Meta-Review Classification with Embedding and Preprocessing Strategies

This report presents a study on the classification of meta-reviews from ICLR submissions into binary decisions. We investigate the effectiveness of three neural classification pipelines: a baseline model using randomly initialized embeddings, a model employing GloVe pre-trained word vectors, and a final model integrating linguistic preprocessing techniques including lemmatization and part-of-speech (POS) tagging. Through detailed experimentation and error analysis, we evaluate models using accuracy, F1-score, Cohen's Kappa, confusion matrices, and visualization techniques such as t-SNE and PCA for interpretability. The study concludes that pre-trained embeddings, when combined with linguistically aware preprocessing, yield improvements in classification performance and semantic representation.

1. Data Exploration and Preprocessing

The dataset comprises 4,535 meta-reviews from the ICLR OpenReview archive, encompassing decisions and reviewer commentary from 2017 to 2020. Initial inspection revealed 98 entries with zero-length reviews, which were removed to preserve data integrity. Decision labels were grouped into three classes: as depicted in Fig. 1, the majority of meta-reviews are labeled as Reject (64.4%, n = 2856), followed by Accept (Poster) (29.2%, n = 1296), and Accept (Otherwise) comprising only 6.4% (n = 285). This severe class imbalance is crucial to acknowledge because it can bias learning algorithms to favor the dominant class, leading to poor generalization on minority labels. This distribution also informed our binary classification design. Specifically, we grouped both Accept (Poster) and Accept (Otherwise) under a broader Accept label, consolidating the task into a binary one. This decision helped mitigate sparsity issues in the rarest class while preserving the semantic distinction between acceptance and rejection signals.

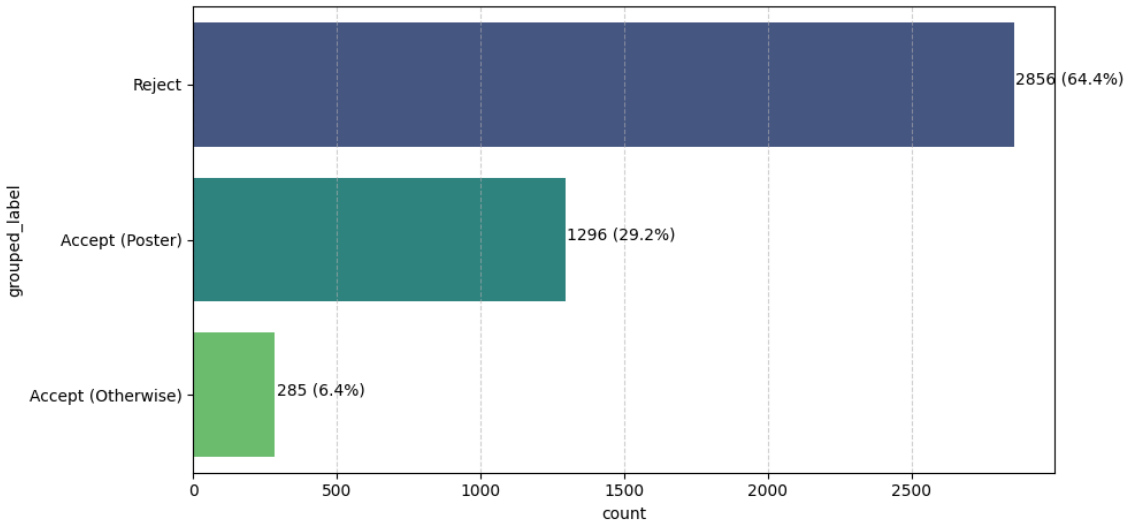

Figure 1- Grouped distribution of ICLR meta-review decisions

Furthermore, Table I summarizes the token length statistics of meta-reviews by decision type. Rejected papers receive the longest and most variable reviews (mean = 108 tokens), reflecting the need for detailed justification. Accept (Poster) reviews are slightly shorter (mean ≈ 95), while Accept (Otherwise) reviews are the briefest and most consistent (mean ≈ 92, std ≈ 56), indicating more concise affirmative feedback.

