SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

A Novel Ensemble Approach for Facial Emotion Recognition and Sentiment Analysis Using Deep Learning and Attention Mechanisms

Mohammad Subhi Al-Batah[1*], Mashhoor Al-Awaishah[2]

[1,2] Department of Computer Science, Faculty of Information Technology, Jadara University, Irbid, Jordan.
[*]Corresponding author: Mohammad Subhi Al-Batah; E-mail: albatah@jadara.edu.jo

Abstract
Facial emotion recognition (FER) and text sentiment analysis are complementary affective-computing tasks, but they are often evaluated independently with inconsistent model comparisons. This study evaluates deep visual backbones and text classifiers under a unified reporting framework. For seven-class FER on a processed AffectNet subset, VGG16, ResNet50, InceptionV3, EfficientNetB0, VGG16 with attention, and a hard-voting ensemble of VGG16, ResNet50, and EfficientNetB0 were compared. For binary sentiment classification on 50,000 IMDB reviews, BERT, LSTM, CNN, and SVM were evaluated. The FER ensemble achieved the highest accuracy (95.2%), precision (94.5%), recall (94.8%), and F1-score (94.7%); VGG16 with attention reached 94.6% accuracy. BERT led the text task with 94.9% accuracy and 94.6% F1-score. The findings support model complementarity and attention-enhanced representation learning while clarifying that the image and text pipelines are parallel rather than cross-modal.

Keywords: facial emotion recognition; sentiment analysis; ensemble learning; attention mechanism; BERT; AffectNet; IMDB.

1. Introduction

Affective computing aims to enable computational systems to recognize, interpret, and respond to human emotional states. Two important branches are facial emotion recognition (FER), which infers affect from visual facial cues, and sentiment analysis, which identifies evaluative polarity from language. These tasks support applications in human-computer interaction, customer experience analysis, assistive technologies, education, and behavioral analytics. Recent reviews show that deep learning has substantially advanced both visual emotion recognition and text-based sentiment modeling, while robustness to pose, illumination, identity, domain shift, and contextual ambiguity remains a central challenge (Kopalidis et al., 2024; Lian et al., 2023).

For FER, convolutional neural networks (CNNs) and transfer learning have become dominant because pretrained visual backbones can reuse hierarchical features learned from large image corpora. VGG16, ResNet50, InceptionV3, and EfficientNet represent distinct architectural strategies for depth, residual learning, multi-scale processing, and efficient model scaling (He et al., 2016; Simonyan and Zisserman, 2015; Szegedy et al., 2016; Tan and Le, 2019). AffectNet is particularly important for evaluation in unconstrained conditions because it contains more than one million web-collected facial images, with a large manually annotated subset for categorical expressions and valence-arousal estimation (Mollahosseini et al., 2019). Attention mechanisms can further emphasize expression-relevant regions or channels, and recent attention-enhanced ensembles have reported strong FER performance by exploiting complementary representations (Khan et al., 2025).

Sentiment analysis has followed a parallel progression from classical classifiers and recurrent networks to transformer-based language models. LSTM networks address long-range sequential dependencies (Hochreiter and Schmidhuber, 1997), CNNs capture local discriminative n-gram patterns (Kim, 2014), and support vector machines (SVMs) remain useful as efficient linear or kernel-based baselines. BERT introduced deep bidirectional contextual representations that can be fine-tuned for text classification and related downstream tasks (Devlin et al., 2019), building on the self-attention mechanism formalized by Vaswani et al. (2017). The 50,000-review IMDB benchmark remains a standard resource for binary sentiment classification (Maas et al., 2011).

Related work by Al-Batah and collaborators also illustrates the value of comparative modeling across image and text classification. Al-Batah et al. (2018) combined Naive Bayes and multilayer perceptron models for

sentiment classification, whereas Al-Batah and Alzboon (2026) used systematic multi-model comparison for image classification. More recent ensemble studies have emphasized complementary learners and careful evaluation rather than reliance on a single model (Al-Batah and Alabbas, 2026; Alzboon and Al-Batah, 2026). These studies motivate a transparent comparison of heterogeneous models while avoiding unsupported claims that performance from one data modality automatically transfers to another.

