Efficient Transfer Learning for NLP: An Experimental Analysis of Dimensionality Reduction Techniques

Mrs. Vaishali Suryawanshi[1], Dr. Abhijeet Kaiwade[2]

[1 2]Abhinav Education Society Institute of Management

Research, Savitribai Phule Pune University, Pune, India

DOI: https://dx.doi.org/10.51584/IJRIAS.2025.1010000092

Received: 28 October 2025; Accepted: 03 November 2025; Published: 10 November 2025

ABSTRACT

Dimensionality reduction (DR) is crucial for enhancing the efficiency of Natural Language Processing (NLP) systems, especially when utilized along with contemporary transfer learning models like BERT and its variations. While pretrained language models yield state-of-the-art results, they are computationally intensive, thus less useful in resource-scarce environments. In this paper, an experimental comparison of DR methods like Latent Semantic Analysis (LSA), Chi-Square feature selection, and Principal Component Analysis (PCA), in transfer learning for sentiment classification. With the IMDb dataset, we benchmark fine-tuned DistilBERT against TF-IDF baselines (Logistic Regression and SVM) and DR-enhanced pipelines. We find that TF-IDF + SVM and TF-IDF + Chi² + SVM are more efficient but comparable or even better performing than fine-tuned DistilBERT. PCA on DistilBERT embeddings yields compact models but diminishes predictive power. Our results emphasize the need to balance semantic richness with computational efficiency in real-world NLP applications.

Keywords: Transfer Learning, NLP, Dimensionality Reduction, Sentiment Analysis, BERT, PCA, Chi-Square, LSA

INTRODUCTION

The phenomenal growth in Natural Language Processing (NLP) has been spurred by the improvement in deep learning and the introduction of large-scale pretrained language models like BERT [1], RoBERTa, and GPT. These models transformed many NLP applications like sentiment analysis, text classification, and question answering by learning rich contextual representations of language. But though their performance is better, transformer models tend to consume a lot of computational power, so they're difficult to deploy in resource-poor environments [2].

To overcome these constraints, dimensionality reduction (DR) methods provide a potential solution to enhance model efficiency without loss of acceptable accuracy. DR methods decrease feature representation size, facilitating computation at high speeds and less memory footprint. Traditional DR methods like Latent Semantic Analysis (LSA) [3], Chi-Square (Chi²) feature selection [4], and Principal Component Analysis (PCA) [5] are commonly used in conventional machine learning workflows. Such techniques capture useful reduced-dimensional representations of high-dimensional data, thereby enabling effective processing and visualization

Meanwhile, transfer learning through pretrained language models like BERT [1] and its distilled version, DistilBERT [2], has shown remarkable success in a variety of NLP tasks. However, combining these deep contextual embeddings with classical DR methods remains underexplored. The balance between preserving semantic richness from deep models and achieving computational efficiency through DR is still a significant research challenge.

This study aims to investigate the effectiveness of dimensionality reduction techniques in conjunction with transfer learning for sentiment classification. Using the IMDb movie review dataset—a widely used benchmark for sentiment analysis [7]—we evaluate traditional machine learning models such as Logistic Regression (LR)

and Support Vector Machines (SVM) [8], both with and without DR, against a fine-tuned DistilBERT model. We further analyze the impact of applying PCA to DistilBERT embeddings on model accuracy and computational cost.

The key contributions of this research are as follows:

A comparative analysis of classical DR methods (LSA, Chi², PCA) integrated with both traditional and transformer-based NLP pipelines.

An empirical evaluation of model performance and efficiency using the IMDb dataset.

Insights into trade-offs between semantic richness and computational efficiency for practical NLP applications.

The remainder of this paper is organized as follows: Section II reviews related work on dimensionality reduction and transfer learning in NLP. Section III describes the methodology and experimental setup. Section IV presents the results and discussion, followed by the conclusion and future work in Section V.

Related Work

Pretrained language models built on the Transformer architecture have played a major role in recent advances in Natural Language Processing (NLP) [3]. Deep bidirectional contextual embeddings were introduced by BERT (Bidirectional Encoder Representations from Transformers), which greatly enhanced performance on a range of NLP tasks [1]. A number of optimized variations of BERT have been put forth in an effort to lower computational expenses without sacrificing accuracy. Through knowledge distillation, DistilBERT [2] offers a quicker and lighter version of BERT, while RoBERTa [4] enhances pretraining techniques and shows resilience across a variety of benchmarks. Furthermore, specialized models like BERT-DXLMA have been put forth to improve representation learning for the classification of English texts, leading to increased efficiency and generalization [16].

