

Cross-Dataset Movie Review Sentiment Classification Using GloVe-Enhanced BiLSTM Attention and Lightweight Hybrid Ensembles

Hongying Guo

*School of Computer Science, University of Technology Sydney, Sydney, Australia
n4218825tuis@163.com*

Abstract. Binary sentiment classification remains a useful test bed for evaluating the trade-off between sparse lexical models, neural sequence models, and lightweight ensembles. This study compares logistic regression, multinomial naive bayes, linear support vector machines, Random Forests, TextCNN, RNN, LSTM, BiLSTM, and BiLSTM with additive attention for English movie-review sentiment classification. The neural models are evaluated with and without 100-dimensional GloVe word vectors. Experiments use the 50,000-review IMDb dataset and the GLUE SST-2 benchmark. A validation partition is reserved for model selection and ensemble-weight selection, while final results are reported over five random seeds. On IMDb, a lightweight ensemble of logistic regression, SVM, and CNN probabilities achieved the best F1-score (0.9005 $\pm$ 0.0011) and ROC-AUC (0.9632 $\pm$ 0.0007). On the external SST-2 test set, BiLSTM with attention and GloVe was strongest, achieving an F1-score of 0.8431 $\pm$ 0.0078 and ROC-AUC of 0.9170 $\pm$ 0.0056. The ablation study shows that GloVe improves the attention model by 0.0243 F1 on IMDb. The findings demonstrate that lightweight lexical ensembles are highly effective in-domain, whereas attention-enhanced sequence modeling provides stronger cross-dataset generalization.

Keywords: sentiment analysis, GloVe, BiLSTM, attention mechanism, ensemble learning

1. Introduction

Sentiment analysis aims to infer attitudes, emotions, or evaluative polarity from natural-language text. It supports applications such as review mining, brand monitoring, recommendation, and decision support. Early work established that supervised classifiers can effectively distinguish positive and negative opinions when trained on domain-specific text [1]. However, performance depends strongly on the representation, the classifier, and the degree to which a model transfers beyond the data distribution on which it was trained.

Traditional machine-learning pipelines represent documents as sparse lexical features and remain strong baselines for review classification. In parallel, neural architectures learn distributed representations and can model local phrase patterns or longer sequential dependencies. TextCNN is effective at extracting salient n-gram-like features [2], while LSTM and BiLSTM networks can

model contextual dependencies across token sequences [3, 4]. Attention mechanisms extend sequence models by allowing the classifier to assign greater weight to informative hidden states [5].

This study addresses three questions: (1) How do traditional, neural, and hybrid methods compare on IMDb movie reviews? (2) Does a GloVe-initialized BiLSTM with attention improve external generalization on SST-2? and (3) Can the value of pretrained embeddings be quantified through a controlled ablation? The contribution is an experimentally reproducible comparison that separates validation-time selection from final testing and reports mean and standard deviation across five random seeds.

2. Related work

Sentiment classification has evolved from lexicon-based and bag-of-words approaches to representation learning. Foundational polarity-classification methods and the broader task of opinion mining are described in [1, 6]. Sparse TF-IDF features paired with linear classifiers remain competitive because sentiment cues often appear as discriminative words, phrases, and negations. Random Forests provide a nonlinear tree-based alternative [7], while SVMs maximize classification margins in high-dimensional feature spaces [8].

Neural approaches reduce manual feature engineering by learning embeddings and task-specific features. GloVe captures global word co-occurrence statistics in pretrained word vectors [9]. Convolutional filters are effective for sentence classification [2], and recurrent networks can aggregate sequential information. BiLSTM representations incorporate both left and right context, which is particularly useful when sentiment depends on modifiers or contrastive clauses. Attention adds a learned weighting operation over hidden states, making it possible to emphasize sentiment-bearing content while retaining the broader sequence context [5]. A broad survey of deep sentiment-analysis methods is available in [10].