Accordingly, the present study evaluates two parallel affective-computing pipelines. The first compares multiple pretrained CNN backbones, an attention-augmented VGG16 model, and a hard-voting ensemble for seven-class FER. The second compares BERT, LSTM, CNN, and SVM for binary sentiment analysis. The contribution is not a multimodal fusion architecture; rather, it is a dual-task experimental study that (i) quantifies the gain from attention and model ensembling in FER, (ii) compares transformer, recurrent, convolutional, and classical text classifiers under the same reported metrics, and (iii) discusses accuracy-efficiency trade-offs and methodological limitations relevant to practical affective-computing systems.

2. Materials and Methods

2.1 Study design

The study used a comparative experimental design with two independent supervised-learning tasks: seven-class facial emotion classification and binary text sentiment classification. Each task was trained and evaluated separately. The two pipelines were treated as independent unimodal experiments, and no direct cross-modal transfer between image and text representations was performed. Figure 1 illustrates the overall study workflow and the separation of the facial emotion recognition and sentiment analysis pipelines.

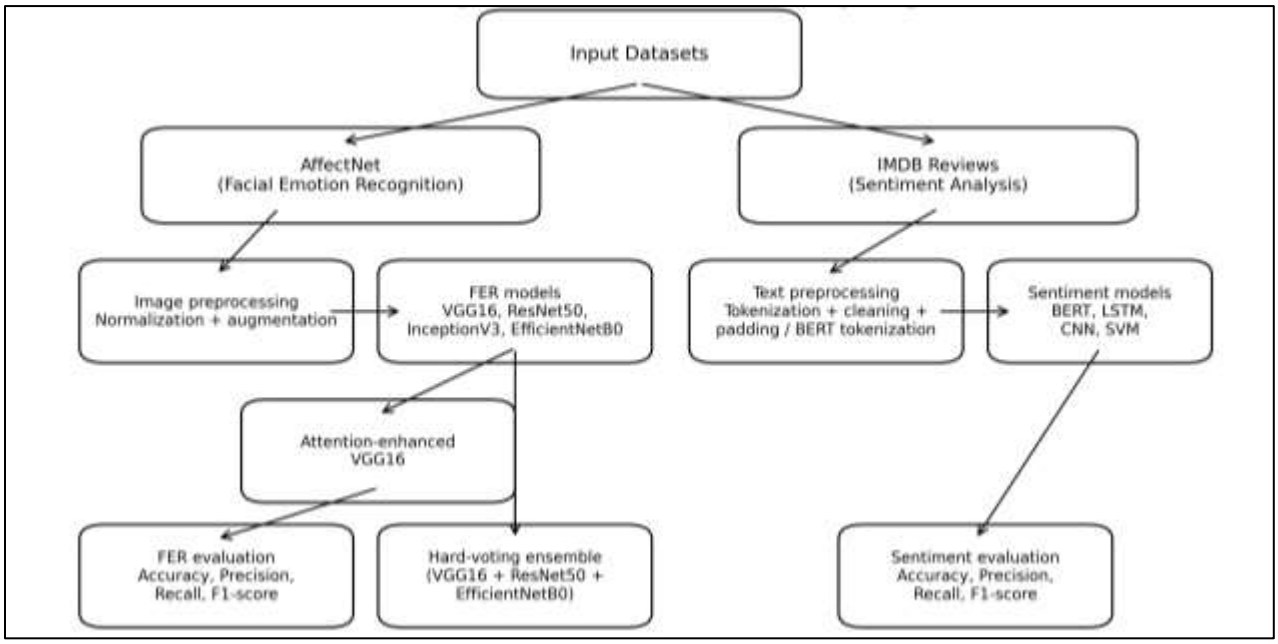

Figure 1. Overall workflow of the dual-task affective computing framework used in this study.

2.2 Datasets

Facial emotion recognition used a processed seven-class AffectNet derivative containing the categories Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise. AffectNet was originally created for in-the-wild affective computing and contains more than one million web-collected facial images, with a substantial manually annotated subset (Mollahosseini et al., 2019). The present experiments used the seven-class processed copy identified in the Data Availability Statement.

