When Do Transformers Actually Win? Sentiment Classification In Low-Data Settings

Mr. Shreyash Dodekar[1], Ms.Anjali Jawale[2], Mrs. Shital Pashankar[3]
[1,2]*MSc Computer Application, School Of Information Technology, Indira University, Wakad, Pune, India*
[3]*Computer Science, School Of Information Technology, Indira University, Wakad, Pune, India*
doi.org/10.64643/ijirt.208452-459

***Abstract*—Nowadays, use of Transformer based language is increasing rapidly to perform NLP (Natural Language Processing) related tasks. DistilBERT is a model which provide strong classification performance can understand the context of sentences or words. The performance of Transformer usually depends on having labelled training data. But, in real world, obtaining labelled data is time consuming, expensive and difficult. This research focuses on "Does a Transformer always provide a meaningful advantage when only a limited amount of labelled data is present?" We are solving this question through a controlled sentiment-classification experiment which uses the Yelp Labelled Reviews dataset and contains customer reviews labelled as positive or negative in binary representation (0 & 1). Multinomial Naive Bayes, Logistic Regression and Linear SVM (Support Vector Machine) are the three traditional approaches used in this model. These models use TF-IDF (Term Frequency Inverse Document Frequency) to convert review text into numerical data called features.**

Further these results will be compared with DistilBERT which is a lightweight Transformer-based language model. A key factor of this research is using labelled training data in different proportions. The experiment will use five training conditions at 10%, 25%, 50%, 75% and 100%. This shows how the model will behave when the training data is Less or gradually increased. The performance of each model will be measured using precision, recall, accuracy, F-1 score.

These measurements will not only evaluate model performance but also the computational cost behind it. The expected outcome of this research is practical understanding of the relationship between training data availability, computational cost and performance. The findings help to determine when a TF-IDF based model will be enough or using a Transformer model may help in improvement. This model will be proved useful when labelled data is limited. Overall, the study aims in making a choice between Traditional machine learning model or Transformer based model for sentiment classification.

***Index Terms*—Sentiment Analysis, NLP, DistilBERT, TF-IDF, Logistic Regression, Multinomial Naive Bayes, Support Vector Machine (SVM), Hyperparameter tuning.**

I. INTRODUCTION

1.1. Background

Sentiment analysis is an NLP (Natural Language Processing) technique which is commonly used to identify opinion based or emotion-based orientation of textual data. It is used to identify whether a piece of text belongs to positive or negative sentiment. Due to rapid growth of online reviews, social media activities, customer feedbacks and various type of user generated content, sentiment classification has become an important application of NLP.

Traditional classification systems often use techniques like TF-IDF for text representation. Machine learning algorithms like Naive Bayes, Linear Support Vector Machine (SVM) and Logistic Regression [7] works on these techniques. This Machine learning algorithms are simple, require low computational cost and are efficient when labelled data is available. TF-IDF mainly focuses on individual words and does not fully establish the relationship between words in a sentence.

The development of pretrained Transformer models is changing NLP-based text classification. Models like BERT and DistilBERT generate representation of text more effectively. DistilBERT is a smaller version of BERT which is computationally more efficient which acts as an alternative for BERT.

An important question rises – Whether a Transformer model is always better when only limited amount of labelled data is available. It is not necessary that a model which performs best on large dataset, will also perform the same with a smaller dataset. Therefore, examining model performance at different training

sizes can provide useful information in choosing appropriate model.

In this study, Traditional Machine learning models like Naive Bayes, Linear Support Vector Machine (SVM), Logistic Regression [7] and Transformer model – DistilBERT are evaluated under different dataset sizes. The study uses 10%,20%,50%,75% and 100% of the available dataset. In addition, we involve hyperparameter tuning to examine whether the performance of DistilBERT increases through appropriate training configurations.