Large pretrained transformers such as BERT have reset many benchmark records [11], but their computational cost can obscure the value of lighter, more interpretable baselines. This paper mainly focuses on a reproducible comparison among classical classifiers, compact neural models, pretrained word embeddings, and a lightweight ensemble. The study uses the IMDb corpus [12] and SST-2 from the Stanford Sentiment Treebank [13], distributed through the GLUE benchmark suite [14].

3. Methodology

3.1. Task definition and preprocessing

The task is binary classification: each review is assigned either negative sentiment (0) or positive sentiment (1). HTML markup was removed, HTML entities were decoded, nonalphabetic characters were normalized to spaces, and text was lowercased. The traditional models used TF-IDF unigram and bigram features with a maximum of 20,000 features, English stop-word removal, and a minimum document frequency of two. Neural models used a 20,000-word vocabulary, an unknown-token symbol, a padding token, and sequences truncated or padded to 200 tokens.

3.2. Traditional machine-learning baselines

Four baselines were trained with the same TF-IDF representation: logistic regression, multinomial naive bayes, linear SVM, and Random Forest. Logistic regression and SVM used balanced class weights. Logistic Regression used a maximum of 1,000 iterations; the Random Forest used 200

trees, balanced class weights, and all available CPU cores. These baselines test whether a proposed neural representation delivers improvement beyond high-quality lexical modeling.

3.3. Neural architectures

The neural comparison includes TextCNN, bidirectional vanilla RNN, unidirectional LSTM, BiLSTM, and BiLSTM with additive attention. The CNN applies filters of widths 3, 4, and 5 over embeddings and max-pools each feature map. The BiLSTM concatenates the final forward and backward states. The attention variant maps each bidirectional hidden state to a learned scalar relevance score, masks padding positions, normalizes the scores with softmax, and forms a weighted context vector from the resulting token weights. A sigmoid output layer converts the final representation into a positive-sentiment probability.

When enabled, neural embeddings were initialized from 100-dimensional GloVe vectors and fine-tuned during training. Dropout was 0.5, the hidden dimension was 128, the batch size was 64, the optimizer was Adam with a learning rate 0.001, and training lasted three epochs. The fixed budget is intentionally shared across neural architectures to make the comparison transparent; it is not a claim that every baseline has received exhaustive hyperparameter optimization.

3.4. Lightweight hybrid ensemble

The ensemble combines calibrated probability-like scores from Logistic Regression, SVM, and CNN through a nonnegative weighted sum whose weights total one. A grid of candidate weights was evaluated only on the validation split, and the selected combination was then evaluated once on the test split. This design avoids selecting ensemble weights directly on test data.

4. Experimental design

4.1. Data and splits

As shown in Table 1, the study used two dataset partitions for model training and evaluation. IMDb contains 50,000 highly-polarized movie reviews, balanced between positive and negative labels [12]. Each run used 20,000 examples for training, 5,000 for validation, and 25,000 for testing. SST-2 contains shorter movie-review sentences with binary labels [13]. Its 67,342 training examples were divided into 53,873 training and 13,469 validation cases, while the 872-example official development set served as an external test set. The five random seeds were 21, 42, 84, 123, and 2026.

Table 1. Dataset partitions used in the study

Dataset	Training	Validation	Test	Purpose
IMDb	20,000	5,000	25,000	In-domain evaluation
SST-2	53,873	13,469	872	External generalization

4.2. Metrics and statistical protocol

Accuracy, precision, recall, F1-score, and ROC-AUC were computed for every run. The F1-score was emphasized because it balances precision and recall, while ROC-AUC measures ranking quality across thresholds [15]. Training time, test prediction time, and serialized model size were recorded

as practical efficiency measures. McNemar tests were additionally generated for single-run paired comparisons on held-out predictions, following the standard use of paired classification outcomes [16, 17]. The primary reported values are five-seed means and standard deviations.

