JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

FINE-TUNING LIGHTWEIGHT TRANSFORMER MODELS FOR TEXT CLASSIFICATION: A COMPARATIVE ANALYSIS OF DISTILBERT, BERT, AND LSTM

Hadiya Ali [1], Maleeha Riaz [2], Imran Ullah [3], Muhammad Saad Hussain [4], Muhammad Abu Bakar Sultan [5]

Affiliations

[1] Department of Computer Science, Abbottabad University of Science and Technology, Abbottabad, Pakistan

[2,4,5] Department of Computer Science, University of Lahore, Sargodha Campus, Pakistan

[3] Department of Computer Science, Hazara University, Dhodial, Mansehra, Pakistan

Corresponding Author's Email

muhammadsaadhussain49@gmail.com

ABSTRACT

Text classification is a fundamental natural language processing task, and transformer-based language models have become the dominant approach. However, large models such as BERT impose substantial computational costs that limit their deployment in resource-constrained settings, motivating interest in lightweight alternatives. This paper presents a comparative analysis of a lightweight transformer (DistilBERT), a full-scale transformer (BERT), a recurrent neural network (bidirectional LSTM), and a classical machine learning baseline (TF-IDF with logistic regression) for binary sentiment classification on the IMDB movie review dataset. All models are evaluated under an identical protocol using a subset of 8000 training and 2000 test reviews, and are compared across accuracy, precision, recall, F1-score, area under the curve, training time, and model size. BERT achieves the highest accuracy of 87.35 percent, followed by the TF-IDF baseline at 86.30 percent and DistilBERT at 85.75 percent, while the bidirectional LSTM reaches 74.50 percent. DistilBERT maintains 98.2 percent of the accuracy of BERT while using 40 percent fewer parameters and training twice as fast. A notable finding is that the classical TF-IDF baseline remains highly competitive under limited data and limited fine-tuning, exceeding DistilBERT in accuracy at a fraction of the computational cost. The results highlight the accuracy-efficiency trade-offs among modern and classical text classification methods and offer practical guidance for model selection.

Keywords: Text Classification, Sentiment Analysis, Transformer, BERT, DistilBERT, LSTM, Fine-Tuning, Natural Language Processing.

I. INTRODUCTION

Text classification is one of the most primitive and basic tasks in Natural Language Processing (NLP) and has many practical applications such as sentiment analysis, spam classification, topic categorisation, and intent recognition [1] [2]. Among the various fields, sentiment analysis has been particularly investigated for its commercial and social impact, allowing organisations to collectively understand the public's opinion from product reviews, social media and customer feedback [3]. The main problem with text classification is that the input texts are variable-length, unstructured, and the task of the classifier is to infer the target label from a representation of such texts; successive advances in the field have been driven by the development of different representations [4].

Initial text classification models were based on sparse bag-of-words and term frequency-inverse document frequency (TF-IDF) representations and used classical ML classifiers like naive Bayes, logistic regression, and support vector machines [5, 6]. While these are easy to calculate, easy to understand and still

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

surprisingly good baselines, they ignore the order of the words in the sentence and have difficulties capturing relationships between words in the sense of the sentence [7]. The semantic limitation was addressed by introducing distributed word representations, such as word2vec [8] and GloVe [9] which represent words as vectors, while sequential modelling of text that captures word order and long-range dependencies was achieved using recurrent neural networks (RNN) especially LSTM network [10].

The most important step was the transformer architecture [12] that abandoned recurrence in favor of self-attention, allowing for very parallel training on massive text corpora. The bidirectional Encoder Representations from Transformers (BERT) [13] revealed that a transformer pre-trained on a masked language modelling task and subsequently fine-tuned on a downstream task is able to attain state-of-the-art results on a broad spectrum of NLP benchmarks. Since then, BERT and its descendants have become the prevailing approach for text classification [14, 31]. But with a large number of parameters (110 million in BERT-base), large memory, and slow fine-tuning and deployment, these models cannot be used in mobile devices or real-time systems with limited resources [15].

Making the transformer small and compact has driven the creation of lightweight transformer models that still achieve high accuracy while being lighter and smaller. DistilBERT [16] is an effort to compress BERT while trying to preserve its ability to understand language with about 97 percent of the capability; this approach was accomplished by using knowledge distillation and reduced the number of parameters by 40 percent, and the running time by 60 percent. Some of the related work includes ALBERT [17] which shares parameters across the layers, and TinyBERT [18] which uses transformer distillation. Although these lightweight models have enjoyed widespread adoption, practitioners have limited empirical understanding of the accuracy-efficiency trade-offs for these lightweight transformers, as well as for full transformers, recurrent models, and classical baselines, under the constrained data and compute budgets often faced by many real-world deployments [19].