Sentiment analysis used the IMDB Large Movie Review Dataset, which contains 50,000 movie reviews with balanced positive and negative labels (Maas et al., 2011). The dataset is a widely used benchmark for binary sentiment classification and enables direct comparison between traditional, neural, and transformer-based text models.

2.3 Image preprocessing and augmentation

Image pixels were normalized before model training. Data augmentation was used to increase visual variability and reduce overfitting, including random rotation of up to 30 degrees, zoom, horizontal flipping, and random cropping. Because the tested pretrained backbones have different architectural input requirements, images were resized to an architecture-compatible input resolution before being passed to each network. This formulation avoids imposing a single resolution that would be incompatible with standard implementations of some backbones, particularly InceptionV3.

2.4 Text preprocessing

The non-transformer text pipeline included tokenization, stop-word removal, lemmatization, and sequence padding where required. Pretrained GloVe vectors were used to provide distributed word representations for the sequence-based neural models (Pennington et al., 2014). BERT was treated separately: it used its own subword tokenizer and contextual embeddings rather than GloVe, consistent with the original BERT architecture (Devlin et al., 2019). The SVM served as a conventional machine-learning baseline using the vectorized text representation produced by the preprocessing pipeline.

2.5 Facial emotion recognition models

Five individual deep-learning configurations were evaluated: VGG16, ResNet50, InceptionV3, EfficientNetB0, and VGG16 augmented with an attention mechanism. The pretrained CNN backbones were initialized from ImageNet weights and fine-tuned for seven-class emotion recognition. VGG16 provides a uniform stack of small convolutions; ResNet50 uses residual connections to support deeper optimization; InceptionV3 employs factorized, multi-scale convolutional blocks; and EfficientNetB0 uses compound scaling to balance network depth, width, and input resolution (He et al., 2016; Simonyan and Zisserman, 2015; Szegedy et al., 2016; Tan and Le, 2019).

For the attention-enhanced configuration, an attention layer was added to the VGG16-based representation so that the classifier could assign greater weight to expression-relevant features. A hard-voting ensemble then combined the class predictions of VGG16, ResNet50, and EfficientNetB0. The final ensemble class was the label receiving the majority of votes from the three component models.

2.6 Sentiment analysis models

Four models were compared on IMDB. BERT was fine-tuned for binary classification using its pretrained bidirectional transformer representation. LSTM provided a recurrent neural baseline for sequential dependencies (Hochreiter and Schmidhuber, 1997). A one-dimensional CNN was used to detect local discriminative patterns in token sequences, following the general principle of CNN-based sentence classification (Kim, 2014). SVM provided a traditional machine-learning baseline.

2.7 Training protocol and computational environment

Both datasets were divided into 80% training and 20% validation partitions. Neural image models used categorical cross-entropy, while binary neural text models used binary cross-entropy. Adam was used for neural-network optimization. The reported experiments used a batch size of 32 and up to 50 training epochs. SVM training followed its standard margin-based optimization and therefore was not governed by neural-network loss, batch, or epoch settings.

Implementation used Python with TensorFlow 2.x and Keras for deep-learning workflows, NumPy and pandas for data handling, and Matplotlib/Seaborn for visualization. Experiments were run in Google Colaboratory using an NVIDIA Tesla T4 GPU and approximately 16 GB of RAM. Hyperparameter selection was based on validation performance using search procedures where applicable. The complete search spaces and random seeds were not retained; this limitation is acknowledged because it restricts exact computational reproduction.

2.8 Evaluation measures

Model performance was summarized using accuracy, precision, recall, and F1-score. Approximate model-training time was also recorded as a hardware-dependent efficiency indicator. Reported precision, recall, and F1-score values are the class-aggregated values generated by the evaluation workflow. The analysis is descriptive: no formal hypothesis testing or confidence-interval estimation was performed, and no statistical-significance claims are therefore made.

3. Results and Discussion

3.1 Facial emotion recognition performance

Table 1 summarizes the FER results. The ensemble of VGG16, ResNet50, and EfficientNetB0 achieved the strongest overall performance, with 95.2% accuracy, 94.5% precision, 94.8% recall, and a 94.7% F1-score. Among the individual backbones, ResNet50 performed best with 94.1% accuracy. Adding attention to VGG16 increased accuracy from 92.4% to 94.6% and F1-score from 91.8% to 94.0%, corresponding to gains of 2.2 percentage points in both measures. The comparative trend is visualized in Figure 2.

