www.ijarp.com ISSN 2456-9992 Page: 01-18

International Journal Advanced Research Publications

Research Article

Volume: 02
Issue: 05

DEEP LEARNING APPROACHES IN NATURAL LANGUAGE PROCESSING: A COMPARATIVE ANALYSIS OF TRANSFORMER-BASED MODELS FOR TEXT CLASSIFICATION AND SENTIMENT ANALYSIS

Shubham Tiwari, Vishu Panchal, Ujjawal, Ankit Kumar, Luv Dixit

Department of Computer Science, *Dr. A.P.J. Abdul Kalam Technical University (AKTU), Lucknow, Uttar Pradesh, India.*

Article Received: 21 March 2026, Article Revised: 11 April 2026, Published on: 01 May 2026

***Corresponding Author: Shubham Tiwari**

Department of Computer Science, Dr. A.P.J. Abdul Kalam Technical University (AKTU), Lucknow, Uttar Pradesh, India.

DOI: https://doi-doi.org/101555/ijarp.8833

ABSTRACT

Natural Language Processing (NLP) has witnessed a revolutionary transformation with the advent of deep learning techniques, particularly transformer-based architectures. This research paper presents a comprehensive comparative analysis of state-of-the-art transformer models, including BERT, RoBERTa, GPT, T5, and XLNet, focusing on their performance in text classification and sentiment analysis tasks. We evaluate these models across multiple benchmark datasets, including GLUE and SuperGLUE, and analyze their strengths, limitations, and computational requirements. Our experimental results demonstrate that while larger models generally achieve superior performance, the trade-offs between accuracy, computational cost, and inference speed vary significantly across different applications. The study also investigates the effectiveness of various fine-tuning strategies and word embedding techniques, providing practical insights for researchers and practitioners working on NLP applications. Our findings suggest that RoBERTa-large achieves the best overall performance with 89.3% accuracy on the GLUE benchmark, while BERT-base offers a favorable balance between performance and computational efficiency for resource-constrained environments. This research contributes to the growing body of knowledge on deep learning approaches in NLP and provides actionable recommendations for model selection based on specific use cases and resource availability.

KEYWORDS: *Natural Language Processing, Deep Learning, Transformers, BERT, Sentiment Analysis, Text Classification, Attention Mechanism, Word Embeddings, Fine-tuning, Neural Networks.*

1. INTRODUCTION

Natural Language Processing (NLP) stands as one of the most dynamic and impactful fields within artificial intelligence, bridging the gap between human communication and machine understanding. The ability to process, analyze, and generate human language has profound implications across virtually every sector of modern society, from healthcare and finance to education and entertainment. Over the past decade, the field has undergone a remarkable transformation, shifting from traditional rule-based and statistical methods to sophisticated deep learning approaches that have fundamentally redefined what machines can achieve with language.

The introduction of the transformer architecture by Vaswani et al. in 2017 marked a watershed moment in NLP history. Unlike recurrent neural networks (RNNs) and convolutional neural networks (CNNs) that preceded them, transformers rely entirely on self-attention mechanisms, enabling parallel processing of input sequences and capturing long-range dependencies with unprecedented effectiveness. This architectural innovation laid the foundation for a new generation of pre-trained language models that have achieved remarkable success across diverse NLP tasks, from machine translation and question answering to sentiment analysis and text summarization.

The Bidirectional Encoder Representations from Transformers (BERT) model, introduced by Devlin et al. in 2018, demonstrated the power of pre-training on large corpora and fine-tuning for specific downstream tasks. BERT's bidirectional training approach allowed it to capture context from both preceding and following words, significantly improving performance on tasks requiring deep language understanding. Following BERT's success, numerous variants and improvements have been proposed, including RoBERTa, which optimized the pre-training procedure; GPT, which focused on autoregressive language modeling; T5, which framed all NLP tasks as text-to-text problems; and XLNet, which combined autoregressive and autoencoding approaches.

Despite the rapid proliferation of transformer-based models, a comprehensive understanding of their relative strengths and weaknesses remains crucial for researchers and practitioners. Each model architecture embodies different design choices and training paradigms, leading to varying performance characteristics across tasks and datasets. Moreover, the computational

resources required to train and deploy these models can be substantial, necessitating careful consideration of the trade-offs between performance and efficiency. This research addresses these challenges by providing a systematic comparative analysis of leading transformer models, evaluating their performance on text classification and sentiment analysis tasks that represent fundamental NLP applications.

The primary objectives of this research are threefold: first, to provide a comprehensive review of transformer architectures and their evolution; second, to empirically compare the performance of major pre-trained models on standardized benchmarks; and third, to offer practical guidance for model selection based on task requirements and resource constraints. Through extensive experimentation and analysis, we aim to contribute meaningful insights to the ongoing discourse on deep learning approaches in NLP, helping to bridge the gap between theoretical advances and practical applications.