In this paper, we aim to fill this gap by conducting a controlled comparative evaluation of four representative text classification approaches from three modelling paradigms: a lightweight transformer (DistilBERT), a full transformer (BERT), a recurrent network (bidirectional LSTM), and a classical baseline (TF-IDF with logistic regression). A standardized training/evaluation protocol is followed for all models using the movie review sentiment data from IMDB [20] in a subset that is suited to a realistic constrained-resource setting. This work has made the following contributions. First, the four models are compared in depth: accuracy/precision/recall/F1-score/Area under the Curve/training time/number of parameters so that a multi-dimensional comparison is made between the four models. Second, the accuracy-efficiency trade-off is explored explicitly, quantifying the accuracy that DistilBERT retains at what saving in computation compared to BERT. Third, the analysis highlights that a classical TF-IDF baseline is still very competitive in the face of limited data and limited fine-tuning, which had a very practical aspect and warns against using large models unthinkingly. The remainder of this paper is organised as follows. The related work is discussed in Section 2. The methodology is described in Section 3. The results are presented and discussed in section 4. In Section 5, conclusions are drawn and suggestions for future work are provided.

II. LITERATURE REVIEW

In this section, literature on the three fields of classical and neural text classification, transformer-based models, and lightweight and efficient transformers is reviewed. Classical text classification pipelines map documents into sparse vectors and use linear or probabilistic classifiers. The bag-of-words model and its TF-IDF weighting scheme [5] are still fundamental, and when used with logistic regression or support vector machines, it gives good and efficient baselines which are still frequently utilized in practice [6, 7, 32]. The combination of naive-bayes features and a support vector machine (SVM) with a carefully tuned set of parameters proved to be a very strong baseline for sentiment classification, and in many cases matched the performance of much more complicated models [21]. Each term is considered as a separate dimension; a drawback of such approaches is they do not account for word order and semantic similarity.

The semantic limitation was solved with distributed representations. Mikolov et al. [8] proposed word2vec that learns dense word embeddings that capture semantic and syntactic regularities while

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

Pennington et al. [9] proposed GloVe method that derives embeddings from global co-occurrence statistics. Deep neural networks have been used on text based on these embeddings. Kim [22] showed that word embedding based CNN achieves good performance for sentence classification. Recurrent neural networks, in particular the LSTM [10] treat text as a sequence, and model long-range dependencies using gated memory cells. Bidirectional LSTMs [11] utilize the sequence in both directions, allowing it to model future and past context, and have been widely used in sentiment analysis and other classification problems. Recurrent models do not discard order information like bag-of-words models do, but they are trained sequentially, which limits parallelism, and they suffer from loss of information when the document they are trained on is very long [12].

Vaswani et al. [12] proposed the transformer architecture which eliminated the recurrence, instead modelling the relationship between all the positions in a sequence in parallel using the multi-head self-attention mechanism. With this design, they managed to efficiently train on very large corpora, and they were able to capture long-range dependencies more effectively than recurrent models. In two stages paradigm, BERT [13] used the transformer encoder for unsupervised pre-training with masked language modelling and next sentence prediction on a large text corpus, followed by supervised fine-tuning to the downstream task. This paradigm resulted in significant gains in the GLUE benchmark and led to the adoption of the pre-training and fine-tuning paradigm in NLP.

Many variations in the transformer were to follow. RoBERTa [23] made modifications to BERT: it replaced the next-sentence-prediction objective with more extensive pre-training. XLNet [24] proposed an objective function based on permutations and the GPT family [25] showed the effectiveness of large autoregressive transformers for generation and few-shot learning. In the case of text classification, fine-tuning a pre-trained transformer (e.g., BERT) with a simple classification head on the pooled representation has become the standard and very efficient way of doing things [14]. Sun et al. [26] investigated best practice for fine tuning BERT for text classification, including the choice of learning rate, which layers to use and how to prevent catastrophic forgetting. Although these models are accurate, they are also computing intensive, and their size is a problem when deploying them in edge devices or in latency sensitive application [15].

The vast amount of work in model compression and efficient architecture [19] is motivated by the computational cost of large transformers. In that regard, Knowledge distillation [27] is applied to train a small student model that mimics the behavior of a large teacher, and DistilBERT [16] uses distillation during pretraining to compress BERT from 12 to 6 transformer layers, preserving ~97% of its performance while being 40% smaller and 60% faster. ALBERT [17, 33] optimises the parameters by cross-layer parameter sharing and factorised embedding parameterisation, while TinyBERT [18] proposes a two-stage transformer distillation framework that further distills the ALBERT model. MobileBERT [28] adapts the model with bottleneck designs for mobile use.

