Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5836

Submitted : April 27, 2025
Accepted : May 8, 2025
Published : May 10, 2025

Sentiment Analysis on Short Social Media Texts Using DistilBERT

Muhammad Sidik Asyaky[1)*], Muhammad Al-Husaini[2)], Hen Hen Lukmana[3)]

[1,2,3)]Informatics, Siliwangi University, Indonesia
[1)]asyaky@unsil.ac.id, [2)]alhusaini@unsil.ac.id, [3)]henhenlukmana@unsil.ac.id

ABSTRACT

Sentiment analysis on short texts from social media, such as tweets, presents unique challenges due to their brevity and informal language. This study explores the effectiveness of transformer-based models, particularly DistilBERT, in performing sentiment analysis on short texts compared to traditional machine learning approaches including Support Vector Machine, Logistic Regression, and Naive Bayes. The objective is to assess whether DistilBERT not only enhances sentiment classification accuracy but also remains efficient enough for quick social media analysis. The models used in this study were trained and evaluated on stratified samples of 10,000, 30,000, and 50,000 tweets, drawn from the Sentiment140 dataset while preserving the original class distribution. The methodology involved data collection and sampling, data splitting, data cleaning, feature extraction, model training, and evaluation using accuracy and F1-score. Experimental results showed that DistilBERT consistently outperformed traditional models in both accuracy and F1-score, and demonstrated competitive results against BERT while requiring significantly less training time. Specifically, DistilBERT trained approximately 1.8 times faster than BERT on average, highlighting its computational efficiency. The best result was achieved by DistilBERT trained on the 50k subset, reaching an accuracy of 85% and an F1-score of 84%. These findings suggest that lightweight transformer models like DistilBERT are highly suitable for real-world sentiment analysis tasks where both speed and performance are critical.

Keywords: DistilBERT; Performance Evaluation; Sentiment Analysis; Short Informal Text; Social Media; Tweet;

INTRODUCTION

The rise of social media platforms such as X has transformed them into major sources of real time information, where millions of users share their opinions, often in short-text form. This surge in user-generated content has led to a vast amount of short-text data being produced daily, capturing public sentiment on a wide range of topics, from politics and business to entertainment (Knittel et al., 2021). Uncovering meaningful patterns within the data can provide valuable insights into public opinion and emerging trends. As demonstrated by (Abei et al., 2025), who analyzed public sentiment on X toward the 2024 Jakarta Governor candidates, classifying tweets into positive and negative sentiments. It concluded that positive sentiments were largely linked to candidates' strong track records and proposed programs that addressed community needs. These insights can inform political communication strategies and highlight the role of social media in democratic engagement. Similarly, (Jusli & Kurniawan, 2024) examined public reactions to Indonesia's electric vehicle tax policy, identifying sentiment polarity and key concerns influencing public opinion. In another study, (Fransiscus & Girsang, 2022) investigated sentiment toward the government's COVID-19 public activity restrictions, revealing public attitudes and reactions during the pandemic. The findings indicated that public sentiment toward public activity restriction in Jakarta was predominantly negative, driven by economic disruptions. These studies underscore the critical role of sentiment analysis in unlocking real time insights into public opinion which enables businesses, policymakers, or researchers to better understand social trends and respond effectively.

Performing sentiment analysis on short text presents unique challenges due to its limited length and lack of context, which reduce semantic cues and increase ambiguity (Barreto et al., 2022). Short texts obtained from social media also often contain slang and informal expressions, adding another layer of difficulty to accurately interpret sentiment (Khan et al., 2025). To address these challenges, many studies have explored various approaches, including the use of machine learning and deep learning models (Tan et al., 2023). The study also points out that while traditional machine learning algorithms such as Naive Bayes (NB), Logistic Regressions (LR) and Support Vector Machines (SVM)

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5836

Submitted : April 27, 2025
Accepted : May 8, 2025
Published : May 10, 2025

combined with Term Frequency-Inverse Document Frequency (TF-IDF) have been commonly employed, they often fall short in detecting sentiment in ambiguous and sarcastic texts. Effective sentiment analysis in short texts requires understanding how words are related to each other, not just their individual occurrences (Mao et al., 2024). As a result, recent research has increasingly focused on deep learning approaches, particularly those that use transformer architectures such as Bidirectional Encoder Representations from Transformers (BERT), which have shown superior performance in understanding the intricacies of short texts (Jim et al., 2024; Karl & Scherp, 2023). Deep learning has also proven effective in other domains, such as visual assistance, where a CNN-based system using TensorFlow was developed to support individuals with Color Vision Deficiency (CVD), achieving high classification accuracy and offering practical benefits (Al-Husaini et al., 2024). These advancements highlight the versatility of deep learning architectures across textual and visual tasks, reinforcing their significance in the development of intelligent applications.

