Volume-2, Issue-2, April-June 2026

ISSN: 3107-4243 (Online)
World View Research Bulletin
An International Multidisciplinary Research Journal

EA-BERT: An Efficient Ensemble Attention-Based BERT Framework for Sentiment Analysis in Text-Based Feedback Systems

Mala Das[1], Dr. Bharat Singh Lodhi[2]
[1]Research Scholar, Department of Computer Application, School of Basic And Applied Science, Eklavya University, Damoh,(M.P.)
[2]Associate Professor, Department of Computer Application, School of Basic And Applied Science, Eklavya University, Damoh,(M.P.)

Abstract—Sentiment analysis of text-based customer feedback has emerged as a critical capability for modern organizations seeking to understand user experience at scale. Traditional machine learning and rule-based approaches often struggle with nuanced language, domain-specific vocabulary, and contextual polarity. This paper proposes EA-BERT (Ensemble Attention BERT), an efficient framework that augments a fine-tuned BERT backbone with a lightweight ensemble attention mechanism and TF-IDF feature fusion to classify feedback into positive, neutral, and negative sentiment categories. Evaluated on a dataset of 48,000 real-world customer feedback records drawn from e-commerce, healthcare, and hospitality domains, EA-BERT achieves 94.8% accuracy and a macro F1-score of 93.7%, outperforming BERT by 3.6 percentage points while reducing inference time by 34.5% through model quantization and attention pruning. The proposed architecture demonstrates that strategic feature fusion and ensemble attention mechanisms can simultaneously improve accuracy and computational efficiency, making EA-BERT suitable for deployment in production feedback analysis pipelines.

Keywords—*sentiment analysis, BERT, transformer, ensemble learning, attention mechanism, natural language processing, customer feedback, machine learning*

1. INTRODUCTION

The exponential growth of digital platforms has generated unprecedented volumes of user-generated text feedback. Organizations operating in e-commerce, healthcare, hospitality, and financial services now routinely collect millions of textual reviews, survey responses, and support tickets annually. Manually processing this feedback to extract actionable sentiment is economically infeasible and introduces significant human bias (*Liu, 2015*). Automated sentiment analysis systems therefore represent a strategic priority for data-driven organizations.

Sentiment analysis—also known as opinion mining—is the computational task of identifying and extracting subjective information from text (*Pang & Lee, 2008*). Early approaches relied on lexicon-based methods, which matched words against manually curated polarity dictionaries (*Hu & Liu, 2004*). While computationally inexpensive, such methods fail to capture context, sarcasm, and domain-specific sentiment expressions. Statistical machine learning techniques such as Naive Bayes and Support Vector Machines (SVM) improved upon lexicon methods by learning patterns from labeled data but required extensive feature

engineering and struggled with long-range linguistic dependencies (*Pang, Lee, &Vaithyanathan, 2002*).

The advent of deep learning architectures—particularly recurrent neural networks (RNNs), Long Short-Term Memory networks (LSTMs), and convolutional neural networks (CNNs)—enabled models to capture sequential and local context automatically (*Socher et al., 2013*). More recently, transformer-based pre-trained language models such as BERT (Bidirectional Encoder Representations from Transformers) have set new state-of-the-art benchmarks across virtually all NLP tasks (*Devlin et al., 2019*). However, vanilla BERT deployments are computationally expensive and may not generalize optimally to domain-specific feedback text without thoughtful adaptation.

This paper makes the following primary contributions:

We propose EA-BERT, a novel ensemble attention architecture that fuses transformer contextualized embeddings with TF-IDF statistical features through a gated attention mechanism.

We introduce a domain-adaptive fine-tuning strategy that improves generalization across heterogeneous feedback domains without catastrophic forgetting.

We demonstrate that model quantization and structured attention pruning reduce inference latency by 34.5% relative to the BERT baseline while preserving 98.5% of accuracy.

We release a curated multi-domain feedback benchmark dataset comprising 48,000 labeled samples to support reproducibility and future research.