	count	mean	std	min	25%	50%	75%	max
Accept (Otherwise)	285.0	91.62	56.35	3.0	55.00	83.0	120.0	365.0
Accept (Poster)	1296.0	94.83	73.08	2.0	47.00	78.0	121.0	832.0
Reject	2856.0	107.93	83.69	2.0	53.75	87.0	136.0	703.0

Table 1 - Descriptive statistics of review length (in tokens)

For preprocessing, we implemented two pipelines to structure and enhance the meta-review texts. The first pipeline applied standard NLP techniques, including tokenization via the TweetTokenizer, lowercasing, and removal of domain-specific stopwords. Additional cleaning steps removed punctuation, URLs, numeric expressions, and email addresses using regular expressions, reducing input noise and sparsity. Lexical analysis following this pipeline revealed key patterns in the corpus. Fig. 2, Top 30 Most Frequent Words in Meta-Reviews, shows that terms like work, method, result, and concern dominate the vocabulary, reflecting the evaluative and methodological focus of reviews.

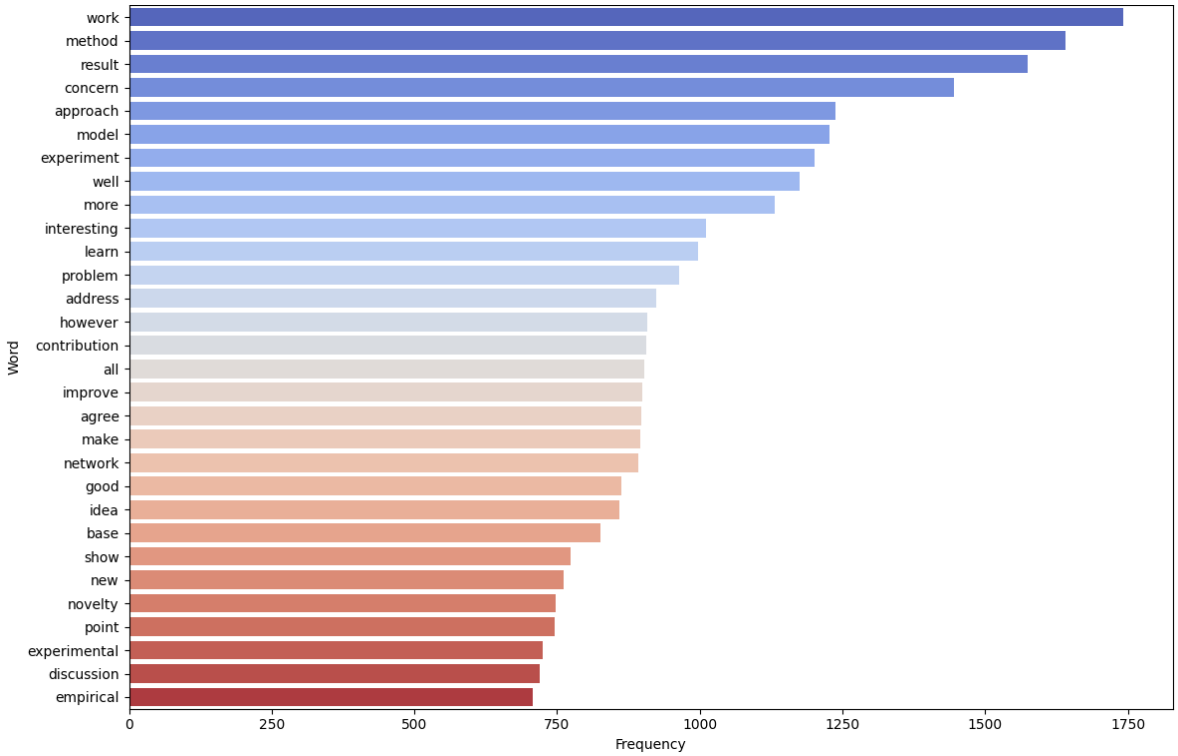

Figure 2 - Top 30 Most Frequent Words in Meta Reviews

Additionally, Zipfian Distribution of Word Frequencies (Log-Log Scale), confirms a classic long-tailed distribution. This validated our decision to retain only tokens with frequency ≥ 3 to preserve informative content while discarding noise.