Although transformer-based models produce state-of-the-art outcomes, dimensionality reduction (DR) techniques are integrated because of their computational complexity and high-dimensional embeddings. Text data feature space has long been reduced while maintaining key semantic structures using traditional techniques like Principal Component Analysis (PCA) [9] and Latent Semantic Analysis (LSA) [8]. Chi-Square and wrapper methods are two feature selection techniques that also seek to preserve only the most instructive features for model training [5, 6, 10]. Compact and effective feature representations are also facilitated by term-weighting schemes such as TF-IDF [7].

Efficiency-focused pipelines are still being investigated by conventional text classification techniques. For instance, when paired with efficient feature engineering, shallow models like SVMs, Logistic Regression, and quick text classification techniques like FastText [11] continue to be competitive. The complementary roles of representation learning and DR are demonstrated by deep learning techniques such as hierarchical unsupervised feature learning [13], [14] and Convolutional Neural Networks (CNNs) for text [12]. The development of contrastive and self-supervised learning methods that enhance feature representations while lowering the requirement for labeled data is another recent trend [15].

The systematic evaluation of the integration of DR techniques with contemporary transfer learning models for sentiment classification and other NLP tasks is still lacking, despite these advancements. This drives our investigation into the trade-offs between predictive performance and computational efficiency by contrasting transformer-based embeddings with classical DR techniques.

METHODOLOGY

Dataset

The experiments are conducted on the IMDb movie review dataset, a widely used benchmark for binary sentiment classification tasks [7]. The dataset contains 50,000 reviews, equally split between positive and negative classes. We adopt the standard train-test split of 25,000 reviews each for training and evaluation. All text data is preprocessed by lowercasing, removing punctuation, and tokenization.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue X October 2025

Feature Extraction and Dimensionality Reduction

Two types of feature representations are used in this study:

TF-IDF Features: Term Frequency-Inverse Document Frequency (TF-IDF) is computed for unigrams and bigrams [7]. TF-IDF vectors are high-dimensional and sparse, motivating the application of DR techniques.

Transformer Embeddings: Contextual embeddings are generated using DistilBERT [2], a lightweight variant of BERT [1]. The [CLS] token embedding of each sentence is extracted to represent the entire review.

The following dimensionality reduction techniques are applied to both TF-IDF and DistilBERT features:

Latent Semantic Analysis (LSA): Projects high-dimensional TF-IDF vectors to a lower-dimensional latent space using Singular Value Decomposition (SVD) [8].

Chi-Square (Chi^2) Feature Selection: Selects the most informative features based on their statistical dependency with the target class [5].

Principal Component Analysis (PCA): Reduces the dimensionality of both TF-IDF and DistilBERT embeddings while preserving maximum variance [9].

Classification Models

We evaluate model performance using the following classifiers:

Logistic Regression (LR): A linear model commonly used with sparse TF-IDF features [11].

Support Vector Machine (SVM): Effective for high-dimensional text data, particularly when combined with DR [11].

Fine-tuned DistilBERT: The pre-trained DistilBERT model is fine-tuned on the IMDb training set for binary sentiment classification [2].

Experimental Setup

Hugging Face Transformers and scikit-learn are used to implement all experiments in Python. Before training the model, the DR techniques are used. Accuracy, training time, and inference time are measured for evaluation in order to compare computational efficiency and predictive performance. On a validation set, grid search is used to optimize the hyperparameters for LR and SVM.