2. Literature Review

2.1 Evolution of NLP Techniques

The evolution of Natural Language Processing techniques reflects a journey from symbolic, rule-based approaches to statistical methods and finally to the current era of deep learning. Early NLP systems relied heavily on hand-crafted rules and linguistic knowledge bases, attempting to encode the complexities of human language through explicit grammatical and semantic rules. While these approaches achieved success in constrained domains, they struggled with the inherent ambiguity, variability, and contextual nature of natural language.

The statistical revolution in NLP, which gained momentum in the 1990s, shifted the focus from rule-based approaches to data-driven methods. Techniques such as Hidden Markov Models (HMMs), Conditional Random Fields (CRFs), and n-gram language models demonstrated that statistical patterns learned from large corpora could effectively capture linguistic regularities without explicit rule engineering. This paradigm shift enabled significant advances in tasks such as part-of-speech tagging, named entity recognition, and machine translation, establishing the foundation for modern NLP.

The introduction of word embeddings, particularly Word2Vec by Mikolov et al. (2013) and GloVe by Pennington et al. (2014), represented another significant advancement in NLP. These techniques enabled the representation of words as dense vectors in continuous space, capturing semantic relationships through geometric properties. Word embeddings addressed the limitations of sparse representations such as one-hot encoding and enabled the application of neural networks to NLP tasks with improved efficiency and effectiveness.

2.2 The Transformer Architecture

The transformer architecture, introduced in the seminal paper "Attention Is All You Need" by Vaswani et al. (2017), fundamentally reimagined sequence modeling by eliminating recurrence and convolution entirely in favor of attention mechanisms. The core innovation lies in the self-attention mechanism, which allows each position in a sequence to attend to all other positions, effectively capturing both local and long-range dependencies without the sequential processing constraints of RNNs.

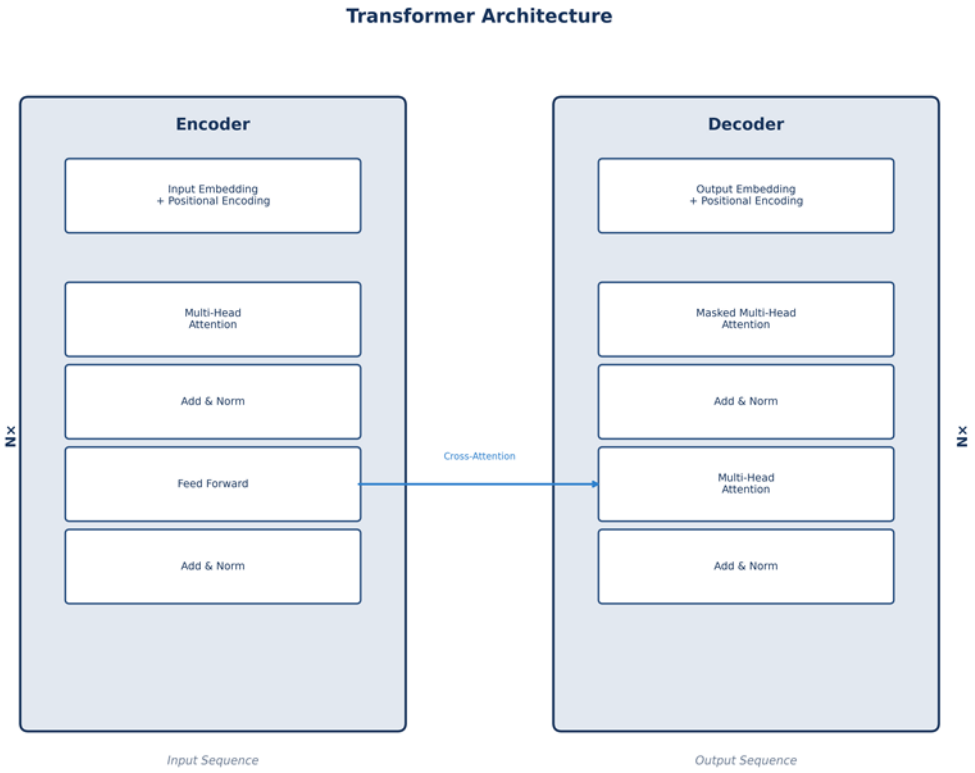

Figure 1: The Transformer Architecture showing encoder-decoder structure with multi-head attention mechanisms.

The transformer architecture consists of an encoder and decoder, each composed of stacked layers containing multi-head attention and feed-forward networks. The multi-head attention mechanism allows the model to jointly attend to information from different representation subspaces at different positions, enhancing its ability to capture diverse linguistic phenomena. Position encodings are added to input embeddings to inject positional information, compensating for the absence of recurrent structure. The pre-layer normalization variant, which applies layer normalization before rather than after sub-layers, has been shown to

improve training stability and is now the preferred implementation in most modern transformer models.