Unlike traditional machine learning models that rely on word frequency or handcrafted features, BERT is a pretrained language model that captures the context of words by reading text bidirectionally using self-attention mechanism. This enables BERT to generate context-aware representations of text (Loureiro et al., 2021), which helps it better understand the nuances in short and informal texts where context is often ambiguous and difficult to infer. However, while BERT delivers impressive accuracy, it's computationally demanding (Sarkar et al., 2024), making it less practical for large-scale real-time analysis. To mitigate this issue, a BERT variant named DistilBERT has been developed to retain most of BERT's performance while being significantly faster and more efficient (Sanh et al., 2019). It leverages a compression technique known as knowledge distillation to reduce model size and training complexity, making it a faster and more practical solution for sentiment analysis, where quick and accurate results are essential.

This research aims to evaluate the effectiveness of DistilBERT in sentiment analysis on short texts by comparing its performance against other models, including SVM, NB, LR, and BERT. The Sentiment140 dataset, which contains millions of tweets annotated with either positive or negative sentiment labels, is used to train and evaluate the models. Through this study, we assess whether DistilBERT not only enhances sentiment classification accuracy but also remains efficient enough for quick social media analysis.

METHOD

This study follows a structured sentiment analysis pipeline to evaluate the performance of DistilBERT in comparison to traditional machine learning models in sentiment analysis on short texts, as commonly outlined in (Palomino & Aider, 2022; Putra et al., 2024). The methodology consists of several stages, starting from data collection and sampling, followed by data splitting, text preprocessing, feature extraction, and concluding with model training and evaluation, as presented in Fig. 1.

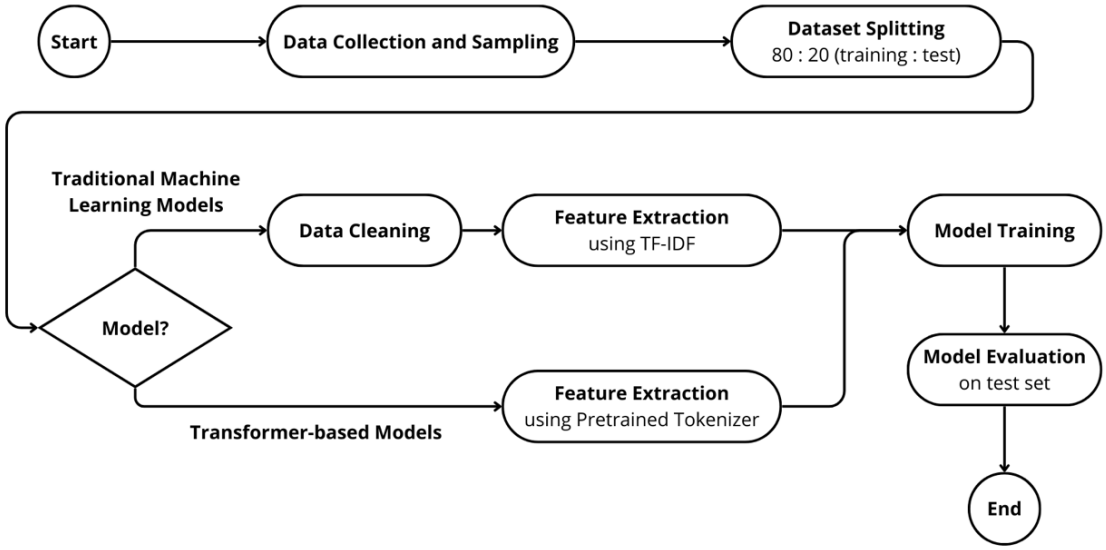


Fig. 1 Research Methodology

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5836

Submitted : April 27, 2025
Accepted : May 8, 2025
Published : May 10, 2025

Data Collection and Sampling

The data used in this study is sourced from the Sentiment140 dataset, a widely used benchmark for sentiment analysis tasks, which is publicly available to download from Kaggle (Kazanova, 2017). This dataset contains millions of short texts called tweets, annotated with either positive or negative sentiment labels. To analyze the models' performance across different dataset sizes and evaluate their scalability, the dataset is sampled into three subsets: 10,000, 30,000, and 50,000 tweets. Each sample subset is carefully selected using stratified sampling to maintain a balanced distribution of sentiment labels across all subsets, which is essential to avoid bias. This sampling approach enables the study to assess how model performance varies with different amounts of training data, a key consideration in real-world scenarios where data availability may be limited (Nguyen et al., 2021).