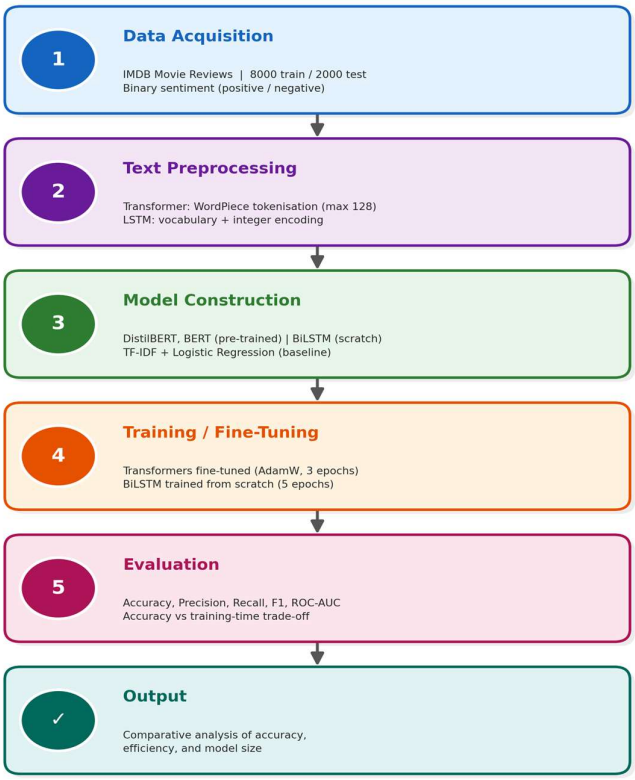
Comparative assessments of the models can be useful in determining which model to use. Other classic and simple neural baselines were revisited by Adhikari et al. [29] for classification of documents and they were found to be competitive with the performance of complex models, which highlights the power of these well-tuned baselines. A few works have been comparing the accuracy of transformers with recurrent and classical models on sentiment datasets, and it is generally accepted that transformers are more accurate but come with a price [30]. But many comparisons are made with sufficient data and lots of fine tuning, which is not the case for many actual deployment efforts. The relative performance of lightweight transformers, full transformers, recurrent networks and classical baselines is less well-understood under limited data and few fine-tuning epochs. The present work provides a controlled comparison, comparing all four paradigms under the same constrained protocol, and analysing the accuracy-efficiency trade-off explicitly, providing practical guidance in addition to the large-data evaluations that dominate the literature.

III. PROPOSED METHODOLOGY

The proposed comparative framework consists of six stages: data acquisition, text preprocessing, model construction, training and fine-tuning, evaluation, and comparative analysis. Fig. 1 presents the overall architecture as a flowchart. Each stage is described in the following subsections.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

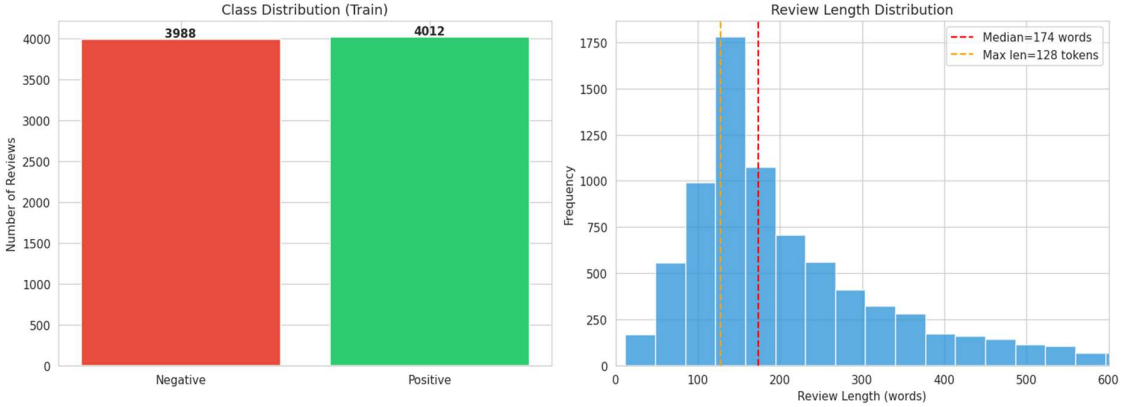
Fig. 1: ARCHITECTURE OF THE PROPOSED COMPARATIVE METHODOLOGY FOR TEXT CLASSIFICATION. THE PIPELINE PROCEEDS FROM DATA ACQUISITION AND PREPROCESSING THROUGH MODEL CONSTRUCTION, TRAINING, AND EVALUATION TO A COMPARATIVE ANALYSIS OF ACCURACY, EFFICIENCY, AND MODEL SIZE.