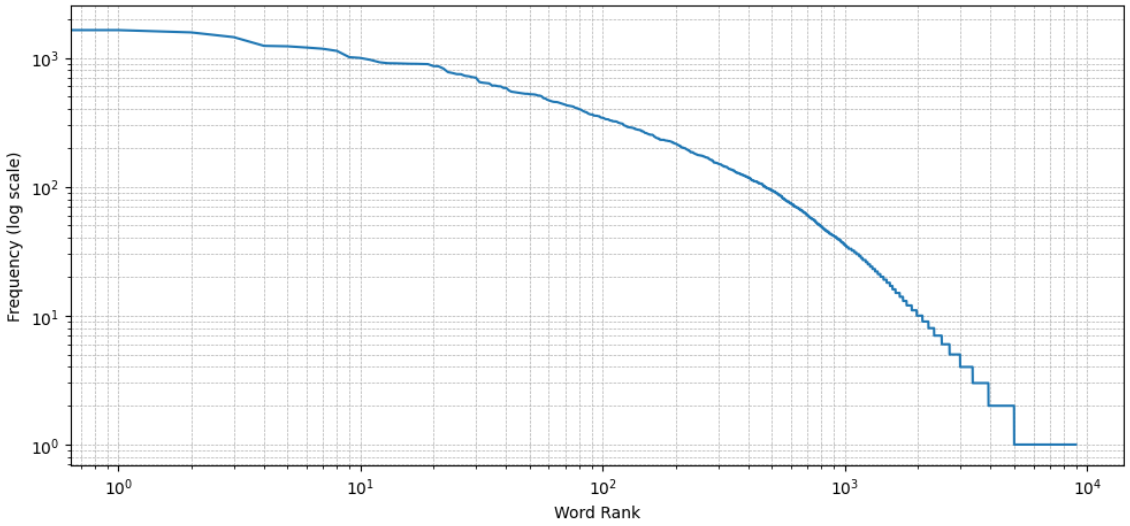

Figure 3 - Zipfian Distribution of Word Frequencies (Log-Log Scale)

The second pipeline introduced linguistically informed enhancements: POS-tagging followed by lemmatization using the WordNetLemmatizer, emoji and contraction handling, and negation marking with mark_negation to preserve polarity cues. This enriched representation improved semantic clarity without excessive computational cost. On average, the baseline pipeline processed reviews in 0.0018 seconds, while the enhanced version required 0.0024 seconds per review.

2. Language Modeling

2.1. Initialization and Architecture

To ensure reproducibility, all seeds were fixed at 42. We partitioned the dataset into training (70%), validation (15%), and testing (15%) sets, ensuring class stratification. A vocabulary was built from tokens with a minimum frequency of 3. Reviews were converted to indexed sequences and structured using a custom EmbeddingBagDataset. Offsets were used to manage variable-length sequences. Our architecture comprised an EmbeddingBag layer (dim=100) to efficiently represent sentences, followed by a 20-dimensional hidden layer activated with ReLU, and a final linear layer for binary classification. The loss function was BCEWithLogitsLoss, and optimization was conducted using Adam (learning rate: 0.01).

2.2. Training and Evaluation

The baseline model was trained for 10 epochs with early stopping based on validation loss. Upon evaluation on the held-out test set, the model achieved an overall accuracy of 73.7%, a macro-averaged F1-score of 61.0%, and a Cohen's Kappa coefficient of 0.41, indicating moderate agreement beyond chance. As illustrated in Fig. 4, the training and validation loss curves began to diverge noticeably after the third epoch.

