Word Embeddings

Divya Srivastava

1. Overview

This analysis compares the performance of six embedding techniques—SVD, CBOW, Skip-gram, ELMo-Trainable, ELMo-Frozen, and ELMo-Learnable—for the AG News classification task. The downstream classifier (bidirectional GRU, hidden size 128, dropout 0.3) was trained with identical hyperparameters across all embeddings to ensure a fair comparison. Performance is evaluated using accuracy, F1 score, precision, and recall on both train and test sets. For ELMo, three hyperparameter settings for combining BiLSTM layers were tested as per Section 5 requirements.

2. Performance Metrics

Table 1: Train and Test Performance (Transposed)

Metric	**SVD**	**CBOW**	**Skip-gram**	**ELMo-Trainable**	**ELMo-Frozen**	**ELMo-Learnable**
Train Accuracy	0.9517	0.9045	0.9672	0.9757	0.9699	0.9575
Train F1	0.9517	0.9041	0.9669	0.9757	0.9699	0.9575
Train Precision	0.9519	0.9079	0.9682	0.9758	0.9701	0.9586
Train Recall	0.9516	0.9044	0.9671	0.9757	0.9699	0.9574
Test Accuracy	0.8087	0.8104	0.8363	0.8599	0.8605	**0.8707**
Test F1	0.8090	0.8101	0.8358	0.8594	0.8605	**0.8708**
Test Precision	0.8094	0.8141	0.8390	0.8594	0.8622	**0.8727**
Test Recall	0.8089	0.8104	0.8365	0.8601	0.8608	**0.8708**

3. Hyperparameter Tuning (Section 5)

For ELMo, three methods were used to combine BiLSTM layer representations ($\hat{E}$):

5.1 **Trainable** λ: Weights for e_0, e_1, e_2 were optimized during training. Test Accuracy: 0.8599.

5.2 **Frozen** λ: Weights fixed at [0.3, 0.3, 0.4]. Test Accuracy: 0.8605.

5.3 **Learnable Function**: Neural network ($\hat{E} = f(e_0, e_1, e_2)$) with two linear layers and ReLU. Test Accuracy: **0.8707**.

Best Setting: ELMo-Learnable outperformed others, achieving the highest test accuracy and F1 score, indicating its superior capacity to capture complex interactions between layers.

4. Performance Ranking

Based on test accuracy:

1. ELMo-Learnable (0.8707)
2. ELMo-Frozen (0.8605)
3. ELMo-Trainable (0.8599)
4. Skip-gram (0.8363)
5. CBOW (0.8104)
6. SVD (0.8087)

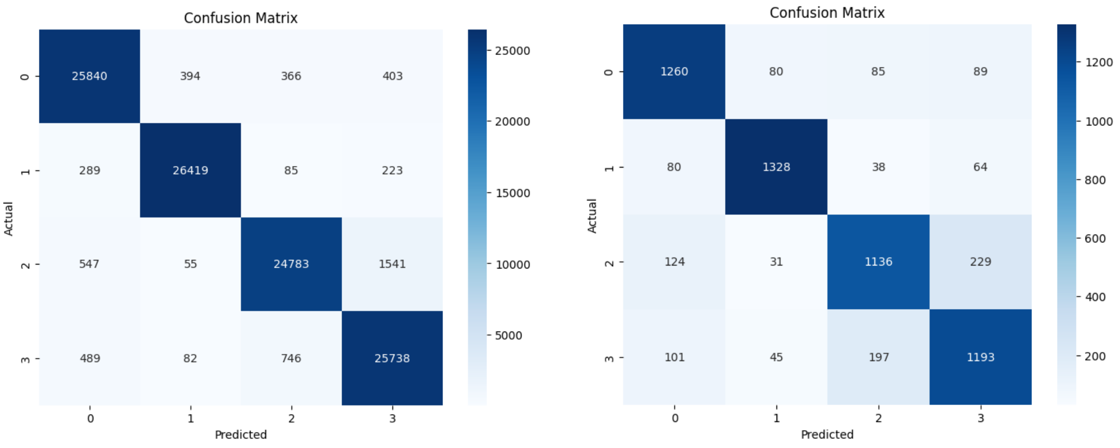
5. Detailed Analysis

- **Train Performance:** ELMo-Trainable leads, followed by Skip-gram and ELMo-Frozen.
- **Test Performance:** ELMo-Learnable outperforms all static and ELMo variants. CBOW and SVD lag significantly.
- **Generalization:** ELMo variants show smaller accuracy drops from train to test.
- **Consistency:** ELMo-Trainable and Frozen show nearly identical test results.