1.2. Motivation

The motivation for this research come from the difficulty of obtaining large amounts of high-quality data for classification. For a researcher working with a small labelled dataset is insufficient. It is not always a good option to select most advanced NLP model. It is necessary to know how much data is actually required for training and whether the additional computational cost highlights a meaningful improvement in model performance. Traditional Machine learning models are comparatively faster and require less computational cost. On the other hand, DistilBERT can capture textual information more precisely where TF-IDF lacks.

1.3. Research problem.

Existing studies examined traditional machine-learning vs Transformers, Training data size proportions and hyperparameter optimization as separate factors. However, limited work on it has evaluated these factors together in a binary classification setting. In addition, there is a lack of comparison between Traditional Machine learning and Transformers across multiple labelled training data.

1.4. Research Questions

The present study points to solve the following research questions:

RQ1: How does size of labelled data affect the performance of traditional machine learning algorithms and DistilBERT for binary sentiment classification?

RQ2: How does DistilBERT compare with traditional TF-IDF based classifiers like Linear SVM, Logistic Regression and Naive Bayes across different training-data sizes?

RQ3: Does hyperparameter optimization improve the performance of DistilBERT for sentiment classification?

RQ4: Which training-data proportion provides the most effective balance between dataset size and sentiment-classification performance?

1.5. Research Objectives

1. To investigate sentiment classification performance using different size of labelled training data.
2. To evaluate traditional machine-learning models like Linear SVM, Logistic Regression and Naive Bayes [7] using TF-IDF text representations.
3. To fine tune DistilBERT for binary sentiment classification.
4. To observe the effect of hyperparameter settings like learning rate and weight decay on DistilBERT performance.
5. To compete the F1-score of traditional machine-learning models with fine-tuned DistilBERT model.
6. To determine whether increasing the amount of training data consistently improves model performance.

II. LITERATURE REVIEW

A. Dataset Size and Low-Data Sentiment Classification

Different levels of data available for training affects performance on twitter dataset. It is an important study as it focuses on the concept that adding more labelled data will always improve a model's performance. To evaluate this [1] divided the amount of training data at different levels to examine traditional machine learning models performance. Their results showed that it is not always necessary that increasing the amount of data will always increase the efficiency of model. However, [1] focused on Twitter classification dataset, the present study uses yelp reviews dataset which are binary sentiment classified. Additionally, the study uses controlled training proportions of 10%, 25%, 50%, 75% and 100% to compare Traditional machine learning algorithms like Logistic Regression, Multinomial Naive Bayes, Linear SVM Vs Transformer algorithm DistilBERT [1]. Text classification works when only small or limited amount of labelled data is available. They evaluated two traditional machine learning models like Naive

Bayes and SVM. Different training sizes and number of experimental combinations were made which concluded that performance of model can be achieved with relatively small training datasets also. The limitation of present research is that it only focuses on traditional machine learning approaches but it does not include pretrained Transformer like DistilBERT, which provides better performance for same limited dataset [2]. Low-resource problem in fine-grained sentiment analysis while using a pretrained model with data augmentation. Their results achieved comparable performance to baseline systems while using 20% training data. This showed how pretrained language representations can prove useful when labelled data is very less. As this study focuses on only one data setting i.e. 20%, the present study keeps the flexibility to grow or shrink labelled data proportion directly [3]. XLM-RoBERTa and BERT used multilingual sentiment analysis for Bengali language. The study explored limited availability of resources for Bengali language and interpreted strong results showing upto 95% accuracy for binary sentiment classification. The study demonstrates the importance of pretrained Transformer models when labelled data is limited. However, the study reflects the main concern as Language scarcity rather than comparing the performance of traditional and Transformer models as the quantity changes [4]. StraRA, is a self-training and task augmentation approach used for few-shot learning. The study shows how AI can learn effectively if you give fine-tuning examples. It drops all extra tricks and uses basic fine-tuning tricks on a simplified model DistilBERT. The study focuses on how to improve the performance with limited labelled data and excluding extra factors. Previous research focuses on shortcut ways to train AI with less data. Unlike it, the present work skips these shortcuts to test how the AI learns on its own as data increases.[5]

B. DistilBERT and Transformer-Based Sentiment Analysis

Comparison of TF-IDF vs DistilBERT based contextual representations for sentiment classification of banking and financial news. The study shows how lightweight AI outperforms Traditional Machine learning methods in real world tasks. The present work takes the idea further by testing how both approaches perform across five different amounts (10%, 25%, 50%, 75% and 100%) in low to high data settings range [6].