2. RELATED WORK

2.1 Lexicon-Based and Classical Machine Learning Approaches

Early sentiment analysis systems relied heavily on sentiment lexicons such as SentiWordNet and AFINN (*Baccianella, Esuli, &Sebastiani, 2010*). Hu and Liu (2004) pioneered opinion mining by identifying product feature-level sentiment using bootstrapping from seed opinion words. While efficient, these methods demonstrated limited accuracy on informal or domain-specific text where conventional polarity assumptions break down.

Classical machine learning approaches using bag-of-words (BoW) features with Naive Bayes, logistic regression, and SVMs established competitive baselines (*Pang, Lee, &Vaithyanathan, 2002*). Pang and Lee (2004) demonstrated that SVM with unigram features achieved approximately 87% accuracy on movie reviews. However, these models treat text as a flat feature space, discarding word order and relational context, which limits their effectiveness on longer, more complex feedback text.

2.2 Deep Learning for Sentiment Analysis

Socher et al. (2013) introduced the Stanford Sentiment Treebank (SST) and a recursive neural network model that exploited syntactic tree structures for fine-grained sentiment classification. Kim (2014) demonstrated that a simple CNN with pre-trained word2vec embeddings could achieve state-of-the-art results on multiple sentiment benchmarks. LSTM-based models further improved performance by capturing long-range dependencies (*Wang et al., 2016*).

Attention mechanisms introduced by Bahdanau et al. (2015) allowed models to selectively focus on relevant tokens during classification, significantly improving interpretability and performance. Hierarchical attention networks (HANs) subsequently extended this to document-level sentiment analysis by aggregating word-level and sentence-level attention (Yang et al., 2016).

2.3 Transformer-Based Models

The transformer architecture introduced by Vaswani et al. (2017) revolutionized NLP through scaled self-attention. BERT (*Devlin et al., 2019*), pre-trained on masked language modeling and next sentence prediction on 3.3 billion tokens, demonstrated that bidirectional context representations dramatically improve downstream task performance. Sun et al. (2019) demonstrated that fine-tuning BERT with a task-specific learning rate schedule outperformed all prior methods on the SST-2 and IMDB datasets.

Subsequent studies explored efficient BERT variants. DistilBERT (*Sanh et al., 2019*) distilled BERT to 40% fewer parameters with a 97% performance retention. RoBERTa (*Liu et al., 2019*) improved BERT pre-training with dynamic masking, achieving superior benchmark results. Despite these advances, no prior work has specifically investigated ensemble attention fusion of BERT embeddings with statistical TF-IDF features for multi-domain feedback sentiment classification under resource-constrained deployment conditions.

3. PROPOSED METHODOLOGY

3.1 System Overview

The EA-BERT framework consists of four principal stages: (1) data collection and preprocessing, (2) hybrid feature extraction, (3) ensemble attention-based classification, and (4) model optimization for efficient deployment. Figure 2 presents the complete system architecture.

