge extraction: a rapid review and case study,” *Journal of Healthcare Informatics Research*, vol. 8, no. 1, pp. 158–179, 2024, <https://doi.org/10.1007/s41666-023-00157-y>.
 - [23] Y. Z. Vakili, A. Fallah, and H. Sajedi, “Distilled BERT model in natural language processing,” in *Proc. 2024 14th Int. Conf. Computer and Knowledge Engineering (ICCKE)*, 2024, pp. 243–250, IEEE, <https://doi.org/10.1109/ICCKE65377.2024.10874673>.
 - [24] M. Soyeb, M. Mahdin, S. S. Sadaf, M. M. Azwad, and A. Abrar, “TYMBERT: Tiny yet mighty—fine-tuned and compressed TinyBERT for high-accuracy, low-resource NLP,” in *Proc. 2025 2nd Int. Conf. Next-Generation Computing, IoT and Machine Learning (NCIM)*, 2025, pp. 1–6, IEEE, <https://doi.org/10.1109/NCIM65934.2025.11160230>.
 - [25] G. Tucudean, M. Bucos, B. Dragulescu, and C. D. Căleanu, “Natural language processing with transformers: a review,” *PeerJ Computer Science*, vol. 10, p. e2222, 2024, <https://doi.org/10.7717/peerj-cs.2222>.
 - [26] A. B. Alawi and F. Bozkurt, “A hybrid machine learning model for sentiment analysis and satisfaction assessment with Turkish universities using Twitter data,” *Decision Analytics Journal*, vol. 11, p. 100473, 2024, <https://doi.org/10.1016/j.dajour.2024.100473>.
 - [27] S. S. Tiwari, H. R. Goyal, A. Shrivastava, K. Yadav, and A. L. N. Rao, “Clinical text analysis with natural language processing: A BERT-based approach,” in *Proc. 2024 Int. Conf. Communication, Computer Sciences and Engineering (IC3SE)*, 2024, pp. 1382–1388, IEEE, <https://doi.org/10.1109/IC3SE62002.2024.10593504>.
 - [28] N. Kumar, P. N. Bhat, B. R. Pushpa, and N. R. Sreekumar, “Hate speech detection and analysis using NLP: A hybrid DB-BiLSTM approach,” in *Proc. 2025 Control Instrumentation System Conference (CISCON)*, 2025, pp. 1–6, IEEE, <https://doi.org/10.1109/CISCON66933.2025.11337424>.
 - [29] S. Zhu, S. Gu, S. Li, L. Xu, and D. Xiong, “Mining parallel sentences from internet with multi-view knowledge distillation for low-resource language pairs,” *Knowledge and Information Systems*, vol. 66, no. 1, pp. 187–209, 2024, <https://doi.org/10.1007/s10115-023-01925-3>.
 - [30] S. Chen, Y. Zhang, and Q. Yang, “Multi-task learning in natural language processing: An overview,” *ACM Computing Surveys*, vol. 56, no. 12, pp. 1–32, 2024, <https://doi.org/10.1145/3663363>.
 - [31] K. L. Tan, C. P. Lee, K. M. Lim, and K. S. M. Anbananthen, “Sentiment analysis with ensemble hybrid deep learning model,” *IEEE Access*, vol. 10, pp. 103694–103704, 2022, <https://doi.org/10.1109/ACCESS.2022.3210182>.
 - [32] M. AnsariShiri and N. Mansouri, “A comparative evaluation of transformer models for medical abstract classification,” *Multimedia Tools and Applications*, vol. 85, no. 3, p. 183, 2026, <https://doi.org/10.1007/s11042-026-21140-w>.

- [33] M. Siino, M. Falco, D. Croce, and P. Rosso, "Exploring LLMs applications in law: A literature review on current legal NLP approaches," *IEEE Access*, vol. 13, pp. 18253–18276, 2025, <https://doi.org/10.1109/ACCESS.2025.3533217>.
- [34] F. Carneiro, D. Vianna, J. Carvalho, A. Plastino, and A. Paes, "BERTweet.BR: A pre-trained language model for tweets in Portuguese," *Neural Computing and Applications*, vol. 37, no. 6, pp. 4363–4385, 2025, <https://doi.org/10.1007/s00521-024-10711-3>.
- [35] A. Maghsoudi et al., "Sentiment Analysis of Insomnia-Related Tweets via a Combination of Transformers Using Dempster-Shafer Theory: Pre- and Peri-COVID-19 Pandemic Retrospective Study," *J. Med. Internet Res.*, vol. 24, no. 12, p. e41517, Dec. 2022, <https://doi.org/10.2196/41517>.
- [36] M. A. Jahin et al., "A hybrid transformer and attention based recurrent neural network for robust and interpretable sentiment analysis of tweets," *Scientific Reports*, vol. 14, no. 1, p. 24882, 2024, <https://doi.org/10.1038/s41598-024-76079-5>.
- [37] L. He, "Enhanced twitter sentiment analysis with dual joint classifier integrating RoBERTa and BERT architectures," *Frontiers in Physics*, vol. 12, p. 1477714, 2024, [doi: 10.3389/fphy.2024.1477714](https://doi.org/10.3389/fphy.2024.1477714).
- [38] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "BERTweet: A pre-trained language model for English Tweets," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Oct. 2020, pp. 9–14, <https://doi.org/10.18653/v1/2020.emnlp-demos.2>.
- [39] F. Barbieri et al., "TweetEval: Unified benchmark and comparative evaluation for tweet classification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, <https://doi.org/10.18653/v1/2020.findings-emnlp.148>.
- [40] S. Minaee, N. Kalchbrenner, Erik Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning-based text classification: A comprehensive review," *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp. 1–40, 2021, <https://doi.org/10.1145/3439726>.