Annotated Urdu sentiment dataset which contains 9,312 reviews are evaluated using rule-based deep-learning, traditional machine-learning and multilingual BERT approaches. The models used in the study includes Logistic Regression, Naive Bayes and Support Vector Machine [7]. It focuses on the challenge of low-resource language processing. This study is useful for present research because it shows how pretrained transformer model can compete strongly with traditional approaches in sentiment classification. It does not implement the changes in training data proportions to evaluate performance gap at each level.[7]

COVID-19 online news dataset was analyzed using DistilBERT for binary sentiment classification. The core objective of this research was to obtain strong NLP performance while avoiding memory requirements and computational cost associated with full BERT. The DistilBERT model nearly achieved 94% accuracy after two epochs. This work uses DistilBERT as a lightweight transformer for sentiment analysis. However, it evaluates a single dataset and does not include multiple training conditions or compare DistilBERT with the exact TF-IDF classifiers [8]. Implemented DistilBERT using positive and negative sentiment classification of COVID-19 survey responses. Their experiments reported the effects of parameters like learning rate and weight decay using different training configurations. This study is more relevant to present research after adding hyperparameter tuning. It shows that DistilBERT performance not only depends on the dataset but also on training configuration [9]. Investigation of Transformer based sentiment classification under limited data settings.

It uses text augmentation and knowledge distillation. The study further observes two main problems – The difficulty of training sentiment classifiers with few labelled data and required computational cost of Transformer models. This study motivates Transformer efficiency and data scarcity. As this method uses text augmentation, it becomes more complex which provides further motivation for simpler experimental design used in present research [10].

C. Low resource and Reproducible Sentiment Classification

According to [11], small changes in an experiment setup can completely change the model's output. The study examines various factors like experiment methodology, implementation details and reported results. It shows how differences in experimental setup can affect replicability. This is an important study for present work because of the methodology used in it. The objective is to compare several models under different data settings, evaluation metrics and experimental procedure [11]. Sentiment classification was system on multilingual version of BERT to evaluate African languages which has very less training data. So, to achieve highest possible accuracy the study used two technical strategies - Supervised adversarial contrastive learning and Sentiment-aware representation learning. Their research proved that pretrained Transformers perform good when they know languages with very little labelled data. In fact, the study shows a drawback of missing a practical threshold implying no specific point were switching to a Transformer actually becomes worthy compared to the computational cost. The present study leaves those extra tricks aside to answer the question: How many labelled data do you actually need before moving from simple traditional machine learning to a faster, lightweight Transformer? [12].

D. Hyperparameter tuning abd fine-tuning of Transformer model.

Hyperparameter tunning settings like learning rate and batch size included to reach top accuracy [13]. Their study says that proper configuration is required for a fair evolution which are usually avoided by using random or default settings. [13] performed fine tuning BERT on a fixed dataset. This new study combines hyperparameter learning rate and weight decay with testing across different amounts of training data. AI models need proper tuning to show their true performance.

Hyperparameters like batch size, learning rate, weight decay, epochs, optimizer, warm-up ratio and dropout change the accuracy of how well a Transformer model can predict financial sentiment. Their study tuned everything which makes it more complex and results into higher computational costs. The present research focuses on learning rate and weight decay which keeps experiments clear, manageable and easy to reproduce.