Table 1. Performance of facial emotion recognition models on the seven-class AffectNet subset.

Model	Accuracy	Precision	Recall	F1-score	Training time (h)
VGG16 (baseline)	92.4%	91.5%	92.0%	91.8%	5.0
ResNet50	94.1%	93.4%	93.7%	93.6%	5.5
InceptionV3	93.5%	92.7%	92.5%	92.6%	6.0
EfficientNetB0	92.9%	92.0%	92.2%	92.1%	4.5
VGG16 + attention	94.6%	93.9%	94.1%	94.0%	6.5
Ensemble (VGG16 + ResNet50 + EfficientNetB0)	95.2%	94.5%	94.8%	94.7%	7.0

The ensemble improved accuracy by 2.8 percentage points and F1-score by 2.9 percentage points relative to baseline VGG16. The result is consistent with the rationale for ensemble learning: architecturally different networks can make partially uncorrelated errors, allowing a voting rule to improve robustness when their strengths are complementary. The attention-enhanced VGG16 result is also directionally consistent with recent FER studies in which spatial or channel attention is used to emphasize emotion-relevant information (Khan et al., 2025). Nevertheless, direct numerical comparison with published studies should be made cautiously because dataset subsets, class balance, train-test protocols, augmentation policies, and evaluation averaging can differ.

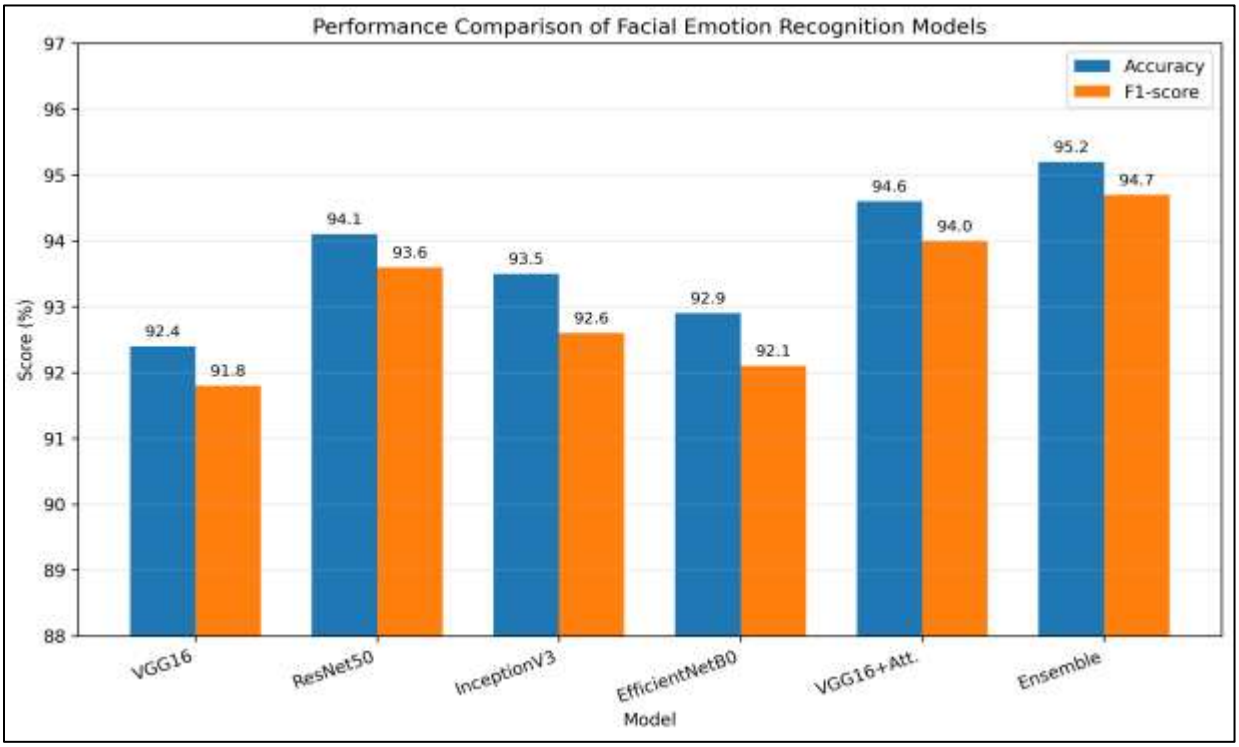

Figure 2. Accuracy and F1-score comparison for the facial emotion recognition models on the seven-class AffectNet subset.