Data Splitting

Each sampled subset is divided into training and testing sets using an 80:20 ratio, with an additional 10% of the training portion set aside as a validation set exclusively for transformer-based models. Each resulting set maintains a balanced distribution of sentiment classes. All models are evaluated on the same test set within each sampled subset to ensure a fair comparison of performance.

Data Cleaning

To prepare the data, a text preprocessing phase is conducted. Two distinct preprocessing strategies are applied to accommodate the differing needs of the models. For traditional machine learning models, the preprocessing involves removing emojis, symbols, URLs, and usernames, converting text to lowercase, removing stop words, and applying lemmatization to reduce words to their base forms. In contrast, minimal preprocessing is used for transformer-based models. Emojis, URLs, and usernames are removed, while the rest of the text remains unchanged to preserve the natural language structure, as transformer-based models rely on contextual information and pretrained language patterns that can be disrupted by overly aggressive text normalization (Siino et al., 2024).

Feature Extraction

In this study, two different feature extraction methods are employed based on the type of model used. For traditional machine learning models, feature extraction is performed using the TF-IDF vectorizer. It transforms text into numerical vectors by measuring the importance of each term within a document relative to its occurrence across the entire corpus. This allows the models to identify patterns in term usage that correlate with sentiment labels. The TF-IDF score for a term t in a document d is computed as:

$$TF - IDF(t, d) = TF(t, d)\ x\ IDF(t) \qquad (1)$$

The Term Frequency (TF) measures how often a term appears within a single document, while the Inverse Document Frequency (IDF) evaluates how distinctive or rare that term is across the entire corpus.

As for transformer-based models, feature extraction is handled by pretrained tokenizer. It converts raw text into a numerical representation that the model can understand. The tokenizer breaks the text into subword units and maps them to unique identifiers from the model's vocabulary. These inputs are passed through embedding layers to generate contextualized vector representations that capture both the meaning of individual words and their relationships within the sentence. This approach allows transformer models to leverage deep contextual understanding learned from large scale pretraining (Devlin et al., 2018).

Model Training

Each model was trained using the same training sets for each of the three sampled subsets. Traditional machine learning models (NB, LR, and SVM) were trained using the TF-IDF feature vectors, while transformer-based models (DistilBERT and BERT) were fine-tuned using their respective tokenized inputs.

The same training process is applied to both DistilBERT and BERT models, the only difference lies in the specific pretrained model used: **distilbert-base-uncased** and **bert-base-uncased** for DistilBERT and BERT respectively. Both models are fine-tuned using Hugging Face's Trainer API with identical training configurations. Each model is trained for 3 epochs with a batch size of 16 for both training and evaluation. The training set is used to update model weights, while the validation set is used at the end of each epoch to monitor performance and detect overfitting. The best model

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5836

Submitted : April 27, 2025
Accepted : May 8, 2025
Published : May 10, 2025

is automatically reloaded based on validation performance. To enhance generalization and prevent overfitting, a weight decay of 0.01 is applied.

As for the traditional machine learning models, we employ MultinomialNB, LinearSVC, and LogisticRegression from the Scikit-learn library, corresponding to NB, SVM, and LR respectively. These models are trained on TF-IDF feature vectors derived from the training set.

Model Evaluation

Each classifier, resulting from model training, is evaluated on the same test set to ensure a fair comparison. The evaluation uses accuracy and F1-score as the performance metrics. Accuracy reflects the overall correctness of predictions, while F1-score provides a balanced measure by combining precision and recall. Precision refers to the proportion of correct positive predictions, and recall measures the model's ability to correctly identify all true positives, both of which contribute to the effectiveness captured by the F1-score. The formula for the accuracy and F1-score is as follows:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$precision = \frac{TP}{TP+FP} \quad (3)$$

$$recall = \frac{TP}{TP+FN} \quad (4)$$

$$F1 = \frac{2\ x\ precision\ x\ recall}{precision+recall} \quad (5)$$

where TP is true positives, FP is false positives and FN is false negatives. Greater accuracy and F1-score indicate better model performance, meaning the model is not only making more correct predictions but is also reliably identifying positive instances while minimizing false positives and false negatives. In addition to these metrics, we also measure the training time for each model to assess the trade-off between performance and computational efficiency.

RESULT

The purpose of this research is to evaluate the effectiveness of DistilBERT for short-text sentiment classification by comparing its accuracy, F1-score, and training efficiency against BERT and traditional machine learning models. The results are thoroughly analyzed to address the research objectives and assess the effectiveness of the applied methods.

Based on the steps described in the Method section, we present the results in two parts: data preparation, model training and evaluation.

Data Preparation

The Sentiment140 dataset contains 1.6 million labeled tweets, annotated as either positive or negative. The sentiment labels are evenly distributed, with 800,000 positive and 800,000 negative tweets as illustrated in Fig. 2.