5. Results

5.1. IMDb in-domain results

Table 2 reports the highest-ranking IMDb methods. The lightweight hybrid ensemble achieved the strongest result, with $F1 = 0.9005 \pm 0.0011$ and $ROC-AUC = 0.9632 \pm 0.0007$. Logistic regression was the strongest individual baseline, demonstrating the continued importance of sparse lexical features for long, polarized movie reviews. Among neural models, BiLSTM with attention and GloVe was best, reaching $F1 = 0.8817 \pm 0.0033$ and $ROC-AUC = 0.9522 \pm 0.0018$. Its F1 was 0.0175 higher than the plain BiLSTM, indicating that attention contributed beyond bidirectional sequence encoding.

Table 2. Five-seed IMDb results (mean $\pm$ standard deviation)

Model	Accuracy	F1	ROC-AUC	Trains
Hybrid Ensemble	0.8994 $\pm$ 0.0012	0.9005 $\pm$ 0.0011	0.9632 $\pm$ 0.0007	17.39
Logistic Regression	0.8882 $\pm$ 0.0013	0.8897 $\pm$ 0.0013	0.9562 $\pm$ 0.0009	6.38
SVM	0.8844 $\pm$ 0.0010	0.8852 $\pm$ 0.0008	0.9541 $\pm$ 0.0005	7.06
BiLSTM + Attention (GloVe)	0.8764 $\pm$ 0.0065	0.8817 $\pm$ 0.0033	0.9522 $\pm$ 0.0018	3.91
CNN (GloVe)	0.8730 $\pm$ 0.0018	0.8743 $\pm$ 0.0017	0.9456 $\pm$ 0.0009	3.95

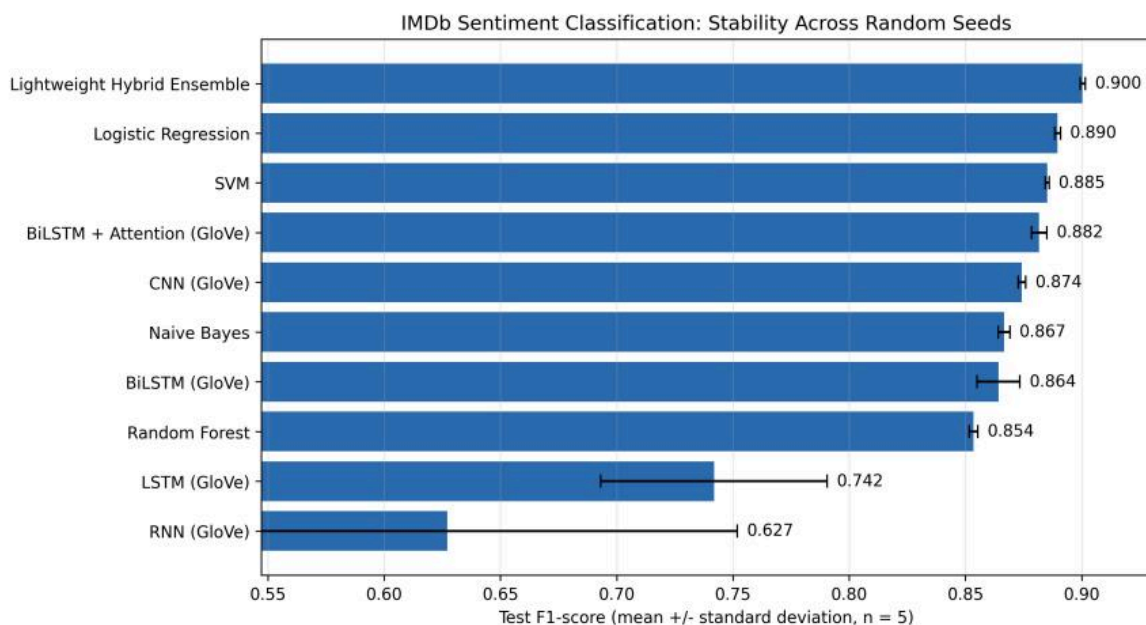


Figure 1. IMDb test F1-score across five random seeds

Figure 1 visualizes the stability of IMDb test F1-scores across five random seeds. The small error bars for the leading methods indicate that the hybrid ensemble, Logistic Regression, and SVM are not only strong but also relatively stable under repeated runs.