A. Dataset and Preprocessing

The study uses the IMDB movie review dataset [20], a widely used benchmark for binary sentiment classification containing 50000 reviews evenly split between positive and negative sentiment. To reflect a realistic resource-constrained setting and to keep transformer fine-tuning tractable, a stratified subset of 8000 training and 2000 test reviews is used, preserving the balanced class distribution. Fig. 2 shows the class balance and the distribution of review lengths, which motivates the choice of a maximum sequence length of 128 tokens.

Fig. 2: DATASET OVERVIEW. LEFT: BALANCED CLASS DISTRIBUTION OF THE TRAINING SUBSET. RIGHT: DISTRIBUTION OF REVIEW LENGTHS IN WORDS, WITH THE MEDIAN AND THE 128-TOKEN MAXIMUM SEQUENCE LENGTH MARKED.



JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

Preprocessing differs by model family. For the transformer models, text is tokenised using the WordPiece tokeniser associated with each pre-trained model, truncated or padded to a fixed length of 128 tokens, and converted into token identifiers and attention masks. For the bidirectional LSTM, a vocabulary of the 20000 most frequent tokens is constructed from the training set, and each review is converted into a sequence of integer indices, truncated or padded to 128 tokens. For the classical baseline, reviews are transformed into TF-IDF vectors over unigrams and bigrams with a vocabulary of 20000 features and English stop words removed. The TF-IDF weight of term t in document d is

$$tfidf(t, d) = tf(t, d) \times \log(N / df(t)) \quad (1)$$

where tf(t, d) is the frequency of term t in document d, N is the total number of documents, and df(t) is the number of documents containing t.

B. Transformer Models

BERT and DistilBERT are transformer encoders built on the multi-head self-attention mechanism [12]. For an input sequence, self-attention computes a weighted combination of value vectors, where the weights are derived from the compatibility of query and key vectors

$$Attention(Q, K, V) = softmax(QK^T / \sqrt{d_k}) V \quad (2)$$

where Q, K, and V are the query, key, and value matrices and d_k is the dimension of the key vectors. Multiple attention heads operate in parallel and their outputs are concatenated, allowing the model to attend to information from different representation subspaces. BERT-base stacks twelve such transformer layers with 110 million parameters, while DistilBERT uses six layers with 66 million parameters, obtained by knowledge distillation from BERT [16]. Both models are pre-trained on large corpora and are fine-tuned here for sentiment classification by adding a linear classification head on top of the pooled representation of the special classification token. Fine-tuning minimises the cross-entropy loss

$$L = - \Sigma \Sigma \, y_{ic} \log(\hat{y}_{ic}) \quad (3)$$

where y_{ic} is the true label and $\hat{y}_{ic}$ the predicted probability of class c for review i, optimised using the AdamW optimiser with a linear learning-rate schedule.

C. Bidirectional LSTM

The bidirectional LSTM processes the embedded token sequence in both forward and backward directions. An LSTM cell maintains a memory state regulated by input, forget, and output gates, computed at each time step t as

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (4)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \quad o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c \cdot [h_{t-1}, x_t] + b_c), \quad h_t = o_t \odot \tanh(c_t) \quad (6)$$

where σ is the sigmoid function and ⊙ denotes elementwise multiplication. The final forward and backward hidden states are concatenated and passed through a dropout layer and a linear classifier. The model uses a trainable embedding layer of dimension 100, a hidden dimension of 128 per direction, and is trained from scratch on the training subset. Algorithm 1 summarises the full experimental procedure.

Algorithm 1: Comparative Text Classification Procedure

Input: IMDB subset (X_train, y_train), (X_test, y_test)

Output: performance metrics for all models

1: for model in {DistilBERT, BERT} do

2: tokenize text with model WordPiece tokenizer (len 128)

3: add linear classification head

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

```
4:    fine-tune with AdamW for E epochs (cross-entropy)
5:    evaluate on test set
6:  build vocabulary from X_train (top 20000 tokens)
7:  encode text as integer sequences (len 128)
8:  train BiLSTM from scratch for E_L epochs
9:  evaluate BiLSTM on test set
10: fit TF-IDF (unigram+bigram) on X_train
11: train logistic regression on TF-IDF features
12: evaluate baseline on test set
13: compare accuracy, F1, AUC, time, and model size
```

D. Experimental Configuration and Metrics

All models are trained and evaluated on an identical data split using an NVIDIA T4 GPU. The transformer models are fine-tuned for three epochs and the bidirectional LSTM is trained for five epochs. Table I summarises the configuration. Performance is measured using accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC), together with training time and parameter count to capture the efficiency dimension. The F1-score is the harmonic mean of precision and recall

$$F1 = 2 \cdot (Precision \cdot Recall) / (Precision + Recall) \quad (7)$$

Table 1: EXPERIMENTAL CONFIGURATION.