E. Traditional ML vs Transformer model.

DistilBERT model aachieved an F1 score of 0.79 (79%) on e-commerce reviews dataset, which easily beats traditional machine learning models. Their results showed strong evidence that lightweight Transformers outperforms traditional methods in real-world sentiment tasks. The drawback of this study is that it uses imbalanced dataset and tests only at one size. The present research tests model on five different level [15].

An experiment was conducted to test how different text cleaning methods like stop-word removal, lowercasing and removing punctuation affects both Traditional machine learning models (like Linear SVM or NB) and Transformers (like BERT or DistilBERT). The study shows a surprising finding that how traditional learning model can sometimes outperform complex Transformers if we give the right preprocessing. [16] Prove that Traditional ML is competitive and Fair benchmarking is essential. The present study solves this drawback by fairly comparing both Traditional model and Transformers at equal training data settings [16].

F. Recent DistilBERT comparison.

Testing of dataset size affects model performance and it was tested by directly comparing three classic traditional algorithms: SVM, Naive Bayes and Logistic Regression against Transformers: DistilBERT. The study uses 10,000, 30,000 and 50,000 tweets available on the public Sentiment140 dataset containing different sample sizes [17]. Their work found that data availability directly changes the performance gap between TF-IDF representations and Transformer architectures.

The present study solves problem based on Domain Generalization and Finer Data Steps which were missing in existing study [17]. Test of five Transformer architectures - BERT, DistilBERT, multilingual BERT, ALBERT and BERT-CNN on a sentiment analysis task in Bahasa Malaysia which uses social media texts with mixed languages and regional slang. Their study achieved 88.96% performing slightly behind heavier variants like full BERT (89.5%). This paper provides strong empirical backing for choosing DistilBERT as primary Transformer model because of its Computational Efficiency and Resource Optimization [18].

G. Hyperparameter optimization as a missing parameter.

BERT, DistilBERT, XLM-RoBERTa, RoBERTa and Optuna were evaluated for hyperparameter search. Their study included tuning factors like batch size (16), learning rate(5e-5), epochs (1), warm-up steps (200) and weight decay (0.01) [19].

III. RESEARCH METHODOLOGY

A. Problem Statement.

Sentiment classification is mostly performed on traditional machine learning techniques like Logistic Regression, Naive Bayes and Support Vector Machines [7] with TF-IDF-based text representations. These methods are relatively simple and computationally efficient but may face difficulty while establishing contextual relationships between words.

Transformer based models like BERT and DistilBERT shows strong performance in various NLP tasks [18]. However, they require high computational cost their performance can depend on amount of labelled training data. Existing studies works on a fixed dataset size and shows lack of comparison between traditional machine learning and Transformer model.

The present study addresses this problem by evaluating Linear SVM, Logistic Regression, Naive Bayes and DistilBERT using 10%, 25%, 50%, 75% and 100% of the available data settings. In addition, we perform hyperparameter tuning to increase DistilBERT model performance.

B. Scope of the study.

The study uses binary sentiment classification dataset containing textual data. The investigation is bounded to the comparison of three traditional machine learning models and lightweight Transformer model DistilBERT. The traditional machine learning model like Naive Bayes, Logistic Regression and Linear SVM use TF-IDF based text representations. The study evaluates DistilBERT using different quantities of training data - 10%, 25%, 50%, 75% and 100%. Hyperparameter optimization is performed to increase the performance of model.

Four unique combinations were made from these tuning settings –

Learning rate: 2×10^{-5} and 5×10^{-5}

Weight decay: 0.01 and 0.05

Evaluation metrics used - Precision, Recall, Accuracy and F1-score. The study does not focus on multilingual sentiment classification.

C. Dataset.

The study uses Yelp sentiment classification dataset which contains 1,000 labelled texts with two sentiment classes.