5.2. SST-2 external generalization

As shown in Table 3, the ranking changes on SST-2. BiLSTM with attention and GloVe achieved the best external F1-score (0.8431 $\pm$ 0.0078) and ROC-AUC (0.9170 $\pm$ 0.0056), followed by BiLSTM with GloVe (F1 = 0.8384 $\pm$ 0.0048). The hybrid ensemble and CNN were competitive but did not surpass the attention model. This pattern indicates that an attention-based sequence representation provides a useful inductive bias for shorter sentences that contain concentrated sentiment cues. The unidirectional LSTM was unstable under the shared three-epoch training budget and is retained as a negative control rather than omitted from the comparison.

Table 3. Five-seed SST-2 results (mean $\pm$ standard deviation)

Model	Accuracy	F1	ROC-AUC	Trains
BiLSTM + Attention (GloVe)	0.8385 $\pm$ 0.0059	0.8431 $\pm$ 0.0078	0.9170 $\pm$ 0.0056	10.30
BiLSTM (GloVe)	0.8319 $\pm$ 0.0046	0.8384 $\pm$ 0.0048	0.9119 $\pm$ 0.0037	9.40
Hybrid Ensemble	0.8158 $\pm$ 0.0145	0.8247 $\pm$ 0.0082	0.9009 $\pm$ 0.0023	9.97
CNN (GloVe)	0.8135 $\pm$ 0.0154	0.8231 $\pm$ 0.0086	0.8947 $\pm$ 0.0035	8.90
Logistic Regression	0.8000 $\pm$ 0.0025	0.8098 $\pm$ 0.0034	0.8783 $\pm$ 0.0023	0.45

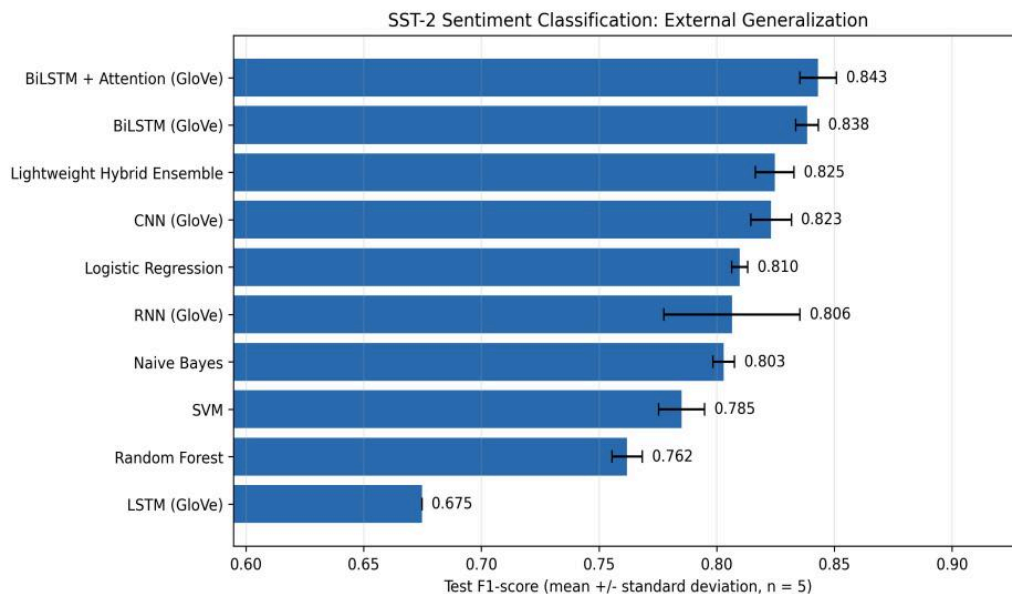


Figure 2. SST-2 external-test F1-score across five random seeds

Figure 2 further illustrates this change in model ranking on SST-2. Compared with the IMDb results, the attention-based BiLSTM with GloVe becomes the strongest method, which suggests that sequence modeling is more useful when the target texts are shorter and more context-sensitive.