Parameter	Value
Training reviews	8000
Test reviews	2000
Max sequence length	128 tokens
Batch size	16
Transformer epochs	3
LSTM epochs	5
Transformer learning rate	2×10^{-5} (AdamW)
LSTM / embedding / hidden	100 / 128
TF-IDF features	20000 (unigram+bigram)
Hardware	**NVIDIA T4 GPU**

IV. RESULTS AND DISCUSSION

This section reports the classification performance of the four models, the accuracy-efficiency trade-off, and a discussion of the findings. All results use the identical test set of 2000 reviews.

A. Classification Performance

Table II reports the performance of all four models. BERT achieves the highest accuracy of 87.35 percent with an F1-score of 0.875 and an AUC of 0.949. The TF-IDF baseline is the second best at 86.30 percent accurately, narrowly ahead of DistilBERT at 85.75 percent, while the bidirectional LSTM attains 74.50 percent. Fig. 3 compares the models across all five metrics, and Fig. 4 presents the confusion matrices. The transformer and TF-IDF models are clustered within a narrow band of accuracy between 85.75 and 87.35 percent, whereas the LSTM trained from scratch lags substantially, reflecting the difficulty of learning useful sequential representations from a limited training set without pre-training.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

Fig 3: PERFORMANCE COMPARISON OF DISTILBERT, BERT, BILSTM, AND THE TF-IDF BASELINE ACROSS ACCURACY, PRECISION, RECALL, F1-SCORE, AND AUC. THE THREE STRONGEST MODELS ARE CLOSELY CLUSTERED, WHILE THE FROM-SCRATCH BILSTM LAGS.

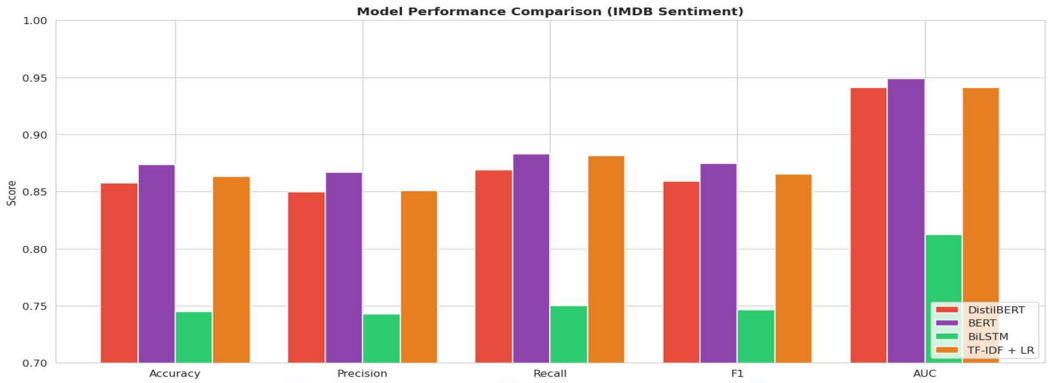


Table 2: CLASSIFICATION PERFORMANCE AND EFFICIENCY OF ALL MODELS ON THE IMDB TEST SET

Model	Acc (%)	F1	AUC	Time (s)	Parameters
BERT	**87.35**	**0.875**	**0.949**	**564**	**110,000,000**
TF-IDF + LR	86.30	0.865	0.941	5	20,000
DistilBERT	85.75	0.859	0.941	281	66,000,000
BiLSTM	74.50	0.746	0.812	10	2,236,034

Fig. 4: CONFUSION MATRICES ON THE TEST SET OF 2000 REVIEWS. BERT AND THE TF-IDF BASELINE PRODUCE THE MOST BALANCED ERROR PROFILES, WHILE THE BILSTM SHOWS SUBSTANTIALLY MORE MISCLASSIFICATIONS