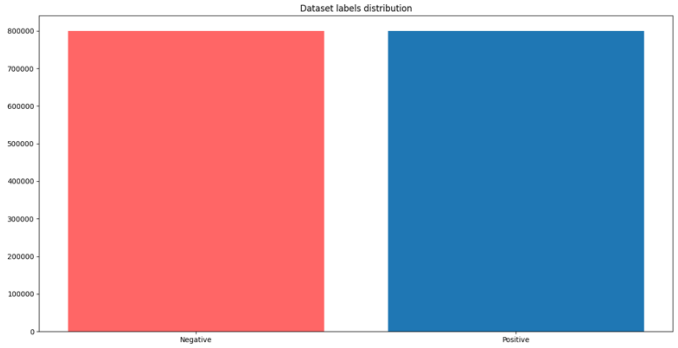


Fig. 2 Label Distribution of 1.6 Million Tweets in the Sentiment140 Dataset

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5836

Submitted : April 27, 2025
Accepted : May 8, 2025
Published : May 10, 2025

Each entry in the dataset includes metadata such as *id*, *date*, *user*, *text* and *label*. However, only the *text* and *label* fields are utilized in this study. Here are example entries from the dataset:

Table 1. Example Entries from the Dataset

Text	Label
@switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. You shoulda got David Carr of Third Day to do it. ;D	0
is upset that he can't update his Facebook by texting it... and might cry as a result School today also. Blah!	0
@MelodyLeaLamb Goooood Morning Melody! Wishing you a Wonderful Mother's Day with your family!	4
in math class, exams in 3day friday.monday.tuesday.wednesday! THAN OFF SCHOOL might do corn tho, need to buy a laptop and new camera	4
is now playing the Philippine National anthem while enjoying his 100% health http://plurk.com/p/t8xil	4

The labels 0 and 4 represent negative and positive sentiments respectively. As you can see from Table 1, the tweets in the Sentiment140 dataset reflect the informal and dynamic nature of short texts from social media. Many tweets include user mentions (e.g., @switchfoot, @MelodyLeaLamb) and URLs linking to external content (e.g., http://twitpic.com/2y1zl). The language used is often informal, with slang, abbreviations, and expressive punctuation (e.g., “shoulda”, “Blah!”, “THAN OFF SCHOOL”). The structure is casual, with grammar that is sometimes fragmented or unstandardized, which presents unique challenges for traditional natural language processing methods.

Three different sample sizes are selected for experimentation: 10,000 (10k), 30,000 (30k), and 50,000 (50k) tweets. Each sample maintains an equal number of positive and negative sentiment labels, ensuring a balanced 50:50 distribution as shown in Fig. 3.

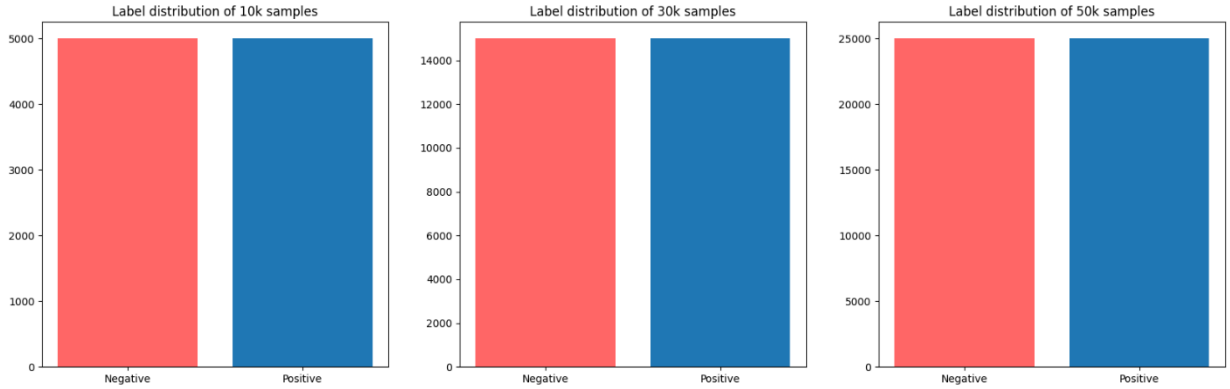


Fig. 3 Label Distribution Across Three Sample Sizes

After sampling, each sampled subset is then split into training and test sets using an 80:20 ratio. To fine-tune transformer-based models, 10% of the training set is further set aside as a validation set. This results in three distinct sets for each sample size: training, validation, and test. The detailed breakdown of the number of tweets in each set for every sampled size is presented in Table 2.