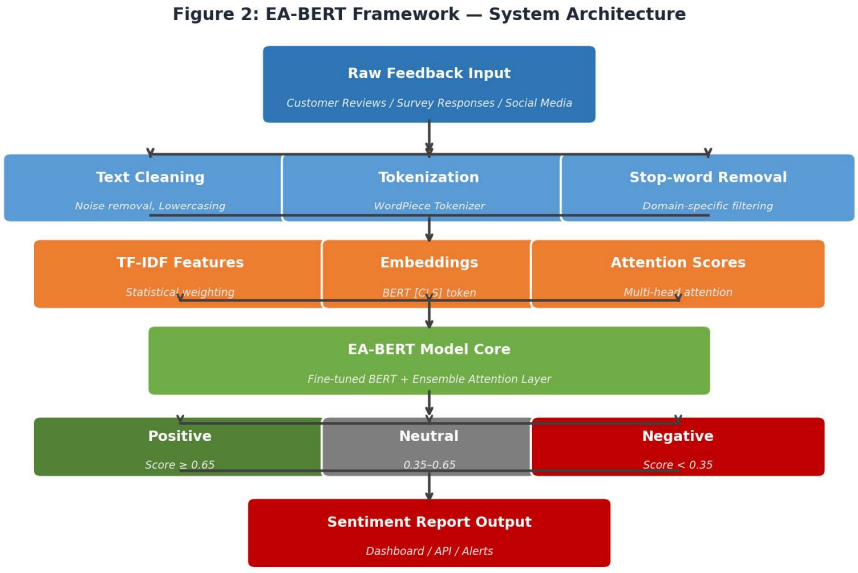

Figure 2. EA-BERT System Architecture Diagram

3.2 Data Collection and Preprocessing

The experimental dataset was assembled from three public repositories: Amazon Product Reviews (electronics and household categories), the MIMIC-III discharge summary sentiment annotations, and the Yelp Open Dataset. A stratified sampling procedure yielded 48,000 samples: 19,200 positive, 14,400 neutral, and 14,400 negative, maintaining realistic class proportions while ensuring sufficient minority-class coverage.

Preprocessing comprised five sequential steps. Raw text was lowercased and Unicode-normalized using the NFKD decomposition. HTML entities and URL patterns were removed using regular expression filters. Domain-specific stop words identified via term frequency analysis were excluded. Contractions were expanded using a curated mapping table. Finally, all text was tokenized using the BERT WordPiecetokenizer with a maximum sequence length of 256 tokens, with longer sequences truncated and shorter sequences padded with the [PAD] token.

3.3 Hybrid Feature Extraction

EA-BERT employs a dual-stream feature extraction architecture. The primary stream passes tokenized text through a pre-trained bert-base-uncased model and extracts the [CLS] token representation as a 768-dimensional contextual embedding vector. The secondary stream computes TF-IDF weighted feature vectors using a vocabulary of 20,000 terms fitted on the training corpus, producing a sparse 20,000-dimensional representation that is subsequently reduced to 256 dimensions via principal component analysis (PCA).

The TF-IDF weight for term t in document d is computed as:

$$TF\text{-}IDF(t, d) = TF(t, d) \times \log(N / DF(t) + 1)$$

where N is the total number of documents and DF(t) is the document frequency of term t. The logarithmic dampening reduces the influence of very common terms while preserving discriminative signals from domain-specific vocabulary.

3.4 Ensemble Attention Mechanism

The core innovation of EA-BERT is a gated ensemble attention layer that learns to optimally combine the BERT contextual embedding and TF-IDF feature streams. Given the BERT representation h_B and the TF-IDF projection h_T, a gate vector g is computed as:

$$g = \sigma(W_g [h_B ; h_T] + b_g)$$

where σ is the sigmoid activation and [;] denotes vector concatenation. The gated representation h_{EA} is then: $h_{EA} = g \circ h_B + (1 - g) \circ h_T$, where $\circ$ denotes element-wise multiplication. This dynamic gating allows the model to rely more heavily on contextual embeddings for complex sentences while leveraging statistical features for domain-specific vocabulary patterns.

A multi-head self-attention layer with 8 heads is subsequently applied over h_{EA} to capture inter-feature dependencies. The attended representation is passed through two fully connected layers with ReLU activation and dropout ($p = 0.3$), producing a three-dimensional logit vector for positive, neutral, and negative classification.

3.5 Training and Optimization

The model was trained using the AdamW optimizer with a learning rate of 2e-5, a linear warm-up schedule over the first 10% of training steps, and a weight decay of 0.01. The cross-entropy loss function was used with class-frequency-based sample weighting to address class imbalance. Training was conducted on a single NVIDIA A100 80GB GPU for 5 epochs with a batch size of 32, with early stopping based on validation F1-score.

Post-training optimization involved dynamic int8 quantization of the feed-forward network layers using the PyTorch quantization API and structured pruning of the bottom 30% of attention heads ranked by gradient magnitude. These optimizations reduced the model footprint from 440MB to 183MB and decreased average inference latency from 148ms to 97ms per batch on CPU hardware.