3.2 Sentiment analysis performance

Table 2 reports the binary IMDB sentiment results. BERT achieved the best performance with 94.9% accuracy and a 94.6% F1-score. CNN ranked second at 91.2% accuracy, followed by LSTM at 89.8% and SVM at 85.6%. BERT therefore improved accuracy by 3.7 percentage points over CNN, 5.1 points over LSTM, and 9.3 points over SVM. Figure 3 presents the relative ranking of the sentiment models.

Table 2. Performance of sentiment analysis models on the IMDB Reviews dataset.

Model	Accuracy	Precision	Recall	F1-score	Training time (h)
BERT	94.9%	94.5%	94.8%	94.6%	3.0
LSTM	89.8%	89.0%	88.9%	88.9%	4.0
CNN	91.2%	90.4%	90.0%	90.2%	3.5
SVM	85.6%	85.0%	85.5%	85.3%	2.0

BERT's advantage is consistent with contextual transformer representations, which model token meaning jointly from left and right context and avoid the fixed-context limitations of many earlier text representations (Devlin et al., 2019; Vaswani et al., 2017). Earlier work by Al-Batah et al. (2018) demonstrated that hybrid neural and probabilistic models can improve sentiment classification over single-model baselines; the present findings extend that comparative perspective to a transformer-based English benchmark. The training-time values should be interpreted only within the reported Colab environment because runtime depends strongly on implementation, sequence length, batch size, hardware utilization, and whether pretrained representations are cached.