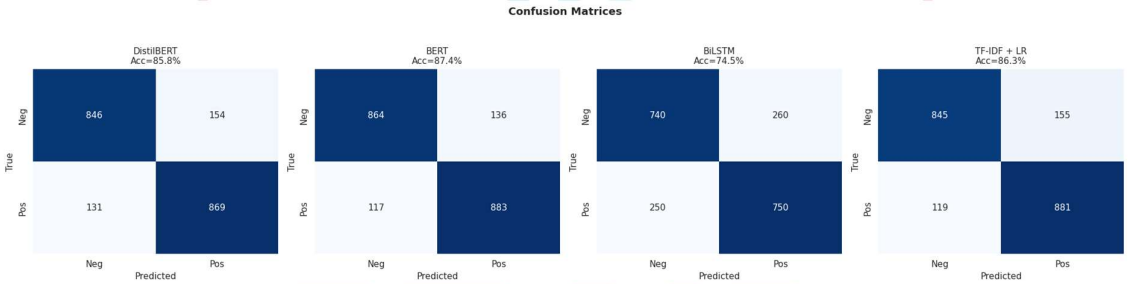
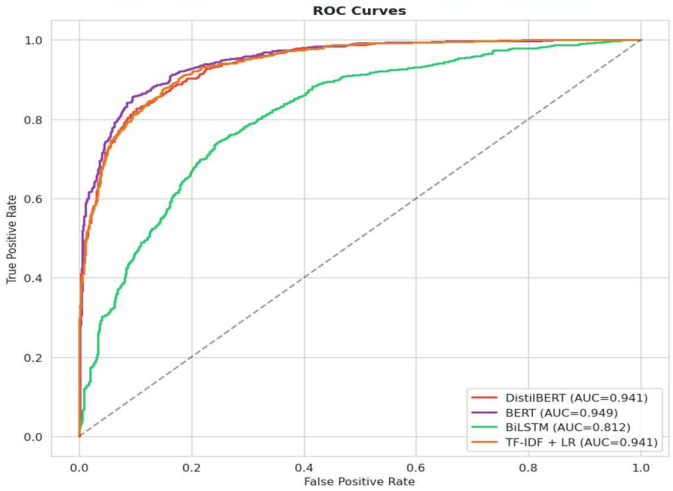


Fig. 5: RECEIVER OPERATING CHARACTERISTIC CURVES FOR ALL MODELS. BERT ACHIEVES THE HIGHEST AUC OF 0.949, with DISTILBERT AND THE TF-IDF BASELINE CLOSE BEHIND AT 0.941 AND THE BILSTM LOWEST AT 0.812.



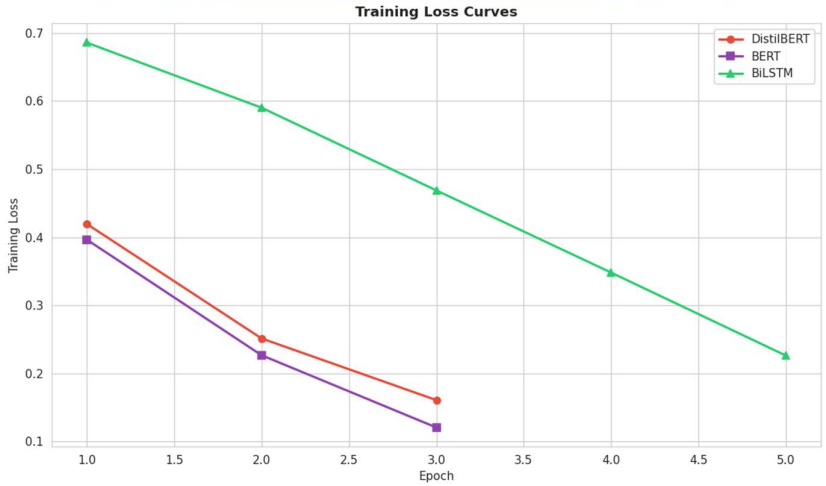
JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

B. Accuracy-Efficiency Trade-Off

The efficiency dimension reveals the central trade-off of this study. Fig. 6 plots test accuracy against training time, and Fig. 7 shows the training loss curves. DistilBERT retains 98.2 percent of the accuracy of BERT (85.75 versus 87.35 percent) while using 40 percent fewer parameters and training in 281 seconds compared to 564 seconds for BERT, a twofold speed-up. This confirms that knowledge distillation preserves most of the predictive capability of the full transformer at a substantially reduced cost, making DistilBERT an attractive choice when compute or latency is constrained.

Fig. 6: ACCURACY VERSUS TRAINING TIME TRADE-OFF. DISTILBERT AND BERT OCCUPY THE HIGH-ACCURACY REGION AT HIGHER TRAINING COST, WHILE THE TF-IDF BASELINE ACHIEVES COMPARABLE ACCURACY AT NEGLIGIBLE COST, AND THE BILSTM IS DOMINATED ON BOTH AXES.

Fig. 7: TRAINING LOSS CURVES. THE TRANSFORMER MODELS CONVERGE TO LOWER LOSS THAN THE BILSTM, AND BERT REACHES THE LOWEST FINAL LOSS, CONSISTENT WITH ITS HIGHEST TEST ACCURACY.