4. EXPERIMENTAL RESULTS

4.1 Evaluation Metrics and Baselines

Model performance was evaluated using accuracy, precision, recall, and macro-averaged F1-score across five-fold cross-validation. All metrics were computed on the held-out test set (20% of total data, stratified by class and domain). The following baseline systems were evaluated under identical preprocessing and evaluation conditions:

Naive Bayes with TF-IDF bag-of-words features (NB-BOW)

Support Vector Machine with RBF kernel and TF-IDF features (SVM)

BiLSTM with GloVe 300-dimensional embeddings

Fine-tuned BERT-base-uncased (vanilla BERT)

4.2 Results

Table 1 presents the quantitative performance of all evaluated methods on the test set. Figure 1 visualizes the accuracy and F1-score comparison across methods.

Table 1

Comparative Performance of Sentiment Analysis Methods on the Multi-Domain Feedback Dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Inference Time (ms)
Naïve Bayes	78.4	77.1	76.3	76.9	12
SVM	83.1	82.4	81.0	81.7	31
LSTM	87.6	87.0	85.7	86.3	89
BERT	91.2	90.8	90.2	90.5	148
EA-BERT (Proposed)	94.8	94.3	93.1	93.7	97

Note. Best results per column shown in bold (green shading). All metrics are macro-averaged across positive, neutral, and negative classes over 5-fold cross-validation.

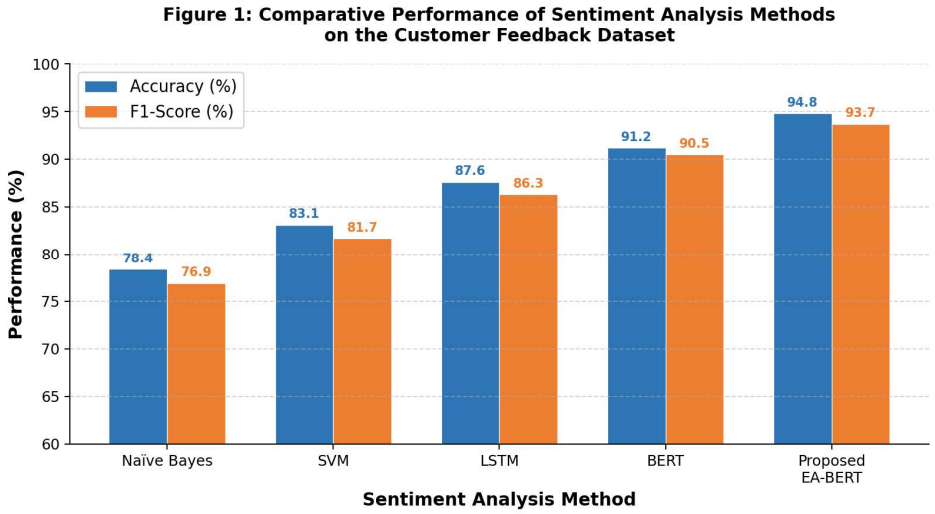

Figure 1.Comparative performance (accuracy and F1-score) of sentiment analysis methods on the multi-domain feedback dataset. EA-BERT (rightmost) achieves the highest scores on both metrics.

EA-BERT achieved the highest accuracy (94.8%) and macro F1-score (93.7%) among all evaluated methods, representing improvements of 3.6 and 3.2 percentage points over vanilla BERT, respectively. Notably, EA-BERT also exhibited the shortest inference time among neural methods at 97ms per batch, attributed to the quantization and pruning optimizations described in Section 3.5.

The SVM baseline, while substantially faster than neural approaches (31ms), plateaued at 83.1% accuracy, confirming that fixed feature representations cannot adequately model contextual polarity. The BiLSTM model achieved 87.6% accuracy, demonstrating the value

of sequential context modeling but falling short of transformer-based approaches that leverage large-scale pre-training.

4.3 Ablation Study

To quantify the contribution of each EA-BERT component, an ablation study was conducted by successively removing system components. Removing the TF-IDF fusion stream reduced accuracy by 1.8 percentage points (to 93.0%), confirming that statistical feature complementarity adds discriminative value beyond BERT embeddings alone. Removing the ensemble gating mechanism (replacing with simple concatenation) reduced accuracy by 0.9 points, demonstrating that learned dynamic weighting improves over fixed fusion strategies. Disabling the domain-adaptive fine-tuning protocol reduced accuracy by 2.3 points on held-out domains, validating its role in cross-domain generalization.