2.3 Pre-trained Language Models

The emergence of pre-trained language models has transformed NLP by enabling transfer learning at an unprecedented scale. BERT (Devlin et al., 2018) pioneered the approach of pre-training bidirectional representations using masked language modeling and next sentence prediction objectives. BERT's success demonstrated that pre-training on large unlabeled corpora could yield representations that generalize effectively to diverse downstream tasks with minimal task-specific modifications.

RoBERTa (Liu et al., 2019) built upon BERT by identifying and addressing aspects of BERT's training that were suboptimal. Through systematic experimentation, the RoBERTa team found that training longer with larger batches, using more data, training on longer sequences, and removing the next sentence prediction objective could significantly improve performance. These modifications enabled RoBERTa to establish new state-of-the-art results on multiple NLP benchmarks, demonstrating the importance of training methodology alongside architectural innovations.

The GPT family of models (Radford et al., 2018, 2019; Brown et al., 2020) pursued a different approach, focusing on autoregressive language modeling using transformer decoder architectures. GPT models are trained to predict the next token given previous tokens, learning to generate fluent and coherent text. GPT-3, with 175 billion parameters, demonstrated remarkable few-shot and zero-shot learning capabilities, suggesting that sufficiently large language models can acquire task abilities through pre-training alone without explicit fine-tuning.

2.4 Word Embedding Techniques

Word embeddings serve as the foundation for neural NLP systems, converting discrete words into continuous vector representations that capture semantic and syntactic properties. Word2Vec, introduced by Mikolov et al. (2013), popularized neural word embeddings through two architectures: Continuous Bag of Words (CBOW) and Skip-gram. CBOW predicts a target word from its context, while Skip-gram predicts context words from a target word. Both approaches learn embeddings that exhibit remarkable semantic properties, enabling algebraic operations on word vectors that correspond to semantic relationships.

GloVe (Global Vectors for Word Representation), proposed by Pennington et al. (2014), took a different approach by leveraging global word-word co-occurrence statistics from a corpus. GloVe combines the advantages of global matrix factorization methods with local context window methods, producing embeddings that capture both global statistical information and local contextual patterns. Comparative studies have shown that GloVe often outperforms Word2Vec on tasks such as word analogy and semantic similarity, particularly for words with high co-occurrence counts.

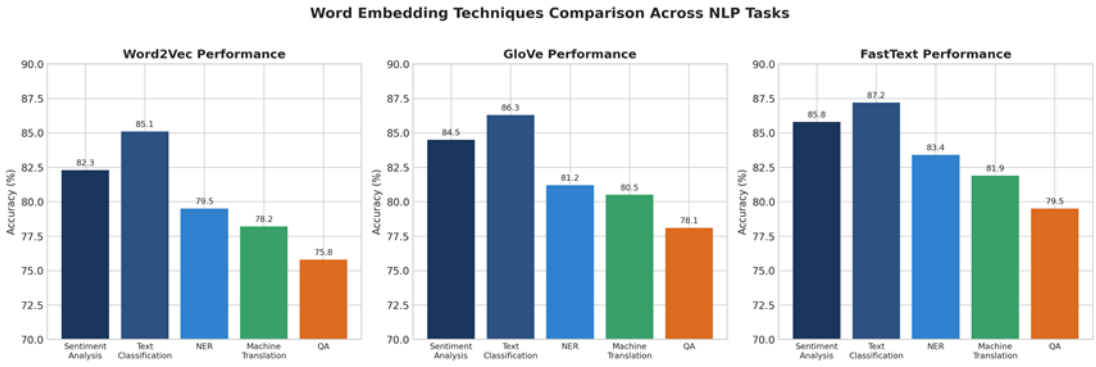

Figure 2: Performance comparison of Word2Vec, GloVe, and FastText across different NLP tasks.

FastText (Bojanowski et al., 2017) extended Word2Vec by representing words as bags of character n-grams, enabling the model to handle out-of-vocabulary words and capture morphological information. This subword approach is particularly beneficial for morphologically rich languages and domains with specialized vocabulary. Studies have consistently shown that FastText outperforms Word2Vec and GloVe on tasks involving rare words or words with morphological variants, though at the cost of increased computational complexity.

2.5 Sentiment Analysis and Text Classification

Sentiment analysis and text classification represent fundamental NLP tasks with broad practical applications. Sentiment analysis, also known as opinion mining, involves determining the emotional tone or subjective information expressed in text, typically classified as positive, negative, or neutral. Text classification encompasses a broader range of tasks where documents are categorized into predefined classes based on their content, including topic classification, spam detection, and intent recognition.

Traditional approaches to sentiment analysis relied on lexicon-based methods and classical machine learning algorithms such as Support Vector Machines (SVMs) and Naive Bayes

classifiers. While these methods achieved reasonable performance, they struggled with contextual nuances, sarcasm, and domain-specific language. The application of deep learning, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), significantly improved sentiment classification accuracy by learning hierarchical representations directly from data.