The most striking result, however, concerns the classical baseline. The TF-IDF and logistic regression model achieves 86.30 percent accuracy, exceeding DistilBERT and approaching BERT, while training in only 5 seconds with a negligible model footprint. This outcome, while counter to the common assumption that transformers dominate all text classification settings, is consistent with prior observations that well-tuned classical baselines remain highly competitive [21], [29], particularly under limited data and limited fine-tuning. With only 8000 training reviews and three fine-tuning epochs, the transformers do not reach their full potential, whereas TF-IDF with logistic regression is well suited to the strong lexical

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

sentiment cues present in movie reviews. This finding has practical significance: for constrained-resource sentiment tasks, a classical baseline may deliver near-transformer accuracy at a tiny fraction of the cost.

C. Discussion

The overall findings demonstrate that when selecting a model for text classification, it is important to consider the context of the system's operation, not necessarily the "bigger is better" approach. If accuracy is the highest priority, and compute is available, BERT is the best option. DistilBERT provides the best of both worlds - if compute or latency is a concern but transformer-style accuracy is required, DistilBERT is an excellent choice, with nearly the same accuracy as BERT but half the training cost. Where resources are scarce or speed or interpretability are primary concerns, the TF-IDF baseline is extremely competitive and can't be ignored. In this case, the from-scratch bidirectional LSTM is outperformed in both accuracy and efficiency, which illustrates the value of pre-training: without such pre-training, a recurrent model can no longer learn to ensure representations are adequate given the limited size of the data set. A drawback to this study is that a smaller training set and fewer epochs of fine-tuning have been employed, limiting the transformers and thus boosting the classical baseline. This is by design, as it represents an actual resource-constrained environment, but it does result in the absolute accuracy of the transformers reported here being less than what could be achieved with the full data sets and longer training.

V. CONCLUSION AND FUTURE WORK

In this paper, a controlled comparative analysis of DistilBERT, BERT, TF-IDF baseline and bidirectional LSTM for binary sentiment classification over IMDB dataset is presented under a resource constrained protocol. BERT was the most accurate at 87.35 per cent, closely followed by TF-IDF baseline at 86.30 per cent and DistilBERT at 85.75 per cent and the from-scratch bidirectional LSTM at 74.50 per cent. DistilBERT reduced the number of parameters by 40% and doubled the training speed, while still maintaining 98.2% of the accuracy of BERT, which shows the advantages of the lightweight DistilBERT. The main takeaway was that the classical TF-IDF baseline was still very competitive under limited data and limited fine-tuning, outperforming DistilBERT and coming close to BERT at minimal computational cost, which calls into question the uncritical use of large models in limited situations. These results collectively give empirical recommendations for selecting text classification models based upon the desired accuracy, efficiency, and interpretability of the specific application.

There are several directions for future work. The most straightforward is to increase the size of the training data and the number of fine-tuning epochs, here they were intentionally limited, as doing so would likely further the gap between the accuracy scores of the four paradigms and better demonstrate the transformers' potential; it would be helpful to experiment and see how the relative ranking of the four paradigms changes as data and compute budget increases. Additional directions include expanding the comparison to other lightweight architectures like ALBERT, TinyBERT and MobileBERT, initialising the recurrent model with pre-trained word-embeddings to isolate the contribution of pre-training, and adding explainability to the predictions of the transformer models, to make their predictions more transparent to deploy.

REFERENCES

[1] C. C. Aggarwal and C. Zhai, “A survey of text classification algorithms,” in Mining Text Data, Springer, 2012, pp. 163–222.
[2] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, “Text classification algorithms: A survey,” Information, vol. 10, no. 4, p. 150, 2019.
[3] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” Foundations and Trends in Information Retrieval, vol. 2, no. 1–2, pp. 1–135, 2008.
[4] B. Liu, Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge University Press, 2015.
[5] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” Information Processing and Management, vol. 24, no. 5, pp. 513–523, 1988.