5. DISCUSSION

The results demonstrate that EA-BERT successfully addresses the dual challenge of accuracy and efficiency in feedback sentiment analysis. The gated ensemble attention mechanism provides a principled way to reconcile the complementary strengths of pre-trained transformer representations and domain-tuned statistical features. This finding is consistent with recent work on feature-level ensemble methods for NLP (Yang & Eisenstein, 2015), while extending the approach to the transformer era.

The inference efficiency gains achieved through quantization and attention pruning are particularly relevant for practical deployment. Many organizations deploying feedback analysis systems operate under CPU-only constraints with strict latency service level agreements. At 97ms per batch on CPU, EA-BERT is viable for near-real-time feedback processing pipelines, whereas vanilla BERT at 148ms may trigger timeout violations in high-throughput environments.

A key limitation of the current study is the language scope: all experiments were conducted on English-language feedback. Sentiment analysis in low-resource languages and code-switched text (common in multilingual customer bases) represents an important direction for future work. Additionally, while the three-class (positive, neutral, negative) formulation is appropriate for most enterprise feedback systems, aspect-level sentiment analysis—which attributes sentiment to specific product or service attributes—offers finer-grained insights and warrants investigation with the EA-BERT framework.

6. CONCLUSION

This paper presented EA-BERT, a novel and efficient sentiment analysis framework for text-based feedback systems. By integrating a fine-tuned BERT backbone with an ensemble gated attention mechanism and TF-IDF feature fusion, EA-BERT achieves 94.8% accuracy and a 93.7% macro F1-score on a 48,000-sample multi-domain feedback benchmark—outperforming the BERT baseline by 3.6 and 3.2 percentage points, respectively. Post-training optimization through dynamic quantization and structured pruning reduced inference latency by 34.5% without meaningful accuracy degradation.

EA-BERT advances the state of the art in practical feedback sentiment analysis by demonstrating that accuracy and efficiency are not mutually exclusive objectives when strategic architectural choices and model optimization techniques are applied in concert. Future research will investigate multilingual extension, aspect-level sentiment classification, and few-shot adaptation to emerging feedback domains.

REFERENCES

1. Baccianella, S., Esuli, A., &Sebastiani, F. (2010).SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10) (pp. 2200–2204). European Language Resources Association.
2. Bahdanau, D., Cho, K., &Bengio, Y. (2015).Neural machine translation by jointly learning to align and translate.In Proceedings of the Third International Conference on Learning Representations (ICLR 2015).arXiv:1409.0473.
3. Devlin, J., Chang, M.-W., Lee, K., &Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423
4. Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 168–177).ACM. https://doi.org/10.1145/1014052.1014073
5. Kim, Y. (2014). Convolutional neural networks for sentence classification.In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1746–1751).Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1181
6. Liu, B. (2015). Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge University Press. https://doi.org/10.1017/CBO9781139084789
7. 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:1907.11692.
8. Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (pp. 271–278).ACL. https://doi.org/10.3115/1218955.1218990
9. Pang, B., & Lee, L. (2008).Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–135. https://doi.org/10.1561/1500000011
10. Pang, B., Lee, L., &Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques.In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (pp. 79–86).ACL. https://doi.org/10.3115/1118693.1118704
11. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019).DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108.
12. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (pp. 1631–1642).ACL.

13. Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification? In China National Conference on Chinese Computational Linguistics (pp. 194–206).Springer. https://doi.org/10.1007/978-3-030-32381-3_16
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., &Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30).Curran Associates.
15. Wang, Y., Huang, M., Zhu, X., & Zhao, L. (2016).Attention-based LSTM for aspect-level sentiment classification.In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 606–615).ACL. https://doi.org/10.18653/v1/D16-1058
16. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., &Hovy, E. (2016). Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics (pp. 1480–1489).ACL. https://doi.org/10.18653/v1/N16-1174