5.3. GloVe ablation

As reported in Table 4, the controlled IMDb ablation isolates the effect of pretrained embeddings. GloVe increased F1 for LSTM by 0.1523, BiLSTM by 0.0710, CNN by 0.0360, and BiLSTM with attention by 0.0243. The attention model remained the strongest neural architecture both with and without GloVe, but its GloVe-initialized version was more accurate and less variable. The simple

RNN showed a negative mean change and high variance, demonstrating that pretrained vectors do not guarantee improvement when the sequence architecture itself is unstable.

Table 4. IMDb GloVe ablation over five seeds

Architecture	Random F1	GloVe F1	Delta F1
LSTM	0.5896 +/- 0.1483	0.7419 +/- 0.0487	+0.1523
BiLSTM	0.7932 +/- 0.0216	0.8642 +/- 0.0092	+0.0710
CNN	0.8383 +/- 0.0107	0.8743 +/- 0.0017	+0.0360
BiLSTM + Attention	0.8574 +/- 0.0069	0.8817 +/- 0.0033	+0.0243
RNN	0.6575 +/- 0.0536	0.6272 +/- 0.1247	-0.0303

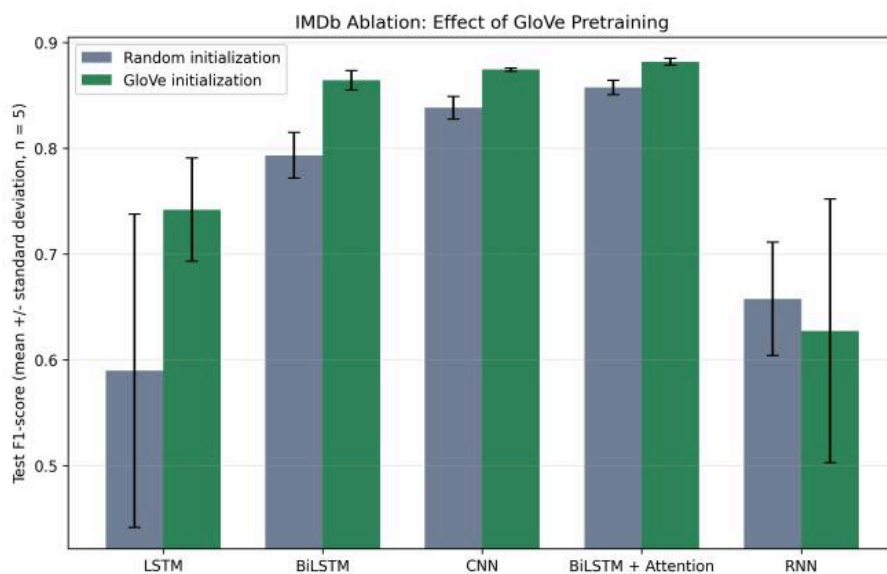


Figure 3. IMDb ablation of random and GloVe embedding initialization

Figure 3 provides a visual comparison between random embedding initialization and GloVe initialization. The improvement is most visible for LSTM-family models, while the RNN result remains unstable, confirming that pretrained embeddings are helpful but cannot fully compensate for weak sequence architecture or limited optimization.