Table 2. Data Splitting Overview

Sampled subset	Training set	Test set	Validation set (10% of training set)
10,000	8,000	2,000	800
30,000	24,000	6,000	2,400
50,000	40,000	10,000	4,000

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5836

Submitted : April 27, 2025
Accepted : May 8, 2025
Published : May 10, 2025

Before training the models, each sampled subset was duplicated and separately preprocessed to accommodate the differing processing needs of traditional machine learning models and transformer-based models. For the former, tweets were cleaned by removing emojis, URLs, and usernames, converting all text to lowercase, eliminating stop words, and applying lemmatization to reduce words to their base forms. This thorough normalization helps the models focus on the most informative linguistic elements. Table 3 provides examples of tweets before and after preprocessing to illustrate the impact of data cleaning.

Table 3. Cleaned Text

Raw text	Cleaned text
@JenniferSuarez Yeah. That sucks...	yeah that suck
http://twitpic.com/7nq2p - This is what I've been moaning about all evening... killing sore	this is what ve been moaning about all evening killing sore
Awwwww zoey is scared of the storm	aww zoey is scared of the storm
@MelodyLeaLamb Goooood Morning Melody! Wishing you a Wonderful Mother's Day with your family!	good morning melody wishing you wonderful mother day with your family
@calst thank you. that's what that unfollow button's for, right?	thank you that what that unfollow button for right

In contrast, minimal preprocessing was applied to the sets used for transformer-based models. While URLs, and usernames were removed, the rest of the text was left unchanged. No tweets were discarded during the cleaning process, so the number of instances in each dataset remained consistent.

To better visualize the content of the dataset after cleaning, word clouds were generated separately for tweets labeled as positive and negative, as illustrated in Fig. 4. These word clouds highlight the most frequently occurring words within each sentiment category, providing an intuitive overview of the dominant terms associated with positive and negative sentiments.



Fig. 4 Word cloud of Negative (Left) and Positive (Right) Sentiment Tweets

Once the data was preprocessed, feature extraction was performed to convert the textual input into numerical representations suitable for each model. The process was carried out based on the approach described in the method section. For traditional machine learning models (NB, SVM, and LR), features were extracted using **TfidfVectorizer** with **ngram_range=(1,2)** and **max_features=500000**. The former allows the model to consider both unigrams (single words) and bigrams (pairs of consecutive words), helping it capture more contextual information from the text (Setiawan et al., 2024). While the latter limits the feature space to the top 500,000 most frequent terms, improving computational efficiency and reducing noise.

As for DistilBERT, we utilized the tokenizer DistilBertTokenizerFast with the distilbert-base-uncased model, which is optimized for speed. It efficiently converts raw text into input IDs, attention masks, and other necessary inputs for the model. BERT followed a similar approach, using its corresponding pretrained tokenizer to generate inputs compatible with the model.

Model Training and Evaluation

This study involves training five models: DistilBERT, BERT, NB, LR, and SVM. For each sampled subset, a corresponding test set is used to evaluate model performance. In addition to performance metrics, training time is

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5836

Submitted : April 27, 2025
Accepted : May 8, 2025
Published : May 10, 2025

recorded to assess computational efficiency.

The analysis of training time shows a clear distinction between traditional machine learning models and transformer-based models. As expected, NB, LR, and SVM completed training in under 0.1 seconds across all sampled subsets. This efficiency is largely attributed to the simplicity of their underlying algorithms, which involve relatively lightweight computations compared to deep learning models. Due to their negligible training time, they are excluded from the graph in Fig. 5, which instead focuses on transformer-based models that require significantly more computational resources. Fig. 5 illustrates the training times for DistilBERT and BERT across the three sampled subsets. As the sampled subset size increases, so does the training time, with BERT consistently taking longer than DistilBERT due to its more complex architecture and larger number of parameters.