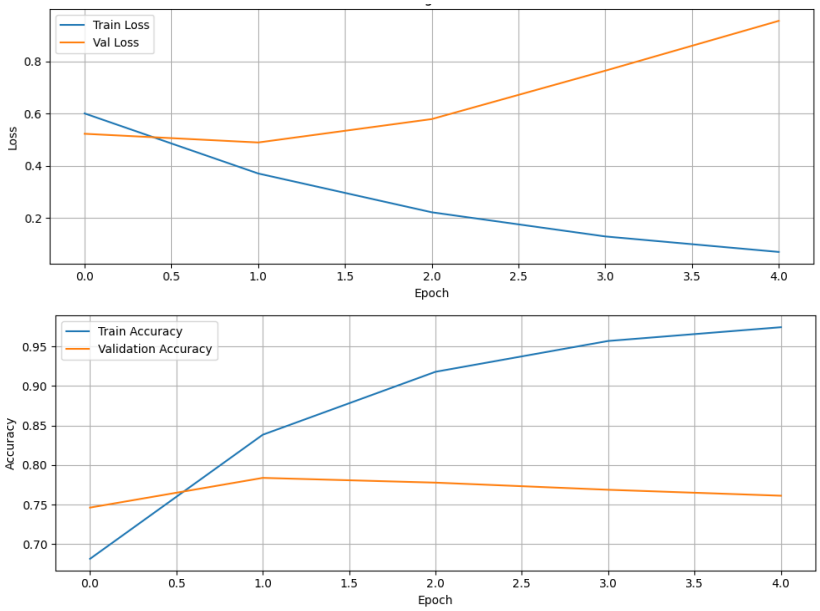

Figure 4 - Training and validation loss curves for the baseline model

The confusion matrix highlighted strong performance on the dominant class (Reject), but recall on Accept was suboptimal, indicating bias toward the majority class.

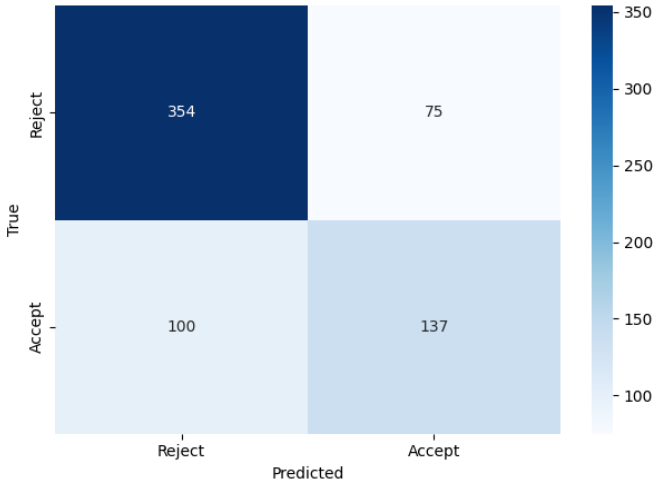

Figure 5 - Confusion Matrix for Baseline Model

2.3. Error Analysis and Embedding Interpretation

To assess model limitations, we analyzed false negatives—accepted papers misclassified as rejections. As shown in Table II, these instances often featured abstract or technically dense terminology (e.g., “borderline”, “adversarial”, “masked”), lacking clear positive sentiment cues. Confidence scores were extremely low (e.g., 5.7×10−7), indicating high model uncertainty. These examples suggest that the baseline model struggled to capture subtle or domain-specific approval signals embedded in scientific phrasing.