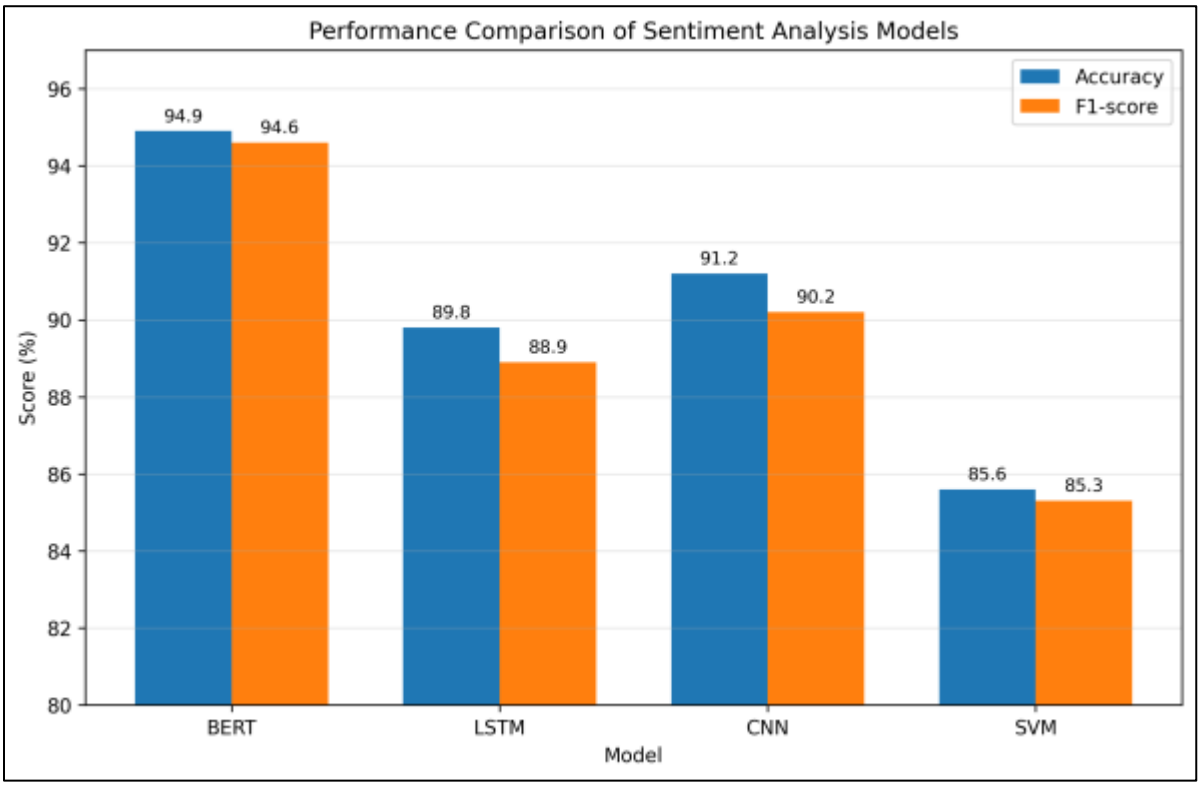

Figure 3. Accuracy and F1-score comparison for the sentiment analysis models on the IMDB Reviews dataset.

3.3 Accuracy-efficiency trade-offs and real-time inference

The highest-performing FER configuration required approximately 7 h of training, compared with 4.5-6.5 h for the individual or attention-augmented models. This illustrates a practical cost of ensembling: improved predictive performance is obtained at the expense of additional training, storage, and inference requirements. EfficientNetB0 offered the lowest reported FER training time (4.5 h) but did not match the ensemble's accuracy. A per-frame inference latency of approximately 50 ms was reported for the FER ensemble on the stated GPU environment. Under strictly serial processing, 50 ms per frame corresponds to an approximate theoretical throughput of 20 frames per second. Actual throughput may vary with preprocessing, batching, asynchronous execution, and hardware utilization. Further deployment studies should therefore report end-to-end latency, preprocessing time, batch size, warm-up procedure, and hardware-specific throughput separately.

3.4 Interpretation in the context of related work

The image results support two complementary design choices. First, attention improves a single backbone by emphasizing salient features. Second, ensemble voting combines distinct inductive biases from VGG16, ResNet50, and EfficientNetB0. Similar benefits of multi-model comparison and complementary learners have been observed in other applied classification studies (Al-Batah and Alabbas, 2026; Al-Batah and Alzboon, 2026; Alzboon and Al-Batah, 2026), although the domains and evaluation protocols differ and should not be treated as directly comparable.

The text results reinforce the strong performance of transformer encoders for contextual sentiment classification. At the same time, CNN, LSTM, and SVM remain useful baselines because they expose the value added by a large pretrained language model. This layered comparison is especially important in practical settings where deployment cost, interpretability, training time, and data volume may be as important as peak accuracy.

Although FER and sentiment analysis both belong to affective computing, they process fundamentally different modalities. Recent multimodal emotion-recognition literature combines face, speech, and text through explicit feature- or decision-level fusion (Lian et al., 2023). The present experiments do not implement such fusion; therefore, the strongest conclusion is that attention, ensemble learning, and transformer pretraining are effective within their respective image and text pipelines. A future multimodal system should train on paired observations and evaluate whether fusion improves performance beyond the unimodal components.

3.5 Limitations

Several limitations should be considered. First, the reported evaluation uses a single 80/20 partition rather than repeated cross-validation or an independent external test set, which limits uncertainty estimation. Second, complete hyperparameter search spaces, random seeds, and class-wise confusion matrices were not retained, preventing full computational reproduction and statistical comparison. Third, AffectNet is class-imbalanced in its original form, and a seven-class processed copy was used here; future work should report the exact retained sample count and class distribution. Fourth, training-time and latency values are hardware- and implementation-dependent. Fifth, the study evaluates image and text tasks in parallel and does not provide paired multimodal fusion or valid image-to-text domain adaptation. These limitations motivate leakage-controlled, seeded, externally validated experiments in future work.