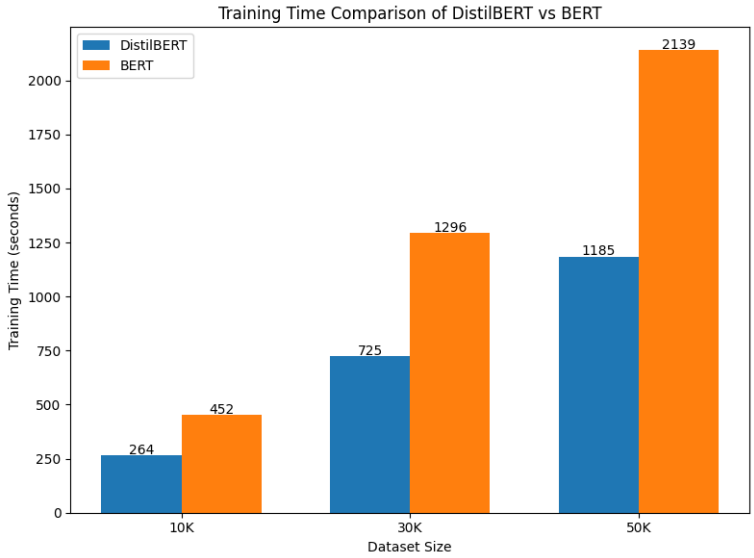


Fig. 5 Training Time Comparison of DistilBERT and BERT

The accuracy scores of each model across the three sampled subsets are presented in Fig. 6. In this figure, the results are grouped by model, with the x-axis representing the sampled subset sizes and the y-axis representing the corresponding accuracy scores. Across all subsets, transformer-based models consistently outperform traditional machine learning models. Notably, DistilBERT achieves slightly higher accuracy than BERT on the 30k and 50k sampled subsets, while their performance is comparable on the 10k subset. Meanwhile, traditional models show relatively stable but lower accuracy, with minor improvements as the sampled subset size increases.

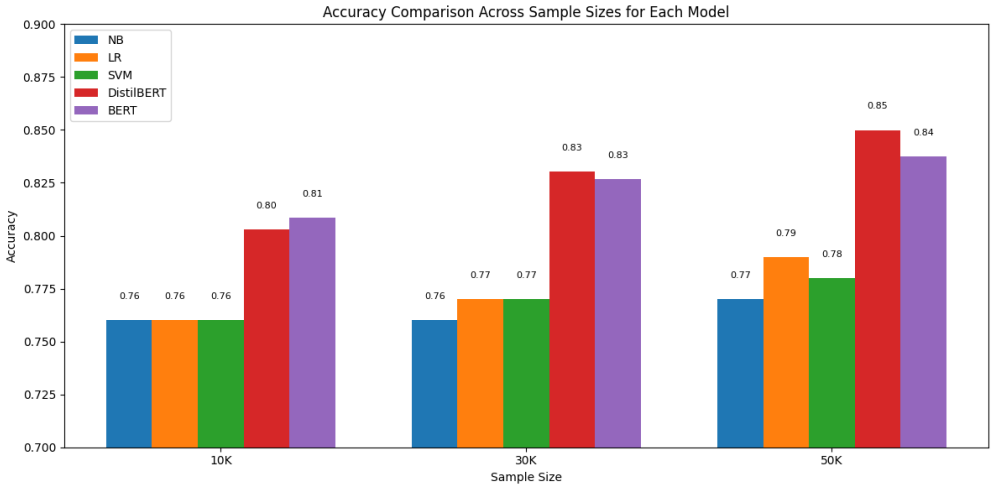


Fig. 6 Accuracy Comparison Across Sample Sizes for Each Model

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5836

Submitted : April 27, 2025
Accepted : May 8, 2025
Published : May 10, 2025

In addition to the graphical presentation, Table 4 provides a detailed summary of the accuracy and F1-scores for each model across the sampled subsets, enabling a more comprehensive comparison of their performance as the dataset size increases.

Table 4. Accuracy and F1-score Comparison Across Sample Sizes for Each Model

Sample size	Model	Accuracy	F1-score
10k	LR	0,76	0,76
	NB	0,76	0,76
	SVM	0,76	0,76
	BERT	**0,809**	0,797
	DistilBERT	0,803	**0,807**
30k	LR	0,77	0,77
	NB	0,76	0,76
	SVM	0,77	0,77
	BERT	0,828	**0,8292**
	DistilBERT	**0,83**	0,8291
50k	LR	0,79	0,79
	NB	0,77	0,78
	SVM	0,78	0,78
	BERT	0,837	0,834
	DistilBERT	**0,85**	**0,844**

To further illustrate the classification performance, the confusion matrix of the DistilBERT model trained on the 50k sampled subset is presented in Fig. 7. This model was selected because it achieved the highest accuracy and F1-score among all tested models.

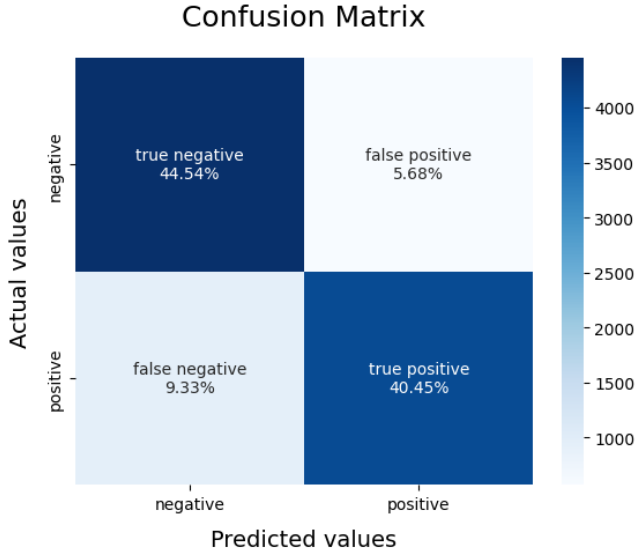


Fig. 7 Confusion Matrix of the DistilBERT Model Trained on the 50k Sampled Subset

