Your Group ID: Group 3

Your UNIs: xl3456, yz4661, lz3073

Your Full Names: Xiaole Liu, Evan Zhou, Leizu Zhang

Public GitHub Repo: https://github.com/DX0ZART/Group3

Model Performance and Hyperparameters

We evaluated a range of traditional machine learning and neural network models for sentiment classification on the SST-2 dataset. For traditional approaches, TF-IDF features were constructed using up to 5,000 features with unigram and bigram representations and a minimum document frequency threshold of two. Among linear models, Logistic Regression and Linear SVM achieved the strongest performance. Logistic Regression, tuned via 5-fold cross-validation, performed best with a regularization parameter of C=1, achieving an accuracy of 0.8188 and a ROC-AUC of 0.8918 on the test set. The Linear SVM, with an optimal C=0.1, slightly outperformed Logistic Regression in accuracy (0.8245) and F1 score (0.8317), though with a marginally lower ROC-AUC. In contrast, tree-based models such as Random Forest and Gradient Boosting showed weaker performance, with accuracies below 0.76, indicating that these models struggled to effectively leverage sparse, high-dimensional TF-IDF features.

Neural network models were trained using padded word sequences of fixed length (100 tokens) and a vocabulary capped at 20,000 words. A simple MLP with trainable embeddings achieved reasonable performance (accuracy 0.8028), while freezing the embeddings caused a dramatic performance drop, suggesting that learning task-specific word representations was crucial. Among deep models, the 1D-CNN achieved the best overall results, with an accuracy of 0.8234 and a ROC-AUC of 0.9142. This architecture benefited from convolutional filters that effectively captured local n-gram patterns and a global max-pooling layer that distilled the most informative features. Recurrent models showed mixed results: a standard LSTM failed to generalize and collapsed to near-random predictions, while a Bi-LSTM significantly improved performance, achieving an accuracy of 0.8119 and a strong ROC-AUC of 0.8971.

Experiments with pre-trained GloVe embeddings further demonstrated that transfer learning can be beneficial, particularly when embeddings are allowed to fine-tune during training.

Training Challenges and Mitigation Strategies

One major challenge encountered during training was the high dimensionality and sparsity of TF-IDF features. This issue particularly affected tree-based models, which are less suited for sparse feature spaces and prone to overfitting in this setting. To address this, we applied vocabulary pruning and relied primarily on linear classifiers with L2 regularization, which are known to perform well under such conditions. Cross-validation was used to carefully tune regularization strength and prevent overfitting.

Overfitting was also a concern for neural network models, especially those with large embedding layers and multiple dense units. This was mitigated through the systematic use of dropout, early stopping based on validation loss, and controlled embedding dimensions. Early stopping proved especially effective in preventing unnecessary training once validation performance plateaued, while dropout helped improve generalization by reducing co-adaptation of neurons.