4. Conclusion

This study presented a dual-task affective-computing evaluation for facial emotion recognition and text sentiment analysis. For FER, a hard-voting ensemble of VGG16, ResNet50, and EfficientNetB0 achieved the strongest reported performance, reaching 95.2% accuracy and a 94.7% F1-score, while attention improved the VGG16 baseline from 92.4% to 94.6% accuracy. For IMDB sentiment analysis, BERT achieved 94.9% accuracy and outperformed CNN, LSTM, and SVM. The findings show that model complementarity, attention, and pretrained contextual representations can improve predictive performance within their respective modalities. Because the image and text experiments were conducted as separate pipelines, the findings should not be interpreted as evidence of direct cross-modal transfer. Future work should use fully documented hyperparameter settings, repeated or nested validation, class-wise error analysis, calibrated probabilities, external datasets, and paired multimodal fusion to assess generalization and deployment readiness.

Data Availability Statement

The datasets used in this study are publicly accessible. The processed AffectNet derivative used for facial emotion recognition is available from Kaggle at: https://www.kaggle.com/datasets/fatihkgg/affectnet-yolo-format. The IMDB Reviews dataset used for binary sentiment analysis is available at: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews. The original IMDB benchmark is described by Maas et al. (2011), and the original AffectNet resource by Mollahosseini et al. (2019).

Ethics Statement

Not applicable. The study used secondary, publicly available datasets and did not recruit participants, collect new identifiable human data, or perform an intervention.

Conflict of Interest

The authors declare that they have no competing financial or personal interests that could have influenced the work reported in this manuscript.

Funding
The authors acknowledge the Deanship of Scientific Research at Jadara University for financial support related to this publication.

Acknowledgments
The authors are grateful to the Deanship of Scientific Research at Jadara University for providing financial support for this publication.

References

1. Al-Batah, M. S., and Alabbas, M. S. (2026). Stability-aware genetic algorithm feature selection and a calibrated hybrid ensemble for imbalanced smart-meter loss detection. Journal of Intelligent Decision Making and Information Science, 3(2), 833-854. https://doi.org/10.59543/jidmis.v3.1736
2. Al-Batah, M. S., and Alzboon, M. S. (2026). Sea animal image classification using machine learning algorithms for accurate and scalable prediction. Discover Artificial Intelligence, 6, 332. https://doi.org/10.1007/s44163-026-01005-9
3. Al-Batah, M. S., Mrayyen, S., and Alzaqebah, M. (2018). Arabic sentiment classification using MLP network hybrid with Naive Bayes algorithm. Journal of Computer Science, 14(8), 1104-1114. https://doi.org/10.3844/jcssp.2018.1104.1114
4. Alzboon, M. S., and Al-Batah, M. S. (2026). A clinically accessible heart disease prediction framework: Multi-model evaluation with ensemble learning and low-code deployment. Discover Applied Sciences, 8, 339. https://doi.org/10.1007/s42452-026-08353-2
5. Cortes, C., and Vapnik, V. (1995). Support-vector networks. Machine Learning, 20, 273-297. https://doi.org/10.1007/BF00994018
6. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019 (pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423
7. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778). https://doi.org/10.1109/CVPR.2016.90
8. Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735
9. Khan, T., Yasir, M., and Choi, C. (2025). Attention-enhanced optimized deep ensemble network for effective facial emotion recognition. Alexandria Engineering Journal, 119, 111-123. https://doi.org/10.1016/j.aej.2025.01.078
10. Kim, Y. (2014). Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (pp. 1746-1751). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1181
11. Kopalidis, T., Solachidis, V., Vretos, N., and Daras, P. (2024). Advances in facial expression recognition: A survey of methods, benchmarks, models, and datasets. Information, 15(3), 135. https://doi.org/10.3390/info15030135
12. Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., and Zong, Y. (2023). A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. Entropy, 25(10), 1440. https://doi.org/10.3390/e25101440
13. Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (pp. 142-150). Association for Computational Linguistics. https://aclanthology.org/P11-1015/
14. Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2019). AffectNet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 10(1), 18-31. https://doi.org/10.1109/TAFFC.2017.2740923
15. Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (pp. 1532-1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162
16. Simonyan, K., and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations. https://arxiv.org/abs/1409.1556

17. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the Inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2818-2826). https://doi.org/10.1109/CVPR.2016.308
18. Tan, M., and Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning (pp. 6105-6114). https://proceedings.mlr.press/v97/tan19a.html
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008. https://proceedings.neurips.cc/paper/7181-attention-is-all-you-need