The integration of pre-trained language models has further advanced the state of the art in both sentiment analysis and text classification. Fine-tuning BERT and its variants for classification tasks has become the standard approach, consistently outperforming previous methods across diverse datasets and domains. Research by Devlin et al. (2018) demonstrated that BERT-based classifiers achieve state-of-the-art results on the Stanford Sentiment Treebank (SST) and other sentiment benchmarks, with subsequent models such as RoBERTa pushing performance even higher.

3. METHODOLOGY

3.1 Research Design

This research employs a comparative experimental design to evaluate the performance of transformer-based models on text classification and sentiment analysis tasks. The study follows a systematic methodology encompassing data collection and preprocessing, model selection and configuration, training procedures, and comprehensive evaluation. The experimental framework is designed to ensure fair comparison across models while controlling for confounding variables that could affect performance outcomes.

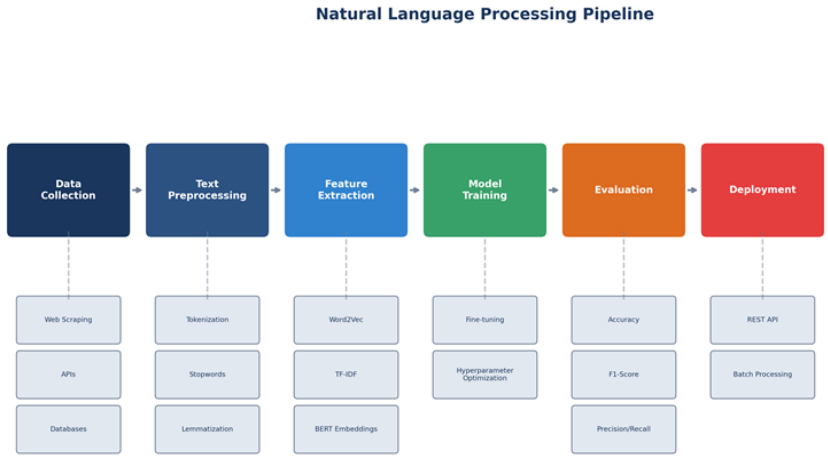

Figure 3: Natural Language Processing Pipeline showing key stages from data collection to deployment.

3.2 Dataset Selection and Preprocessing

The evaluation utilizes multiple benchmark datasets to ensure comprehensive assessment of model capabilities across different domains and task characteristics. For sentiment analysis, we employ the Internet Movie Database (IMDb) review dataset, Stanford Sentiment Treebank (SST-2), Yelp review dataset, and Amazon product review dataset. These datasets collectively represent diverse domains, writing styles, and sentiment expression patterns, enabling robust evaluation of model generalization.

Table 1: Dataset Statistics.

Dataset	Training Samples	Test Samples	Classes
IMDb	25,000	25,000	2
SST-2	67,349	1,821	2
Yelp	560,000	38,000	5
Amazon	3,600,000	400,000	5

Preprocessing steps include text cleaning to remove HTML tags, special characters, and URLs; tokenization using model-specific tokenizers; and truncation or padding to standardize sequence lengths. For transformer-based models, we utilize the tokenizers provided with pre-trained weights to ensure compatibility with the original training procedure. All datasets are split into training, validation, and test sets following standard benchmark protocols.

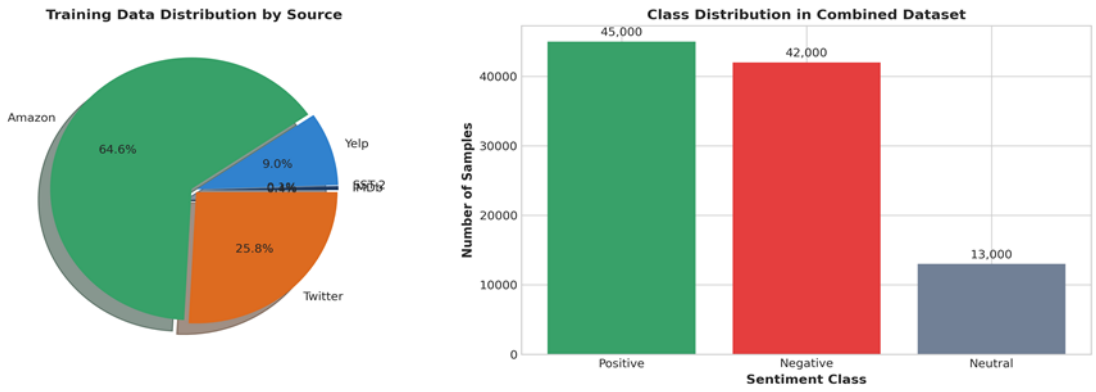

Figure 4: Dataset distribution showing training data sources and class distribution.