The confusion matrix provides a detailed view of the model's predictions, showing the distribution of true positives, true negatives, false positives, and false negatives, and offering additional insights into the types of errors made during sentiment classification. The model correctly identified negative sentiments 44.54% of the time and positive sentiments 40.45% of the time. However, it misclassified 9.33% of positive examples as negative and 5.68% of negative examples as positive.

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5836

Submitted : April 27, 2025
Accepted : May 8, 2025
Published : May 10, 2025

DISCUSSIONS

The results of this study highlight the advantages of transformer-based models, particularly DistilBERT, for sentiment analysis on short texts. In addition to achieving strong predictive performance, DistilBERT demonstrated significantly faster training times compared to BERT across all sampled subsets. For instance, on the 10k subset, DistilBERT required 4 minutes to train, while BERT required 7,5 minutes. On the 30k subset, DistilBERT completed training in 12 minutes, whereas BERT took 21,6 minutes. For the 50k subset, DistilBERT trained in 19,8 minutes, and BERT in 35,6 minutes. On average, DistilBERT trained approximately **1,8 times** faster than BERT. This substantial reduction in training time, combined with its competitive performance, makes DistilBERT a highly practical choice for scenarios where computational resources or training time are constrained.

In terms of predictive performance, DistilBERT consistently demonstrated strong results across all sampled subsets. It slightly outperformed BERT in both the 30k and 50k subsets in terms of accuracy and F1-score, suggesting that its lighter architecture did not compromise its effectiveness on larger datasets. On the 10k subset, DistilBERT remained highly competitive with BERT, showing only a minimal difference in performance. Both transformer-based models significantly outperformed traditional machine learning models (LR, NB, and SVM), which consistently achieved lower accuracy and F1-scores regardless of the dataset size. Notably, the best result was achieved by DistilBERT trained on the 50k subset, reaching an accuracy of **85%** and an F1-score of **84%**. These results reinforce the strength of transformer-based approaches in handling the complexity and variability inherent in short informal texts.

Overall, the best DistilBERT model, trained on the 50k subset, shows a good balance between detecting both classes but exhibits a slightly higher tendency to miss positive sentiments compared to negative ones. The relatively low false positive rate (**5.68%**) indicates that when the model predicts a positive sentiment, it is more likely to be correct, which is important in many sentiment-sensitive applications. This behavior suggests that the model is more conservative when identifying positive sentiments, favoring precision over recall in this class.

CONCLUSION

This study demonstrates the effectiveness of transformer-based models, particularly DistilBERT, in performing sentiment analysis on short texts. Across multiple sampled subsets, DistilBERT consistently achieved strong accuracy and F1-scores, outperforming traditional machine learning models such as Naive Bayes, Logistic Regression, and SVM. It also showed competitive performance against BERT, delivering slightly higher accuracy on larger datasets (30k and 50k subset) while maintaining substantially faster training times. These results highlight DistilBERT's suitability for scalable and time-sensitive sentiment analysis applications.

Overall, the findings support the use of lightweight transformer models like DistilBERT for practical sentiment analysis tasks, especially when balancing efficiency and performance is crucial. Future work could involve domain-specific fine-tuning, expanding the analysis to multilingual datasets, and investigating additional machine learning techniques for further comparison.

REFERENCES

Abei, F., Sulaeman, A. A., & Suprapto, S. (2025). Twitter Sentiment Towards 2024 Jakarta Governor Candidates With Naïve Bayes Algorithm. *Journal of Computer Networks, Architecture and High Performance Computing*, *7*(1), 265–277. https://doi.org/10.47709/cnahpc.v7i1.5358

Al-Husaini, M., Nurpazrin, Y. S., Rizal, R., & Lukmana, H. H. (2024). Convolutional Neural Networks and TensorFlow-Based Enhance Correction for Color Visual Deficiency (CVD) Impaired. *2024 Ninth International Conference on Informatics and Computing (ICIC)*, 1–6. https://doi.org/10.1109/ICIC64337.2024.10957170

Barreto, S., Moura, R., Carvalho, J., Paes, A., & Plastino, A. (2022). Sentiment analysis in tweets: an assessment study from classical to modern text representation models. *Data Mining and Knowledge Discovery*. http://arxiv.org/abs/2105.14373

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. http://arxiv.org/abs/1810.04805