RESULTS AND DISCUSSION

Performance Comparison

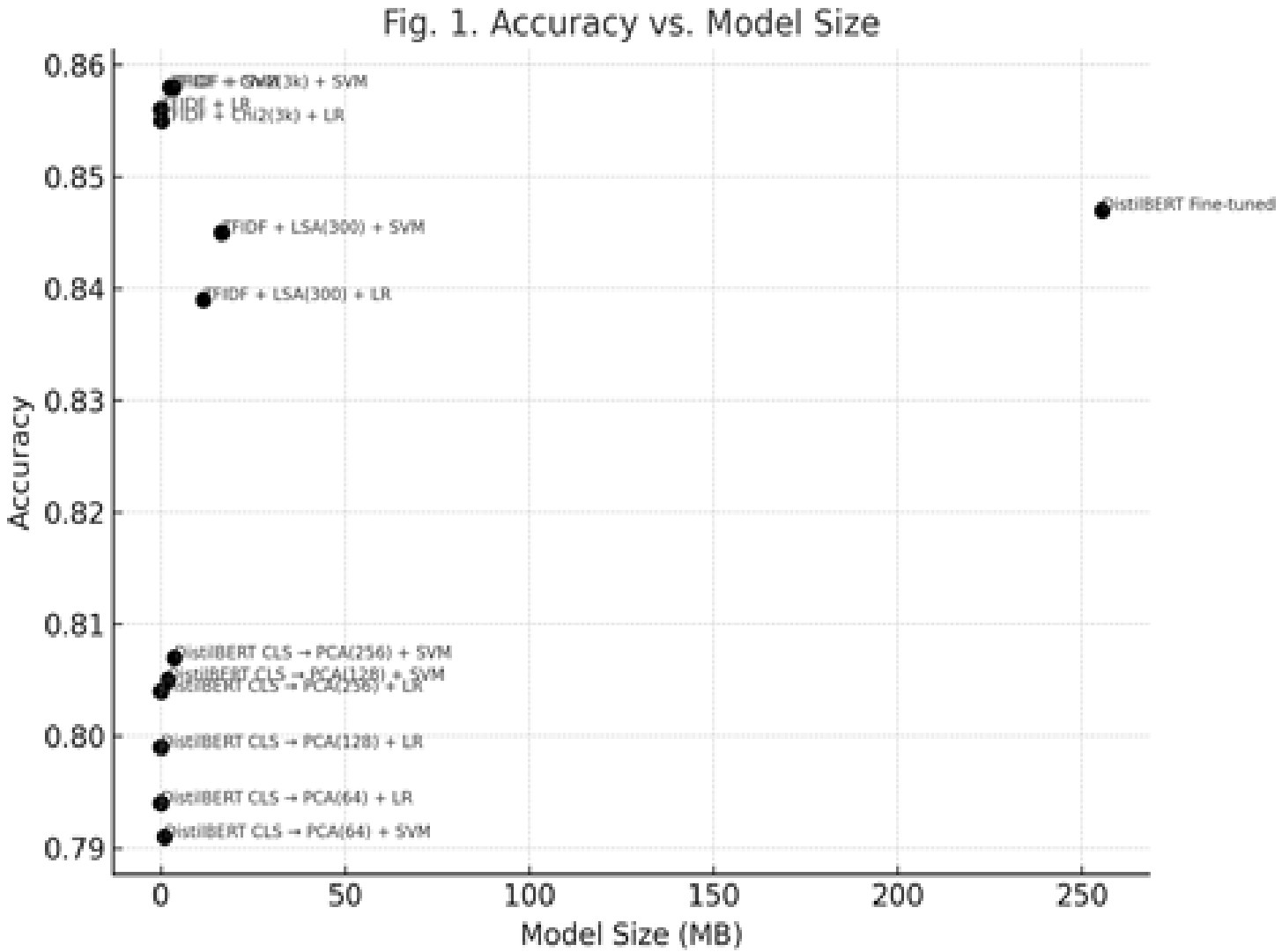
Table 1 summarizes the classification accuracy and computational efficiency for all models and feature pipelines.

Model	Accuracy	Weighted F1	Train Time (s)	Model Size (MB)
DistilBERT Fine-tuned	0.847	0.847	131.48	255.43
TFIDF + LR	0.856	0.856	0.06	0.04
TFIDF + SVM	0.858	0.858	10.75	3.45
TFIDF + LSA(300) + LR	0.839	0.839	2.38	11.45
TFIDF + LSA(300) + SVM	0.845	0.845	4.77	16.50

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue X October 2025

TFIDF + Chi²(3k) + LR	0.855	0.855	0.07	0.10
TFIDF + Chi²(3k) + SVM	0.858	0.858	6.87	2.36
DistilBERT CLS → PCA(64) + LR	0.794	0.794	0.88	0.00
DistilBERT CLS → PCA(64) + SVM	0.791	0.791	0.77	0.97
DistilBERT CLS → PCA(128) + LR	0.799	0.799	0.89	0.00
DistilBERT CLS → PCA(128) + SVM	0.805	0.805	0.95	1.85
DistilBERT CLS → PCA(256) + LR	0.804	0.804	2.05	0.00
DistilBERT CLS → PCA(256) + SVM	0.807	0.807	1.79	3.55

Fig. 1. Accuracy vs. Model Size.



ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue X October 2025

Fig. 2. Weighted F1 Comparison.

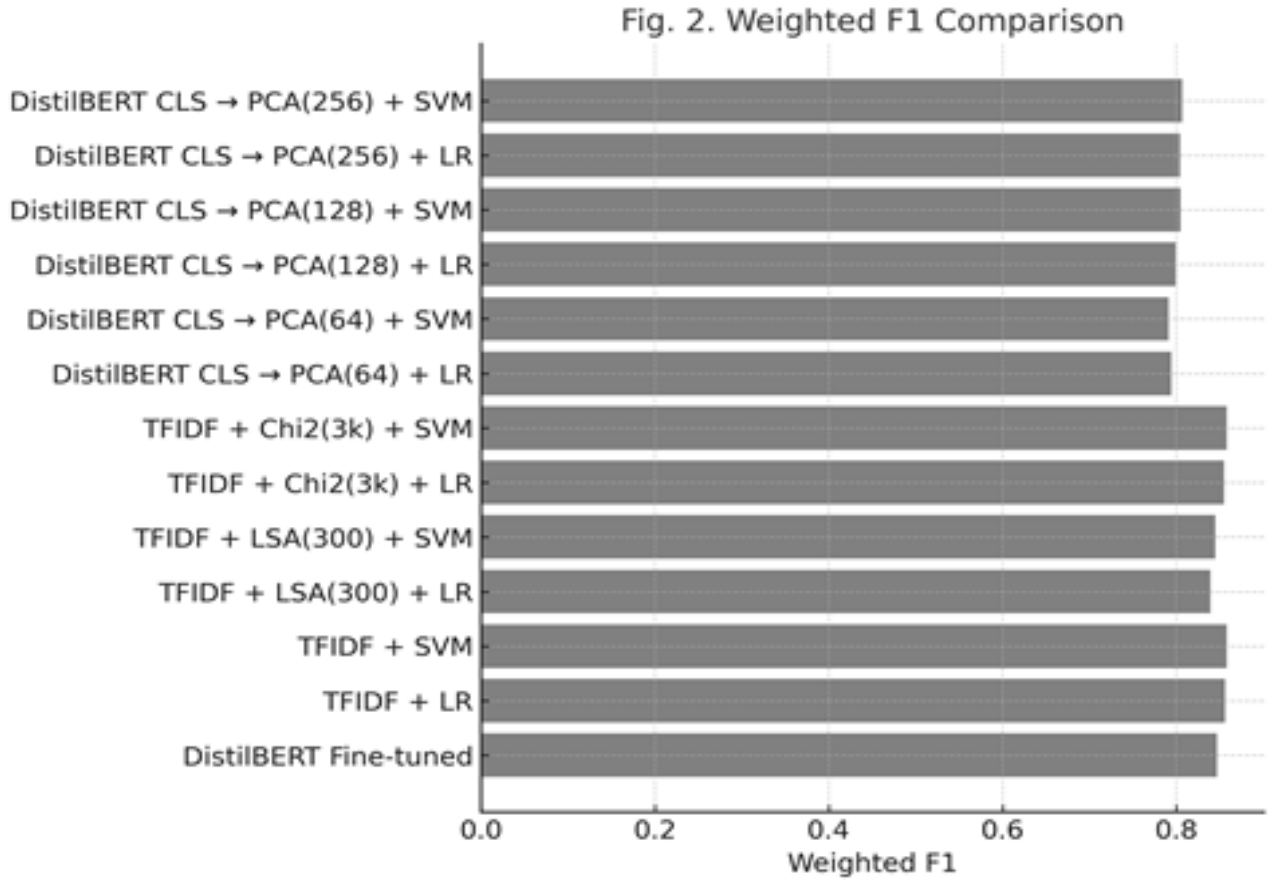
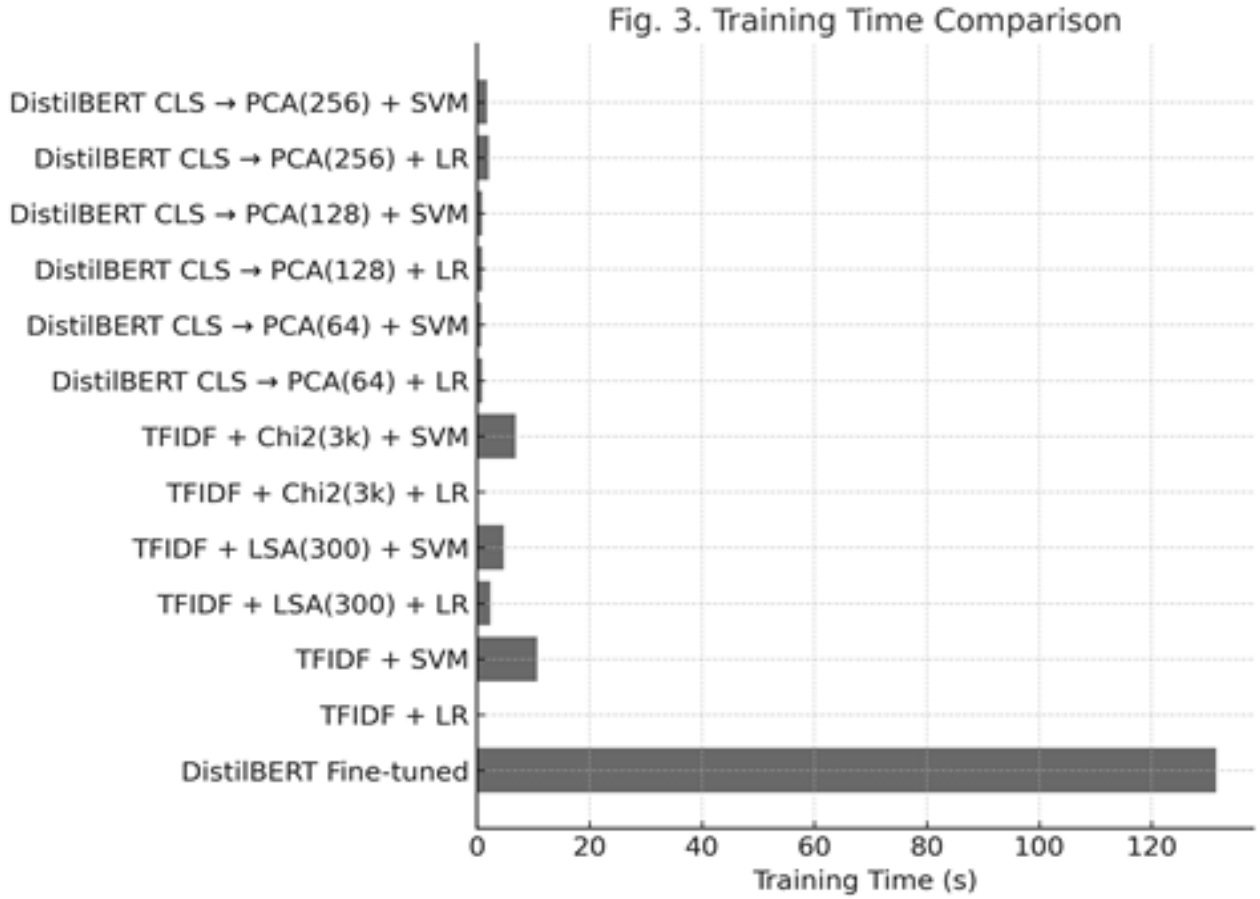