Table 1: Dataset

PROPERTY	DESCRIPTION
Total Samples	1,000
Positive Samples	500
Negative Samples	500
Number Of Classes	2

D. Data Preprocessing.

The model was prepared for classification before training. The missing values were checked beforehand and the sentiment were represented as binary classes. For traditional machine learning, the textual data is converted into numerical features using TF-IDF. For DistilBERT the raw data was processed using Tokenizer, it converts each text into token IDs.

E. Traditional Machine Learning Models.

Three traditional algorithms were selected for research work. Logistic Regression is simple and interpretable for binary text classification.

Naive Bayes has good computational efficiency, it predicts the data point using probability. Linear SVM uses a decision boundary which separates the two sentiment classes as positive and negative.

F. Transformer Model.

DistilBERT is a lightweight BERT model which is smaller and computationally lighter than the original BERT.

G. Hyperparameter Optimization.

Four different combinations were used to evaluate the highest performing metrics (epoch - 3).

Table 2

Learning rate	Weight decay
2e-5	0.01
2e-5	0.01
5e-5	0.05
5e-5	0.05

H. Evaluation Metrics.

TP = True Positive.
TN = True Negative.
FP = False Positive.
FN = False Negative.

1. $Accuracy\ score = \frac{TP+TN}{TP+TN+FP+FN}$

2. $\text{Precision score} = \frac{TP}{FP+TP}$

3. $Recall\ \text{score} = \frac{TP}{FN+TP}$

4. $F1\ \text{score} = 2 * \frac{\text{Precision+Recall}}{\text{Precision} * \text{Recall}}$

I. Experimental Procedure.

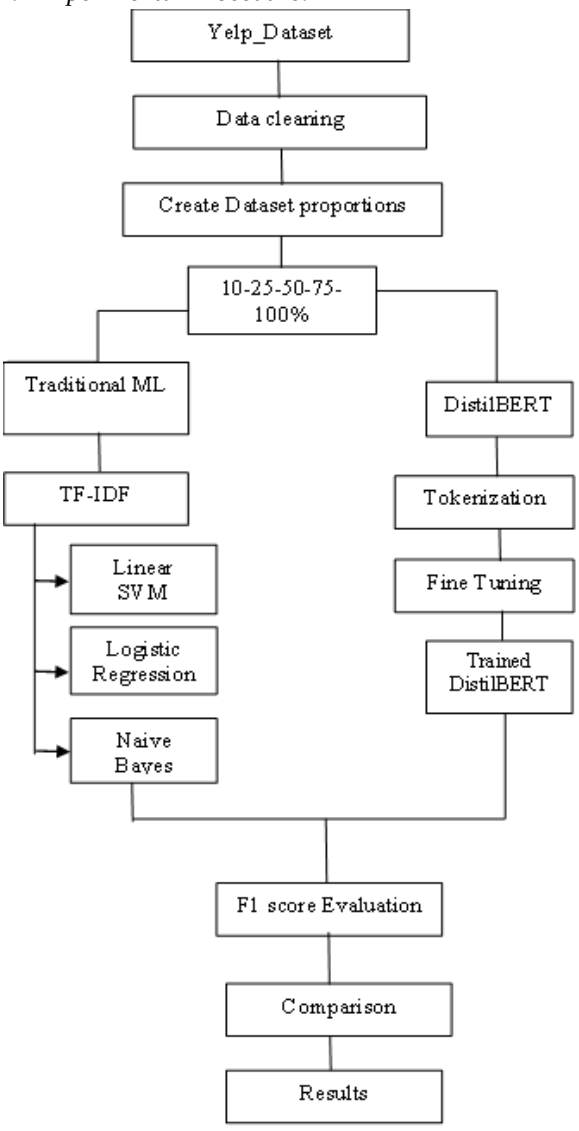

Figure 1 Architecture of Proposed Model

IV. RESULTS AND DISCUSSIONS

A. Traditional ML Performance.

From *table 3* we can see the highlighted part is the algorithm with highest performance at different data settings.