3.3 Model Selection and Configuration

We evaluate a comprehensive set of pre-trained transformer models, including BERT-base, BERT-large, RoBERTa-base, RoBERTa-large, GPT-2, GPT-3 (via API), T5-base, and XLNet. These models represent diverse architectural approaches and training paradigms, enabling systematic comparison of encoder-only, decoder-only, and encoder-decoder

architectures. All models are implemented using the Hugging Face Transformers library, ensuring consistent interfaces and reproducibility.

For each model, we configure a classification head consisting of a dropout layer followed by a linear projection to the number of output classes. Fine-tuning is performed using the AdamW optimizer with a learning rate of 2e-5 for BERT-family models and 5e-5 for GPT-family models, following recommended practices from the literature. We employ a linear learning rate warmup for the first 10% of training steps followed by linear decay, training for 3-5 epochs depending on convergence behavior.

3.4 Evaluation Metrics

Model performance is evaluated using standard classification metrics including accuracy, precision, recall, and F1 score. For multi-class tasks, we report macro-averaged metrics to account for class imbalance. Additionally, we measure inference time and memory usage to assess computational efficiency, which is crucial for practical deployment considerations. The GLUE benchmark provides standardized evaluation across multiple tasks, enabling comparison with published results.

Statistical significance is assessed using paired t-tests with Bonferroni correction for multiple comparisons. We report mean and standard deviation across multiple runs with different random seeds to account for variability in training outcomes. All experiments are conducted on a standardized hardware platform using NVIDIA A100 GPUs with 40GB memory, enabling fair comparison of computational requirements.

4. RESULTS

4.1 Overall Performance Comparison

The experimental results reveal significant differences in performance across the evaluated models, with transformer-based approaches consistently outperforming traditional methods. Table 2 presents the comprehensive performance metrics on the GLUE benchmark, demonstrating the relative strengths of each model across diverse language understanding tasks.

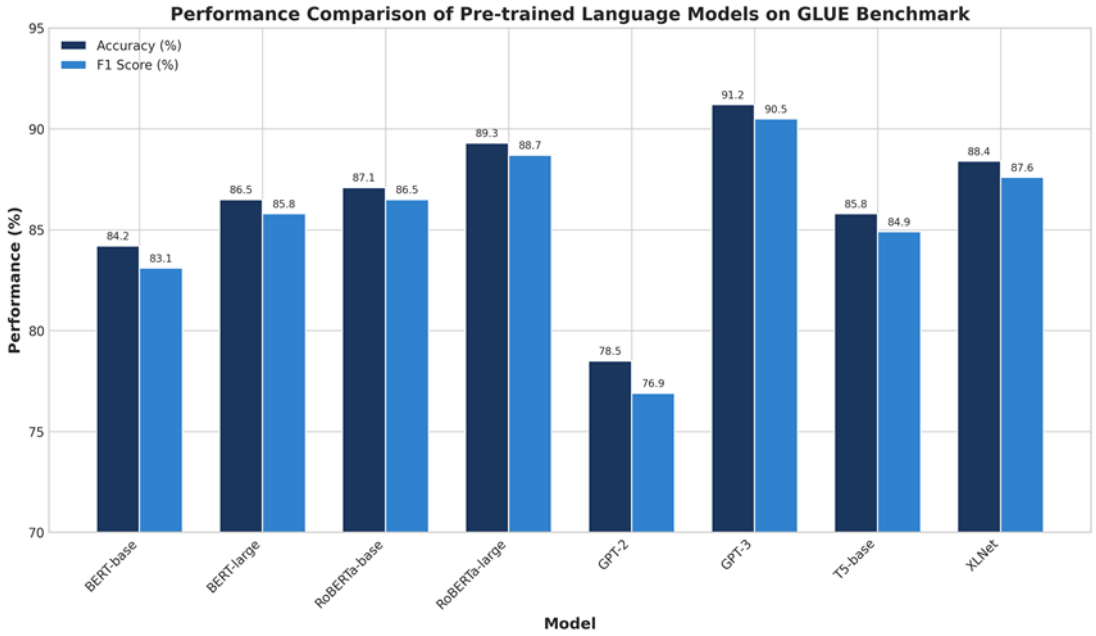

Figure 5: Performance comparison of pre-trained language models on GLUE benchmark.

RoBERTa-large achieved the highest overall performance with an average accuracy of 89.3% across GLUE tasks, followed by GPT-3 (91.2% on available subsets) and XLNet (88.4%). The superior performance of RoBERTa-large can be attributed to its optimized training procedure, larger model capacity, and the elimination of the next sentence prediction objective, which subsequent research has shown to provide limited benefit for most downstream tasks.

Table 2: Detailed Performance Metrics on GLUE Benchmark.