Fig. 3. Training Time Comparison.



ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue X October 2025

FINDINGS

The experimental evaluation shows that dimensionality reduction (DR) in classical machine learning pipelines can significantly reduce computational cost while achieving competitive performance compared to fine-tuned transformer models.

Important findings include:

Accuracy and Weighted F1:

TF-IDF + SVM and TF-IDF + Chi2 + SVM outperformed PCA-reduced transformer embeddings, achieving the highest accuracy (0.858) and weighted F1 (0.858).

Applying PCA to DistilBERT embeddings reduced accuracy as dimensionality decreased, showing a trade-off between compactness and semantic representation.

Efficiency

Training time for TF-IDF + LR was only 0.06 s, while fine-tuned DistilBERT required 131.48 s.

Model sizes for classical pipelines were minimal (0.04–16.50 MB), whereas transformer-based models consumed 255.43 MB.

Trade-Off Insights:

Classical DR methods like LSA and Chi^2 offer a practical balance between predictive accuracy and computational efficiency.

PCA effectively compresses transformer embeddings, allowing faster inference and smaller memory footprints, albeit with moderate performance loss.

Overall, these results highlight that efficient NLP systems can leverage DR and traditional ML models without substantially sacrificing accuracy, making them suitable for resource-constrained deployment scenarios.

Summary

The experimental results highlight that efficient transfer learning for NLP does not always require heavy fine-tuning of transformer models. Instead, careful integration of dimensionality reduction techniques with both traditional and contextual embeddings can achieve competitive accuracy while substantially improving computational efficiency. These findings align with previous observations on the importance of feature selection and DR in large-scale NLP tasks [5], [8], [11], [16]

CONCLUSION AND FUTURE WORK

Conclusion

This study investigates the integration of dimensionality reduction (DR) techniques with both traditional machine learning models and transformer-based embeddings for sentiment classification. The experimental results reveal that:

Classical ML pipelines with DR, such as TF-IDF + Chi^2 + SVM, achieve comparable accuracy and weighted F1 scores to fine-tuned DistilBERT while drastically reducing training time and model size.

PCA applied to DistilBERT embeddings reduces dimensionality and computational cost, but at a slight cost to predictive performance.

Overall, a balance between semantic richness and computational efficiency can be achieved by combining DR with appropriate feature representations, making NLP models more suitable for resource-constrained environments.

These findings highlight the practical value of DR techniques in developing efficient and scalable NLP systems.

Future Work

Future research can explore several directions to extend this work:

Broader NLP Tasks: Apply DR-enhanced pipelines to multi-class sentiment analysis, topic classification, and other NLP tasks beyond binary sentiment.

Advanced Dimensionality Reduction: Investigate methods such as autoencoders, variational autoencoders (VAE), and contrastive self-supervised learning [15] for compressing transformer embeddings more effectively.

Dynamic Feature Selection: Implement task-specific or adaptive feature selection strategies to optimize the trade-off between accuracy and computational efficiency.

Cross-Lingual and Multilingual Models: Evaluate the impact of DR on transformer models in low-resource languages and cross-lingual transfer scenarios.

Real-Time Applications: Explore deployment of DR-enhanced NLP models in real-time systems where low latency and memory efficiency are critical.

By pursuing these directions, NLP systems can become both high-performing and computationally efficient, widening their applicability in practical, real-world scenarios.

REFERENCES

1. 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,doi: 10.18653/v1/N19-1423
2. V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” arXiv preprint arXiv:1910.01108, 2019.https://doi.org/10.48550/arXiv.1910.01108
3. A. Vaswani, N. Shazeer, N. Parmar, et al., “Attention is all you need,” in Proc. NeurIPS, 2017. https://doi.org/10.48550/arXiv.1706.03762
4. Y. Liu, M. Ott, N. Goyal, et al., “RoBERTa: A robustly optimized BERT pretraining approach,” arXiv preprint arXiv:1907.11692, 2019. https://doi.org/10.48550/arXiv.1907.11692
5. I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” J. Machine Learning Research, vol. 3, pp. 1157–1182, 2003.
6. H. Almuallim and T. G. Dietterich, “Learning with many irrelevant features,” in Proc. AAAI, 1991.
7. G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” Information Processing & Management, vol. 24, no. 5, pp. 513–523, 1988.
8. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” JASIS, vol. 41, no. 6, pp. 391–407, 1990.
9. K. Pearson, “On lines and planes of closest fit to systems of points in space,” Philosophical Magazine, vol. 2, no. 11, pp. 559–572, 1901.
10. R. Kohavi and G. H. John, “Wrappers for feature subset selection,” Artificial Intelligence, vol. 97, no. 1–2, pp. 273–324, 1997.
11. A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of tricks for efficient text classification,” in Proc. EACL, 2017.
12. R. Johnson and T. Zhang, “Effective use of word order for text categorization with CNNs,” in Proc. NAACL-HLT, 2015.

13. M. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies," in Proc. CVPR, 2007.
14. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015.
15. A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," Applied Intelligence, vol. 51, pp. 2065–2089, 2021.
16. X. Mao, Z. Li, Q. Li, and S. Zhang, "BERT-DXLMA: Enhanced representation learning and generalization model for English text classification," Neurocomputing, 2024,doi: 10.1016/j.neucom.2024.129325.