Table 3

Dataset	Linear SVM F1	Logistic Regression F1	Naive Bayes F1
10%	67.34%	68.00%	70.75%
25%	68.42%	65.48%	69.84%
50%	76.53%	74.32%	76.29%
75%	79.81%	80.19%	81.16%
100%	81.34%	81.95%	81.52%

From the above results we can observe that Naive Bayes performs better at 3 data proportions.

B. Traditional ML vs DistilBERT Performance.

From *table 4* we can observe the highlighted part is the algorithm with highest performance at different data settings.

Table 4

Training Data	Linear SVM F1 (%)	Logistic Regression F1 (%)	Naive Bayes F1 (%)	DistilBERT F1 (%)
10%	67.34	68.00	70.75	83.81
25%	68.42	65.48	69.84	92.23
50%	76.53	74.32	76.29	92.00
75%	79.81	80.19	81.16	93.07
100%	81.34	81.95	81.52	93.53

From the above results we can observe that DistilBERT performs better at all data proportions.

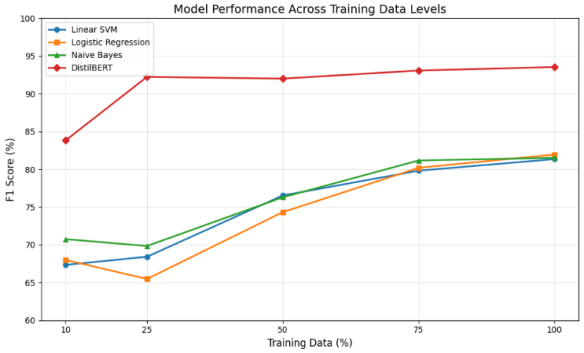

Figure 2

C. Computational cost of Traditional ML vs DistilBERT.

Table 5

Training Data	Linear SVM (sec)	Logistic Regression (sec)	Naive Bayes (sec)	DistilBERT (sec)
10%	0.0058	0.0213	0.0057	3.2808
25%	0.0082	0.0334	0.0087	10.1103
50%	0.0152	0.0279	0.0050	11.8439
75%	0.0159	0.0851	0.0080	17.4711
100%	0.0233	0.1244	0.0081	17.1967

From *table 5* we can observe that DistilBERT requires higher computational cost compared to other Traditional Machine learning model at every data proportions.

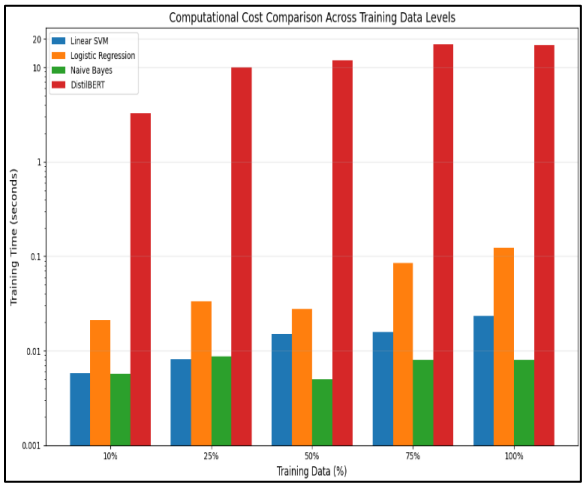

Figure 3

D. Traditional ML Performance after adding Hyperparamter.

There is a clear difference in performance after tuning Hyperparameters.

Hyperparameters used:

Table 6

Model	Hyperparameter tuned	Values tested
Linear SVM	C	1, 10
Logistic Regression	C	0.1, 10
Naive Bayes	alpha	0.1, 0.5, 1.0

Table 7

Training Data	Linear SVM F1	Logistic Regression F1	Naive Bayes F1	Best Traditional ML
10%	67.29 %	68.20%	68.84 %	Naive Bayes
25%	74.61 %	71.35%	72.54 %	Linear SVM
50%	78.61 %	80.40%	81.44 %	Naive Bayes
75%	83.17 %	82.93%	81.77 %	Linear SVM
100%	84.54 %	86.01%	84.26 %	Logistic Regression