[6] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in Proc. ECML, 1998, pp. 137–142.
[7] S. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in Proc. ACL, 2012, pp. 90–94.
[8] M. Chaudhari, "Spray-Deposited Fluorine-Free Polymer Coatings via Fatty-Acid/Oxide Interphases for Extreme Water Repellency," in Proceedings of the International Conference on Sustainable Micro-Nano Materials & Innovative Technology (ICSUMMIT 2026), Vadodara, India, Feb. 13–14, 2026, Atlantis Highlights in Engineering, vol. 44, Atlantis Press, Jul. 2026. doi: 10.2991/978-94-6239-727-9_26.
[9] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in Proc. EMNLP, 2014, pp. 1532–1543.
[10] F. Ahmed, "Cloud Security Posture Management (CSPM): Automating Security Policy Enforcement in Cloud Environments," *ESP International Journal of Advancements in Computational Technology*, vol. 1, no. 3, pp. 157–166, 2023. https://philarchive.org/rec/AHMCSP
[11] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," Neural Networks, vol. 18, no. 5–6, pp. 602–610, 2005.
[12] S. T. Hasan, "Graph neural network-based optimization for home health aide-patient matching," *Spanish Journal of Innovation and Integrity*, vol. 54, pp. 268–281, 2026.
[13] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
[14] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?," in Proc. China National Conf. on Chinese Computational Linguistics, 2019, pp. 194–206.
[15] S. T. Hasan, "AI-driven equity analytics in New York City home care workforce management: Identifying bias in scheduling, case assignment, workload distribution, and career advancement," *Journal of Business Insight and Innovation*, vol. 4, no. 2, pp. 119–131, 2025.
[16] M. Chaudhari, "Vapor-Phase Thin-Film Polymerization for Microbial-Resistant Surface Functionalization," in Proceedings of the International Conference on Sustainable Micro-Nano Materials & Innovative Technology (ICSUMMIT 2026), Vadodara, India, Feb. 2026, doi: 10.2991/978-94-6239-727-9_27.
[17] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," in Proc. ICLR, 2020.
[18] X. Jiao et al., "TinyBERT: Distilling BERT for natural language understanding," in Findings of EMNLP, 2020, pp. 4163–4174.
[19] M. A. Jawed, "Managed aquifer recharge with advanced treated municipal effluent: Modelling the impact on groundwater chemistry and native microbial ecology," International Journal of Research & Technology, vol. 12, no. 3, pp. 128–138, 2024.
[20] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in Proc. ACL, 2011, pp. 142–150.
[21] S. Wang and C. D. Manning, "Fast dropout training," in Proc. ICML, 2013, pp. 118–126.
[22] M. A. Jawed, "Uncertainty-aware geostatistical reconstruction of regional groundwater hydraulics under sparse monitoring conditions: A case study of the Evergreen Underground Water Conservation District (EUWCD), Texas, USA," Bishop International Journal of Mathematics and Computer Science, vol. 1, no. 1, pp. 57–71, 2025.
[23] U. Iqbal, "AI-enhanced network optimization for electric vehicle charging infrastructure expansion in the United States using graph theory and demand analytics," *Journal of Engineering and Computational Intelligence Review*, vol. 2, no. 2, pp. 112–129, 2024.
[24] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in Proc. NeurIPS, 2019, pp. 5753–5763.
[25] T. Brown et al., "Language models are few-shot learners," in Proc. NeurIPS, 2020, pp. 1877–1901.

[26] C. Sun, L. Huang, and X. Qiu, “Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence,” in Proc. NAACL-HLT, 2019, pp. 380–385.
[27] U. Iqbal, "AI-powered supplier risk intelligence: Predicting financial and geopolitical supply chain disruptions in U.S. critical industries," *Journal of Engineering and Computational Intelligence Review*, vol. 3, no. 2, pp. 173–193, 2025.
[28] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, “MobileBERT: A compact task-agnostic BERT for resource-limited devices,” in Proc. ACL, 2020, pp. 2158–2170.
[29] A. Adhikari, A. Ram, R. Tang, and J. Lin, “Rethinking complex neural network architectures for document classification,” in Proc. NAACL-HLT, 2019, pp. 4046–4051.
[30] M. Munikar, S. Shakya, and A. Shrestha, “Fine-grained sentiment classification using BERT,” in Proc. Int. Conf. on Artificial Intelligence for Transforming Business and Society, 2019.
[31] M. Asif and M. Rafiq-uz-Zaman, “The silent disengagement: A quantitative analysis of leadership, recognition, and workload as predictors of quiet quitting among knowledge workers,” *Al-AASAR Journal*, vol. 3, no. 1, pp. 271–303, 2026, doi: 10.63878/aaj1525.
[32] M. Asif, “Financial performance of startups linked to universities: Evidence from developing economies,” *Journal of Applied Linguistics and TESOL*, vol. 8, no. 3, pp. 2736–2763, 2025, doi: 10.63878/jalt2003.
[33] M. Asif, A. Ali, and F. A. Shaheen, “Assessing the effects of artificial intelligence in revolutionizing human resource management: A systematic review,” *Social Science Review Archives*, vol. 3, no. 4, pp. 2887–2908, 2025, doi: 10.70670/sra.v3i3.1055.