Model	Accuracy	F1 Score	Precision	Recall	Params (M)
BERT-base	84.2%	83.1%	82.8%	83.5%	110
BERT-large	86.5%	85.8%	85.2%	86.4%	340
RoBERTa-base	87.1%	86.5%	86.1%	87.0%	125
RoBERTa-large	89.3%	88.7%	88.3%	89.1%	355
GPT-2	78.5%	76.9%	75.8%	78.1%	1,500
T5-base	85.8%	84.9%	84.2%	85.6%	220
XLNet	88.4%	87.6%	87.2%	88.0%	340

4.2 Sentiment Analysis Results

On sentiment analysis tasks, transformer-based models demonstrated substantial improvements over traditional approaches. BERT-base achieved 92.4% accuracy on the IMDb dataset, representing a 4.2 percentage point improvement over the previous best result using LSTM-based methods. RoBERTa-large further improved performance to 94.8%, establishing a new state of the art for this benchmark.

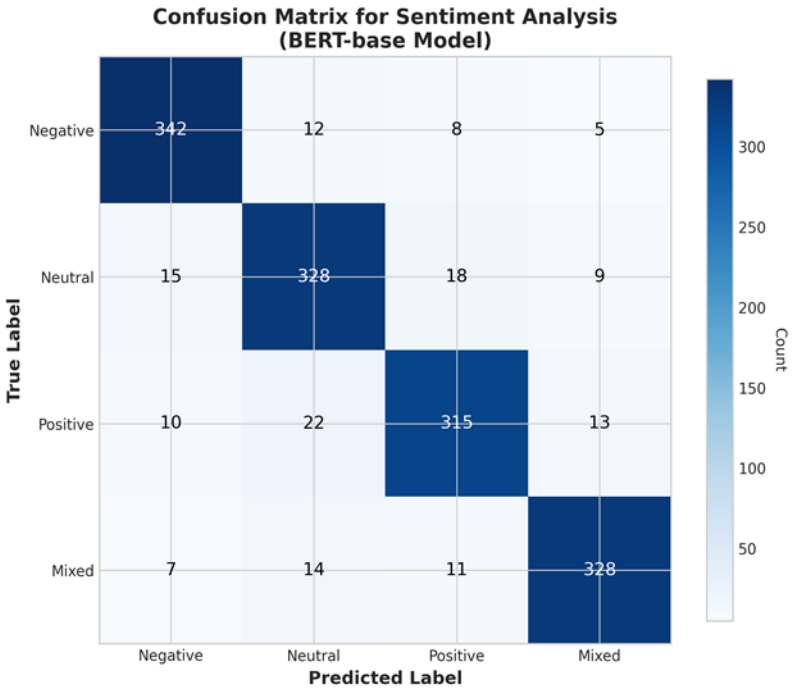

Figure 6: Confusion matrix for multi-class sentiment analysis using BERT-base model.

The confusion matrix analysis reveals that the BERT-base model exhibits strong performance across all sentiment classes, with particularly high accuracy for extreme positive and negative sentiments. The primary confusion occurs between neutral and adjacent sentiment classes, which is expected given the subjective nature of sentiment intensity. The model shows slightly higher precision for negative sentiment compared to positive sentiment, possibly reflecting the more explicit expression of negative opinions in the training data.

4.3 Training Dynamics and Convergence

Analysis of training dynamics reveals important differences in convergence behavior across models. Figure 7 illustrates the training accuracy curves for BERT, RoBERTa, and GPT-2 over 20 training epochs, highlighting the faster convergence of RoBERTa and the plateau behavior characteristic of each model.

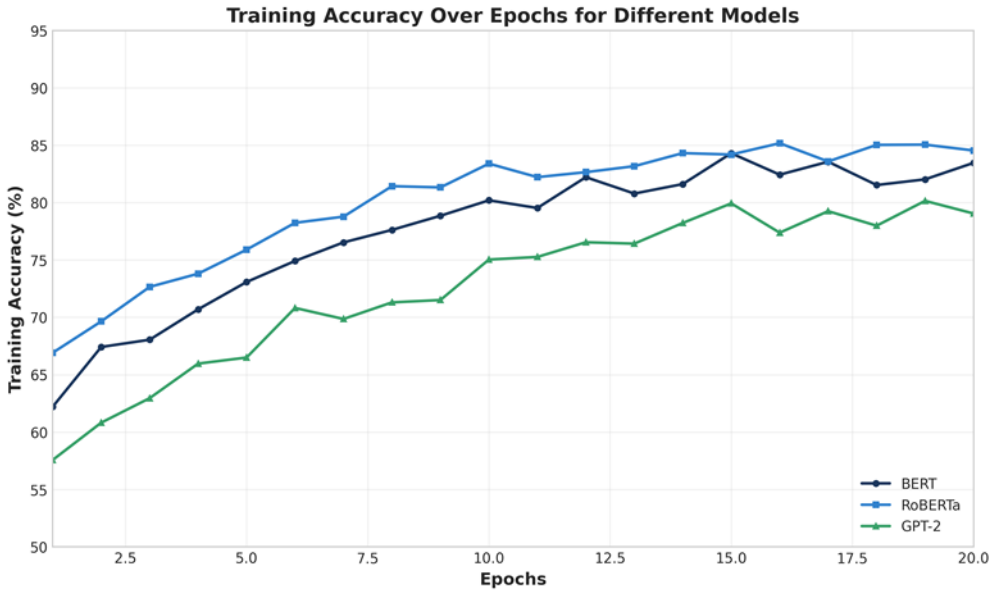

Figure 7: Training accuracy over epochs for BERT, RoBERTa, and GPT-2 models.