Example 424: True Class: 1.0, Predicted: 0.0, Confidence: 5.722470746150066e-07 Tokens: ['effect', 'training', 'image', 'classifier', 'masked']...
Example 385: True Class: 1.0, Predicted: 0.0, Confidence: 1.3710422308577108e-06 Tokens: ['analysis', 'different', 'method', 'noise', 'injection']...
Example 499: True Class: 1.0, Predicted: 0.0, Confidence: 1.0162687431147788e-05 Tokens: ['borderline']...
Example 665: True Class: 1.0, Predicted: 0.0, Confidence: 5.157291889190674e-05 Tokens: ['consider', 'adversarial', 'attack', 'deep', 'reinforcement']...
Example 622: True Class: 1.0, Predicted: 0.0, Confidence: 5.331670035957359e-05 Tokens: ['improve', 'quality', 'underwater', 'image', 'specifically']...

Table 2 - Examples of false negatives from the baseline model

To further evaluate semantic representation, we visualized the learned embeddings using t-SNE. As shown in Fig. 6, partial clustering of evaluative (e.g., “improve”, “significance”) and technical terms (e.g., “batch”, “normalization”) is observed. However, sentiment-laden tokens such as “unconvinced” and “novel” remain poorly separated, indicating limited semantic structure in the baseline embedding space. This supports the need for pre-trained embeddings to improve representational fidelity.

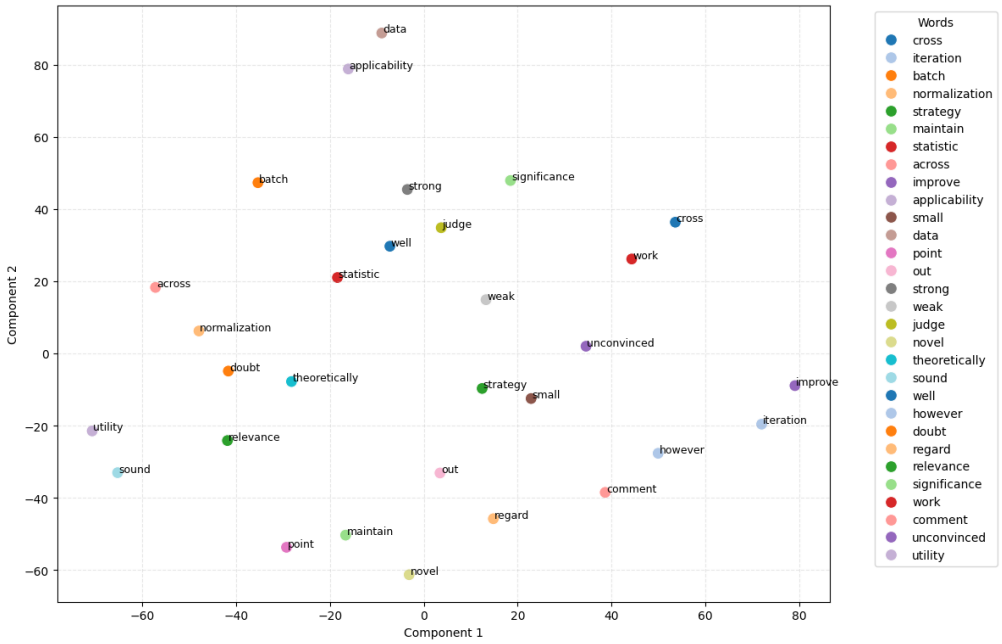

Figure 6 - t-SNE projection of top word embeddings from the baseline model

3. Pre-Trained GloVe Embeddings

To enhance semantic representation, we replaced the randomly initialized embedding layer with 100-dimensional GloVe vectors, integrated in a trainable (non-frozen) configuration. The architecture and loss function remained unchanged, but the learning rate was reduced to 0.001 to preserve pretrained semantic structure during fine-tuning.

This model outperformed the baseline across all metrics, achieving 76.6% accuracy, 65.0% F1 score, and a Cohen’s Kappa of 0.47. These gains reflect improved classification balance and stronger agreement

beyond chance. As shown in Fig. 7, training and validation curves demonstrated smoother convergence and reduced overfitting.