↓ Performance Loss

Table 8

Dataset	SVM Change	Logistic Regression Change	Naive Bayes Change
10%	−0.05% ↓	+0.20% ↑	−1.91% ↓
25%	+6.19% ↑	+5.87% ↑	+2.70% ↑
50%	+2.08% ↑	+6.08% ↑	+5.15% ↑
75%	+3.36% ↑	+2.74% ↑	+0.61% ↑
100%	+3.20% ↑	+4.06% ↑	+2.74% ↑

We can clearly observe that using Hyperparameters in Traditional ML models increased the accuracy for every data proportion except for the two highlighted in *Table 8*.

E. DistilBERT before tuning vs DistilBERT after tuning.

Four Combinations of tuning settings were applied on this model; it gave the following result.

Table 9

Learning rate	Weight decay
2e-5	0.01
2e-5	0.01
5e-5	0.05
5e-5	0.05

Table 10

Dataset Size	DistilBERT Before Tuning F1 (%)	DistilBERT After Tuning F1 (%)	Improvement (%)
10%	83.81%	87.68%	+3.87%
25%	92.23%	92.39%	+0.16%
50%	92.00%	92.11%	+0.11%
75%	93.07%	97.48%	+4.41%
100%	93.53%	92.16%	−1.37%

From fig 10, we can see the increase in f1 score at 10,20,50,75% where 75% has highest increment. However, unexpectedly we can see a drop in performance (-1.37%) of DistilBERT after tuning.

V. CONCLUSION & FUTURE SCOPE

The experimental results before hyperparameter tuning shows a huge advantage of using DistilBERT model over Traditional machine learning techniques. DistilBERT has higher accuracy compared to every traditional machine learning model at all data proportions. DistilBERT results are more optimized, but require higher computational cost compared to traditional machine learning models. After hyperparameter tuning the results show a gradual increase in performance at 10%, 25%, 50% and 75% but shows a small drop at 100% data level. However, the results show that increasing the amount of training data does not automatically produce improvement in DistilBERT performance (e.g., at 75% f1 score is 97.48% which decreased to 92.16%). 75% training data shows higher performance compared to other data levels with a gradual increase of +4.41%. A learning rate of 2e-5 performed best at 25%, 50% and 100% whreas 5e-5 performed best at 10% and 75%. Therefore, fine tuning configuration is necessary for better performance of a model.

The present study can be further expanded by increasing the sample size of dataset (present sample size = 1000). Future research can focus on other lightweight transformer models to check the highest accuracy and low computational cost. More hyperparameters like batch size, maximum sequence length and warm up steps can be included in further studies. Future work can also focus on multilingual datasets to achieve more flexibility of models.

REFERENCES

[1] T. H. Nguyen, H. H. Nguyen, Z. Ahmadi, T. A. Hoang, and T. N. Doan, “On the impact of dataset size: A Twitter classification case study,” in *Proc. IEEE/WIC/ACM Int. Conf. Web Intelligence and Intelligent Agent Technology*, Dec. 2021, pp. 210–217.

[2] M. Riekert, M. Riekert, and A. Klein, “Simple baseline machine learning text classifiers for small datasets,” SN Computer Science, vol. 2, no. 3, p. 178, 2021.

[3] H. Y. Lu, J. Yang, C. Hu, and W. Fang, “One for ‘All’: A unified model for fine-grained sentiment analysis under three tasks,” PeerJ Computer Science, vol. 7, p. e816, 2021.

[4] A. Bhowmick and A. Jana, “Sentiment analysis for Bengali using transformer-based models,” in Proc. 18th Ints. Conf. Natural Language Processing (ICON), Dec. 2021, pp. 481–486.