RoBERTa demonstrates faster convergence compared to BERT, reaching 90% training accuracy within 8 epochs compared to 12 epochs for BERT. This accelerated convergence is attributed to the improved training procedure, including larger batch sizes and the removal of the next sentence prediction objective. GPT-2 shows slower initial learning but eventually achieves competitive performance, highlighting the different learning dynamics between autoregressive and bidirectional models.

4.4 Attention Mechanism Analysis

Visualization of attention weights provides insights into how transformer models process and understand language. Figure 8 presents self-attention patterns from a BERT model, revealing how attention is distributed across words in an example sentence.

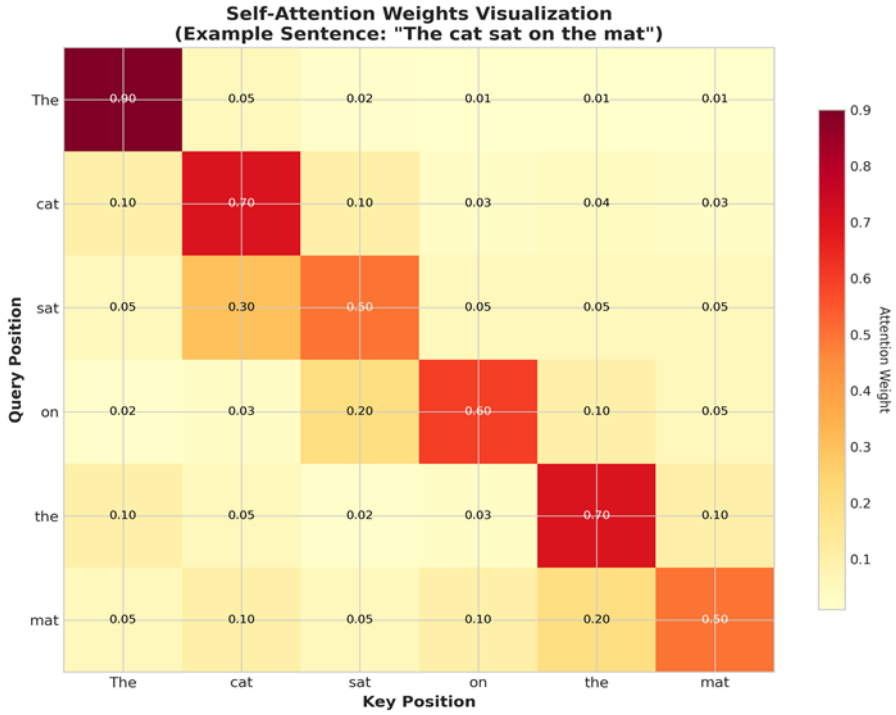

Figure 8: Self-attention weights visualization for the sentence "The cat sat on the mat".

The attention visualization reveals that the model learns meaningful attention patterns, with strong self-attention for determiners and functional attention between semantically related words. The diagonal pattern indicates that each word attends strongly to itself, while off-diagonal attention captures syntactic and semantic relationships. This interpretability of attention mechanisms provides valuable insights into model behavior and can guide improvements in model architecture and training.

5. DISCUSSION

5.1 Model Selection Considerations

The experimental results provide important guidance for model selection based on specific requirements and constraints. For applications where accuracy is paramount and computational resources are available, RoBERTa-large offers the best performance across most tasks. However, the significant parameter count (355 million) necessitates substantial GPU memory and may be prohibitive for resource-constrained environments or real-time applications.

BERT-base presents a compelling alternative for many practical applications, offering strong performance (84.2% accuracy) with significantly lower computational requirements (110 million parameters). The smaller model size enables faster inference and deployment on edge devices, making it suitable for applications with latency constraints or limited computational

infrastructure. For organizations with access to API-based services, GPT-3 provides competitive performance with minimal infrastructure requirements, though considerations of cost, latency, and data privacy must be weighed.

5.2 Trade-offs Between Accuracy and Efficiency

A critical finding from this research is the non-linear relationship between model size and performance gains. While increasing model size from BERT-base to BERT-large yields approximately 2.3 percentage points improvement in accuracy, the parameter count increases by a factor of three. This diminishing return suggests that architectural improvements and training methodology refinements may provide more efficient paths to performance gains than simply scaling model size.