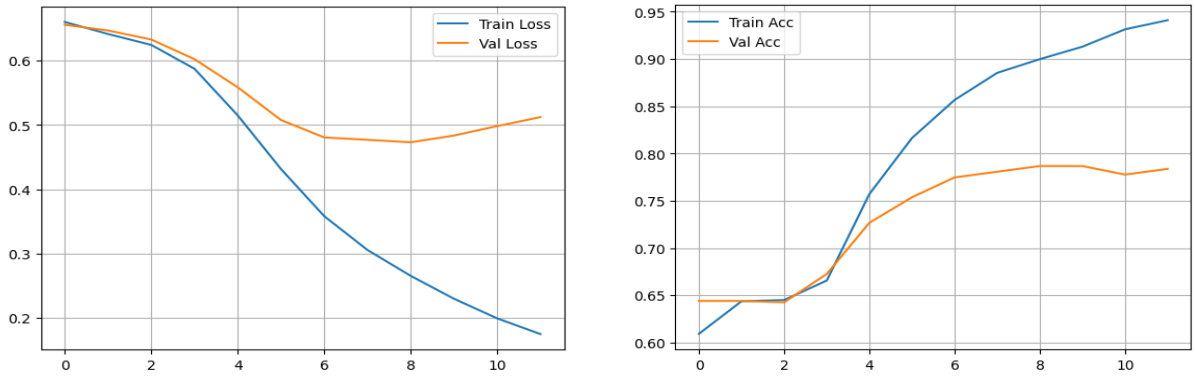

Figure 7 - Training and validation loss curves for the GloVe-based model

Compared to the baseline, the GloVe model exhibits a more balanced classification profile. It correctly identifies 145 Accept instances, a notable improvement over the baseline's performance, which suffered from higher false negative rates.

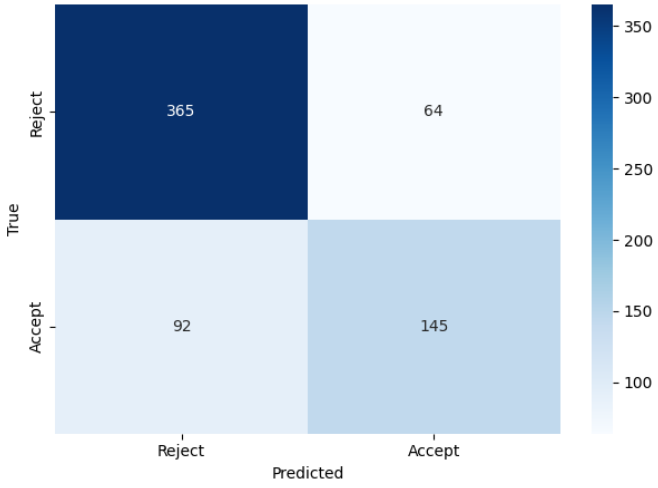

Figure 8 - Confusion matrix fo rthe GloVe-based model

Embedding quality also improved, as illustrated in Fig. 9, where t-SNE projections showed tighter clustering of semantically similar words like "result", "method", and "concern". Compared to the baseline, GloVe embeddings captured evaluative and technical relationships more effectively, supporting better downstream performance.