6. Discussion

The results support a nuanced conclusion rather than a universal ranking. IMDb reviews are longer and strongly polarized, allowing the hybrid ensemble to exploit complementary lexical evidence from logistic regression, SVM and CNN. In contrast, SST-2 sentences are shorter and more context-sensitive. The BiLSTM-attention model appears better suited to this setting because it summarizes both directions of context while weighting discriminative token positions. The difference between the two datasets is informative: a model selected solely on IMDb would not necessarily be the best choice for external deployment.

The GloVe ablation provides direct evidence for pretrained lexical knowledge, especially for LSTM-family architectures. At the same time, the RNN result cautions against treating embedding initialization as a substitute for architecture choice and optimization. The unidirectional LSTM also displayed high variance or prediction collapse under the shared limited epoch budget. This is a

limitation of the current fixed-budget comparison, but it strengthens the reporting standard: unstable baselines remain visible rather than being silently removed.

Several limitations should guide future work. First, the datasets are English-only and binary, so results may not generalize to fine-grained, multilingual, or domain-shifted sentiment tasks. Second, model hyperparameters were intentionally held close to a common configuration rather than tuned exhaustively per architecture. Third, transformer baselines were not included in the main comparison. Future work should add a compute-matched transformer baseline, use validation-based Bayesian or grid search, test additional review domains, and evaluate calibration and robustness to negation, sarcasm, and mixed sentiment.

7. Conclusion

This paper presented a reproducible comparison of traditional machine learning, neural sequence models, GloVe initialization, and lightweight ensembling for movie-review sentiment classification. Five-seed evaluation showed that the hybrid ensemble was best on IMDb, while BiLSTM with attention and GloVe was best on the external SST-2 test set. The ablation confirmed that pretrained embeddings improve several neural architectures, including the proposed attention model. The central practical implication is that in-domain performance and external generalization should be reported together: sparse lexical ensembles remain exceptionally effective for polarized long reviews, whereas attention-enhanced sequence models offer a stronger choice when portability across sentiment datasets is required.

However, this study still has several limitations: the experiment conclusions may not directly generalize to multilingual, fine-grained, or domain-specific sentiment analysis tasks; the neural models may not have reached their best possible performance; this study did not include transformer-based models, which limit the completeness of the comparison; this paper mainly reports quantitative results, while qualitative error analysis is still limited. Future work could include more detailed analysis of misclassified reviews, especially examples involving negation, sarcasm, contrastive expressions, and mixed sentiment.

References

- [1] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), 1-135. <https://doi.org/10.1561/15000000011>
- [2] Kim, Y. (2014). Convolutional neural networks for sentence classification. *EMNLP*. <https://doi.org/10.3115/v1/D14-1181>
- [3] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- [4] Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11), 2673-2681. <https://doi.org/10.1109/78.650093>
- [5] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representations*. <https://arxiv.org/abs/1409.0473>
- [6] Liu, B. (2012). Sentiment analysis and opinion mining. Morgan & Claypool. <https://doi.org/10.2200/S00416ED1V01Y201204HLT016>
- [7] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
- [8] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273-297. <https://doi.org/10.1007/BF00994018>
- [9] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *EMNLP*. <https://doi.org/10.3115/v1/D14-1162>
- [10] Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *WIREs Data Mining and Knowledge Discovery*, 8(4), Article e1253. <https://doi.org/10.1002/widm.1253>

- [11] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL-HLT. <https://doi.org/10.18653/v1/N19-1423>
- [12] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. ACL. <https://aclanthology.org/P11-1015/>
- [13] 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. EMNLP. <https://aclanthology.org/D13-1170/>
- [14] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). GLUE: A multi-task benchmark and analysis platform for natural language understanding. International Conference on Learning Representations. <https://arxiv.org/abs/1804.07461>
- [15] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427-437. <https://doi.org/10.1016/j.ipm.2009.03.002>
- [16] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12, 153-157. <https://doi.org/10.1007/BF02295996>
- [17] Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation, 10(7), 1895-1923. <https://doi.org/10.1162/089976698300017197>