[5] T. Vu, M. T. Luong, Q. Le, G. Simon, and M. Iyyer, “STraTA: Self-training with task augmentation for better few-shot learning,” in Proc. 2021 Conf. Empirical Methods in Natural Language Processing, Nov. 2021, pp. 5715–5731.

[6] V. Dogra, A. Singh, S. Verma, Kavita, N. Z. Jhanjhi, and M. N. Talib, “Analyzing DistilBERT for sentiment classification of banking financial news,” in Intelligent Computing and Innovation on Data Science: Proc. ICTIDS 2021, Singapore: Springer Nature Singapore, 2021, pp. 501–510.

[7] L. Khan, A. Amjad, N. Ashraf, and H. T. Chang, “Multi-class sentiment analysis of Urdu text using multilingual BERT,” Scientific Reports, vol. 12, no. 1, p. 5436, 2022.

[8] S. K. Akpatsa, H. Lei, X. Li, V. H. K. S. Obeng, E. M. Martey, P. C. Addo, and D. D. Fiawoo, “Online News Sentiment Classification Using DistilBERT,” Journal of Quantum Computing, vol. 4, no. 1, p. 1, 2022.

[9] M. Jojoa, P. Eftekhar, B. Nowrouzi-Kia, and B. Garcia-Zapirain, “Natural language processing analysis applied to COVID-19 open-text opinions using a DistilBERT model for sentiment categorization,” AI & Society, vol. 39, no. 3, pp. 883–890, 2024.

[10] X. Gong, W. Ying, S. Zhong, and S. Gong, “Text sentiment analysis based on transformer and augmentation,” Frontiers in Psychology, vol. 13, p. 906061, 2022.

[11] G. M. Di Nunzio and R. Minzoni, “A thorough reproducibility study on sentiment classification: Methodology, experimental setting, results,” Information, vol. 14, no. 2, p. 76, 2023.

[12] D. Hu, L. Wei, Y. Liu, W. Zhou, and S. Hu, “UCAS-IIE-NLP at SemEval-2023 Task 12: Enhancing generalization of multilingual BERT for low-resource sentiment analysis,” in Proc.

17th Int. Workshop Semantic Evaluation (SemEval-2023), Jul. 2023, pp. 1849–1857.

[13] S. Y. Ng, K. M. Lim, C. P. Lee, and J. Y. Lim, "Sentiment analysis using DistilBERT," in Proc. 2023 IEEE 11th Conf. Systems, Process & Control (ICSPC), Dec. 2023, pp. 84–89.

[14] D. K. Nasiopoulos, K. I. Roumeliotis, D. P. Sakas, K. Toudas, and P. Reklitis, "Financial sentiment analysis and classification: A comparative study of fine-tuned deep learning models," International Journal of Financial Studies, vol. 13, no. 2, p. 75, 2025.

[15] K. Taneja, J. Vashishtha, and S. Ratnoo, "Transformer based unsupervised learning approach for imbalanced text sentiment analysis of e-commerce reviews," Procedia Computer Science, vol. 235, pp. 2318–2331, 2024.

[16] M. Siino, I. Tinnirello, and M. La Cascia, "Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers," Information Systems, vol. 121, p. 102342, 2024.

[17] M. S. Asyaky, M. Al-Husaini, and H. H. Lukmana, "Sentiment analysis on short social media texts using DistilBERT," Journal of Computer Networks, Architecture and High-Performance Computing, vol. 7, no. 2, pp. 524–533, 2025.

[18] M. A. Zulkalnain, A. R. Syafeeza, W. H. M. Saad, and S. Rahaman, "Evaluation of Transformer-Based Models for Sentiment Analysis in Bahasa Malaysia," Journal of Telecommunication, Electronic and Computer Engineering (JTEC), vol. 17, no. 1, pp. 29–33, 2025.

[19] G. Eser and C. Sahin, "Sentiment analysis and rating prediction for app reviews using transformer-based models," *International Journal of Advanced Natural Sciences and Engineering Research*, vol. 8, no. 4, pp. 372–379, 2024.