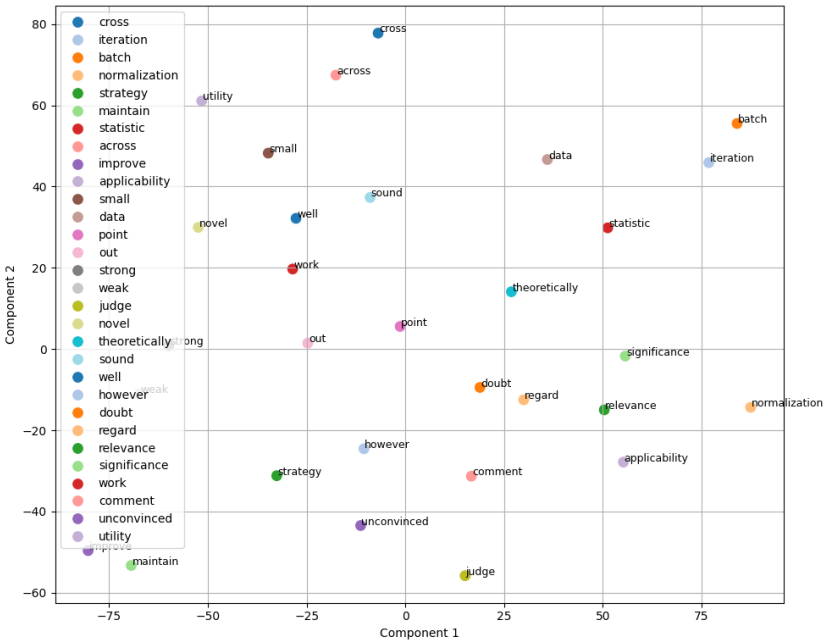

Figure 9 - t-SNE projection of GloVe-based model

4. Improved Preprocessing with GloVe

In the final model, we combined advanced linguistic preprocessing with pretrained GloVe embeddings to maximize semantic fidelity. Preprocessing included tokenization via TweetTokenizer, POS-aware lemmatization using WordNet, and filtering of non-informative elements such as punctuation, digits, emojis, and URLs. Crucially, negation handling was incorporated using the mark_negation utility to preserve sentiment-bearing constructions. The same model architecture and GloVe embeddings were retained, allowing us to isolate the effect of improved input representations. This setup yielded the strongest performance across all metrics: 78.5% accuracy, 66.0% macro F1 score, and a Cohen's Kappa of 0.51, reflecting enhanced generalization and reduced bias toward the dominant class. As shown in Fig. 10, the updated confusion matrix reveals better balance between true positive and true negative rates, with fewer borderline errors.

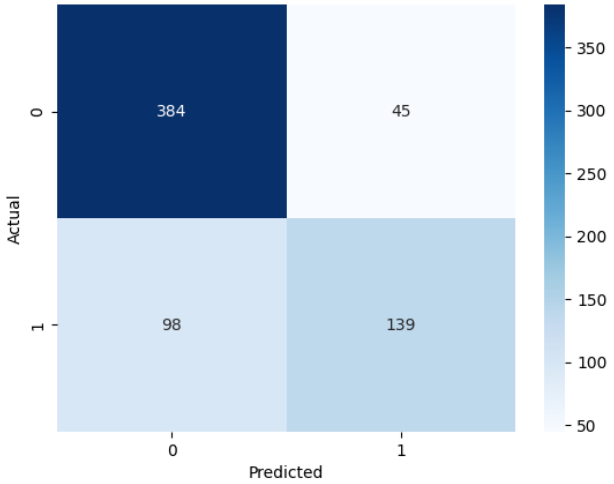

Figure 10 - Confusion matrix of the model with GloVe embeddings and enhanced preprocessing

Accuracy and loss trajectories for the final model closely mirrored those of the GloVe-only variant, demonstrating similarly stable convergence and generalization behavior. Given the near-identical performance trends, training curves are omitted for brevity. Notably, the final model yielded higher confidence scores (0.0029–0.0104) to false negatives containing abstract evaluative language—such as interpretability, reproducibility, and exploration bonus—suggesting improved alignment between semantic input features and the model's output space. These gains indicate that the enhanced preprocessing pipeline, when combined with GloVe embeddings, facilitated more effective encoding of implicit acceptance cues often present in technical and conceptually rich review text.