Fransiscus, & Girsang, A. S. (2022). Sentiment Analysis of COVID-19 Public Activity Restriction (PPKM) Impact using BERT Method. *International Journal of Engineering Trends and Technology*, *70*(12), 281–288. https://doi.org/10.14445/22315381/IJETT-V70I12P226

Jim, J. R., Talukder, M. A. R., Malakar, P., Kabir, M. M., Nur, K., & Mridha, M. F. (2024). Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review. *Natural Language Processing Journal*,

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5836

Submitted : April 27, 2025
Accepted : May 8, 2025
Published : May 10, 2025

6, 100059. https://doi.org/10.1016/j.nlp.2024.100059

Jusli, D. T. A., & Kurniawan, R. (2024). Analysis Of Opinion Sentiment Towards Electric Vehicle Tax On Social Media X Using The Support Vector Machine (SVM) Method. *Journal of Computer Networks, Architecture and High Performance Computing*, *6*(4), 1792–1808. https://doi.org/10.47709/cnahpc.v6i4.4739

Karl, F., & Scherp, A. (2023). Transformers are Short Text Classifiers: A Study of Inductive Short Text Classifiers on Benchmarks and Real-world Datasets. *Machine Learning and Knowledge Extraction*. https://doi.org/https://doi.org/10.1007/978-3-031-40837-3_7

Kazanova. (2017). *Sentiment140 dataset with 1.6 million tweets*. Kaggle. https://www.kaggle.com/datasets/kazanova/sentiment140

Khan, J., Ahmad, K., Jagatheesaperumal, S. K., & Sohn, K. A. (2025). Textual variations in social media text processing applications: challenges, solutions, and trends. *Artificial Intelligence Review*, *58*(3). https://doi.org/10.1007/s10462-024-11071-z

Knittel, J., Koch, S., Tang, T., Chen, W., Wu, Y., Liu, S., & Ertl, T. (2021). Real-Time Visual Analysis of High-Volume Social Media Posts. *IEEE Transactions on Visualization and Computer Graphics*, 879–889. https://doi.org/10.1109/TVCG.2021.3114800

Loureiro, D., Rezaee, K., Pilehvar, M. T., & Camacho-Collados, J. (2021). Analysis and Evaluation of Language Models for Word Sense Disambiguation under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license. *Computational Linguistics*, *47*(2). https://doi.org/10.1162/COLI

Mao, Y., Liu, Q., & Zhang, Y. (2024). Sentiment analysis methods, applications, and challenges: A systematic literature review. In *Journal of King Saud University - Computer and Information Sciences* (Vol. 36, Issue 4). King Saud bin Abdulaziz University. https://doi.org/10.1016/j.jksuci.2024.102048

Nguyen, T. H., Nguyen, H. H., Ahmadi, Z., Hoang, T. A., & Doan, T. N. (2021). On the Impact of Dataset Size:A Twitter Classification Case Study. *ACM International Conference Proceeding Series*, 210–217. https://doi.org/10.1145/3486622.3493960

Palomino, M. A., & Aider, F. (2022). Evaluating the Effectiveness of Text Pre-Processing in Sentiment Analysis. *Applied Sciences (Switzerland)*, *12*(17). https://doi.org/10.3390/app12178765

Putra, R. F., Mukhlis, I. R., Datya, A. I., Pipin, S. J., Reba, F., Al-Husaini, M., Mandowen, S. A., Zain, N. N. L. E., & Judijanto, L. (2024). *ALGORITMA PEMBELAJARAN MESIN (Dasar, Teknik, dan Aplikasi)*. Sonpedia Publishing Indonesia. www.buku.sonpedia.com

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *EMC*. http://arxiv.org/abs/1910.01108

Sarkar, S., Babar, M. F., Hassan, M. M., Hasan, M., & Santu, S. K. K. (2024). Processing Natural Language on Embedded Devices: HowWell Do Transformer Models Perform? *ICPE 2024 - Proceedings of the 15th ACM/SPEC International Conference on Performance Engineering*, 211–222. https://doi.org/10.1145/3629526.3645054

Setiawan, Y., Maulidevi, N. U., & Surendro, K. (2024). The Optimization of n-Gram Feature Extraction Based on Term Occurrence for Cyberbullying Classification. *Data Science Journal*, *23*(1). https://doi.org/10.5334/dsj-2024-031

Siino, M., Tinnirello, I., & La Cascia, M. (2024). Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. *Information Systems*, *121*. https://doi.org/10.1016/j.is.2023.102342

Tan, K. L., Lee, C. P., & Lim, K. M. (2023). A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. In *Applied Sciences (Switzerland)* (Vol. 13, Issue 7). MDPI. https://doi.org/10.3390/app13074550

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.