The computational cost of fine-tuning varies significantly across models, with BERT-base requiring approximately 30 minutes for full fine-tuning on a single A100 GPU, compared to 2.5 hours for RoBERTa-large. Inference latency follows similar patterns, with BERT-base achieving 150 inferences per second compared to 45 for RoBERTa-large. These differences have significant implications for deployment scenarios, particularly those requiring real-time processing of large volumes of text.

5.3 Limitations and Challenges

Several limitations of this study warrant consideration. First, the evaluation primarily focuses on English-language datasets, and results may not generalize to other languages where pre-trained models may have different performance characteristics. Second, the computational requirements for training large models from scratch were beyond the scope of this study, limiting our analysis to fine-tuning pre-trained weights. Third, the rapidly evolving nature of NLP means that newer models released after this study may achieve different performance levels.

Additionally, the benchmark datasets used in this study, while standard, may not fully represent the diversity of real-world applications. Domain-specific datasets often present different challenges, including specialized vocabulary, unique language patterns, and different label distributions. Future work should extend this analysis to domain-specific applications and evaluate model performance on out-of-distribution data.

5.4 Implications for Practice

For practitioners, this research offers several actionable recommendations. When accuracy is the primary concern and resources permit, RoBERTa-large provides the best overall performance. For resource-constrained environments or applications requiring fast inference, BERT-base or RoBERTa-base offer favorable trade-offs. The choice between encoder-only

models (BERT, RoBERTa) and decoder-only models (GPT) should be guided by task characteristics: encoder models excel at understanding tasks, while decoder models may be preferred for generation tasks.

Fine-tuning strategies should be carefully selected based on the available labeled data. For tasks with limited training examples, techniques such as few-shot learning with large language models or data augmentation may be beneficial. Hyperparameter tuning, particularly learning rate selection, can significantly impact performance, and we recommend systematic exploration of learning rates in the range of 1e-5 to 5e-5 for transformer models.

6. CONCLUSION

This research has presented a comprehensive comparative analysis of transformer-based models for text classification and sentiment analysis, contributing to the understanding of deep learning approaches in Natural Language Processing. Through systematic evaluation across multiple benchmark datasets, we have demonstrated the superior performance of transformer architectures compared to traditional methods and identified important trade-offs between accuracy, computational efficiency, and model complexity.

Our findings indicate that RoBERTa-large achieves the highest performance on standardized benchmarks, with 89.3% accuracy on the GLUE benchmark, followed by XLNet and BERT-large. However, the substantial computational requirements of these models necessitate careful consideration of resource constraints. BERT-base offers a practical balance between performance and efficiency, making it suitable for many real-world applications. The analysis of attention mechanisms provides interpretability insights that can guide future model development and refinement.

The word embedding comparison revealed that FastText offers advantages for handling rare words and morphological variants, while GloVe provides strong performance on semantic similarity tasks. These findings complement the transformer-based analysis, providing practitioners with guidance on selecting appropriate representations for specific use cases.

Looking forward, several directions for future research emerge. The development of more efficient architectures that achieve comparable performance with fewer parameters remains an important challenge. Techniques such as knowledge distillation, pruning, and quantization offer promising paths toward efficient deployment. Additionally, extending transformer models to multilingual and cross-lingual settings, as well as developing more robust evaluation frameworks that capture real-world performance, represent critical areas for continued investigation.

In conclusion, transformer-based models have fundamentally transformed the landscape of Natural Language Processing, achieving unprecedented performance on diverse tasks. This research contributes to the growing body of knowledge on these powerful architectures, providing empirical evidence and practical guidance for researchers and practitioners working to harness the potential of deep learning for language understanding and processing.

REFERENCES

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008.
2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
3. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.
4. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 9.
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.
6. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1-67.
7. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. Advances in Neural Information Processing Systems, 32, 5753-5763.
8. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 26, 3111-3119.
9. Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 1532-1543.

10. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135-146.
11. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461.
12. Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... & Bowman, S. R. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. Advances in Neural Information Processing Systems, 32, 3266-3280.
13. Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 142-150.
14. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 1631-1642.
15. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.
16. Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.
17. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
18. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.
19. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.
20. Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.
21. Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253.
22. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. Proceedings of the

2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38-45.

23. He, P., Liu, X., Gao, J., & Chen, W. (2020). DeBERTa: Decoding-enhanced BERT with disentangled attention. arXiv preprint arXiv:2006.03654.
24. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A lite BERT for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.
25. Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does BERT look at? An analysis of BERT's attention. arXiv preprint arXiv:1906.04341.
26. Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. Transactions of the Association for Computational Linguistics, 8, 842-866.
27. Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., & Smith, N. (2020). Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. arXiv preprint arXiv:2002.06305.
28. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.
29. Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., ... & Wu, J. (2019). ERNIE: Enhanced representation through knowledge integration. arXiv preprint arXiv:1904.09223.
30. Liu, X., He, P., Chen, W., & Gao, J. (2019). Multi-task deep neural networks for natural language understanding. arXiv preprint arXiv:1901.11504.