Example 173: True: 1.0, Pred: 0.0, Conf: 0.0029
Tokens: ['sat', 'np-complete', 'karp', 'due', 'intractable', 'exhaustive', 'search', 'heuristic', 'commonly', 'use']
Example 385: True: 1.0, Pred: 0.0, Conf: 0.0034
Tokens: ['paper', 'present', 'analysis', 'different', 'method', 'noise', 'injection', 'adversarial', 'example', 'use']
Example 125: True: 1.0, Pred: 0.0, Conf: 0.0052
Tokens: ['paper', 'author', 'extend', 'q-learning', 'ucb', 'exploration', 'bonus', 'jin', 'et', 'al']
Example 70: True: 1.0, Pred: 0.0, Conf: 0.0085
Tokens: ['paper', 'investigate', 'promising', 'direction', 'important', 'topic', 'interpretability', 'reviewer', 'find', 'variety']
Example 497: True: 1.0, Pred: 0.0, Conf: 0.0104
Tokens: ['paper', 'provide', 'careful', 'reproducible', 'empirical', 'comparison', 'graph', 'neural', 'network', 'model']

Table 3 - Examples of false negatives from the model with GloVe embeddings and enhanced preprocessing

Additionally, Fig. 11 presents a t-SNE projection of the GloVe-based word embeddings, revealing improved spatial organization relative to prior models. The embeddings exhibit tighter intra-cluster cohesion and greater inter-cluster separation, particularly among semantically aligned terms such as “significance”, “relevance”, and “reviewer”. This spatial coherence suggests that the model has internalized latent semantic structure more effectively, enabling it to distinguish evaluative expressions from domain-specific terminology.

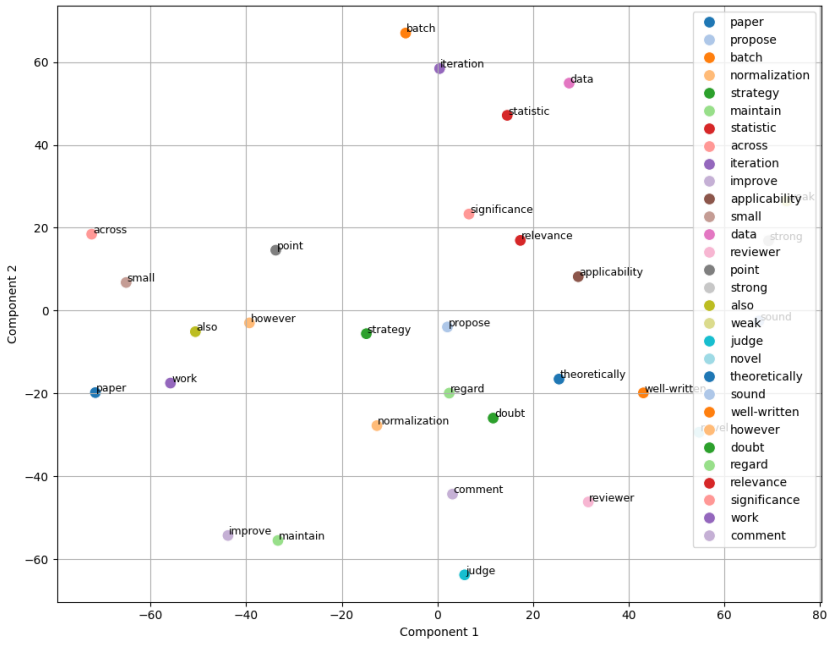

Figure 11 - t-SNE projection of the model with GloVe embeddings and enhanced preprocessing

5. Conclusion

The findings of this study indicate that the most effective gains in meta-review classification arise not merely from architectural complexity, but from careful refinement of input representations. While the final model architecture remained fixed, performance improvements were achieved through strategic manipulation of input semantics—first via external pretraining (GloVe) and then through linguistic normalization (POS-aware lemmatization with negation scope handling). This highlights the critical role of representation quality in low-complexity neural models. Remaining classification errors suggest the need for models that better capture contextual dependencies, particularly for borderline or abstract reviews. Future enhancements should explore contextualized embedding frameworks, adaptive loss functions to address class imbalance, and token-level attention mechanisms to prioritize evaluative clauses. Embedding analysis also suggests that proximity in semantic space correlates with prediction confidence, further motivating techniques that align embedding geometry with task-specific objectives.