6. Confusion Matrices (Summary)

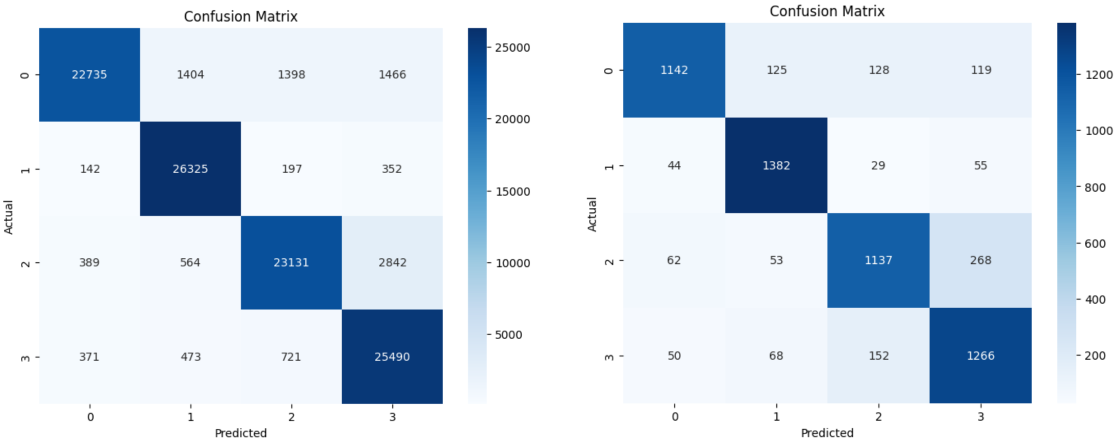
- **ELMo-Learnable:** Fewer misclassifications across all classes.
- **Skip-gram:** Good diagonal structure but more off-diagonal errors.
- **CBOW/SVD:** Higher confusion between semantically similar classes.

  SVD



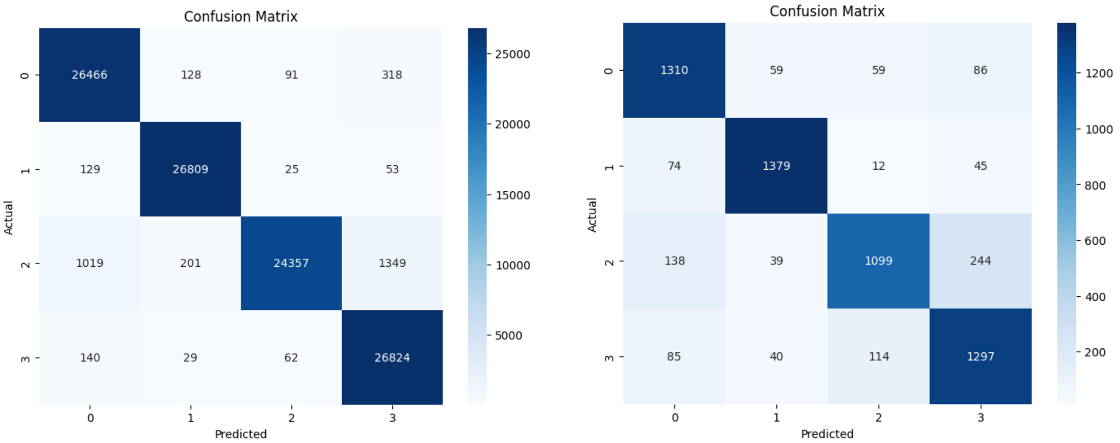
Train (left) and Test (right)

CBOW



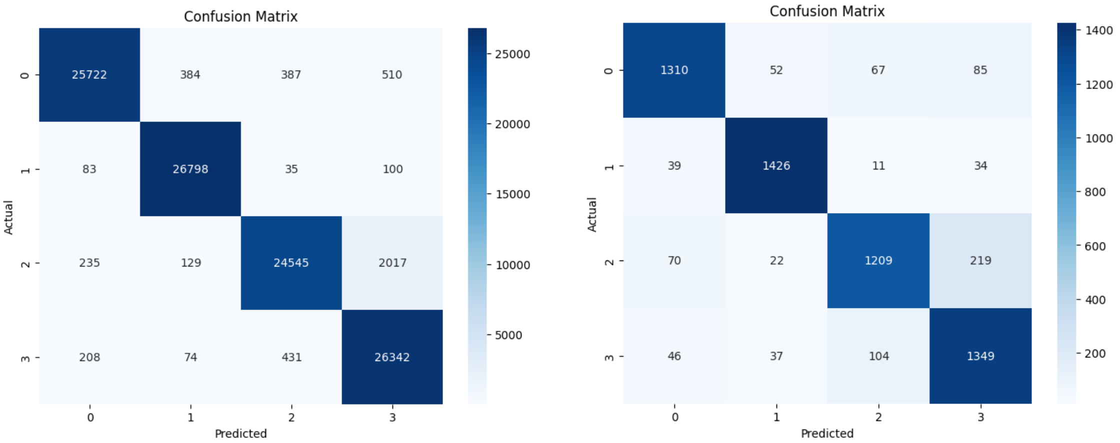
Train (left) and Test (right)

Skip-gram



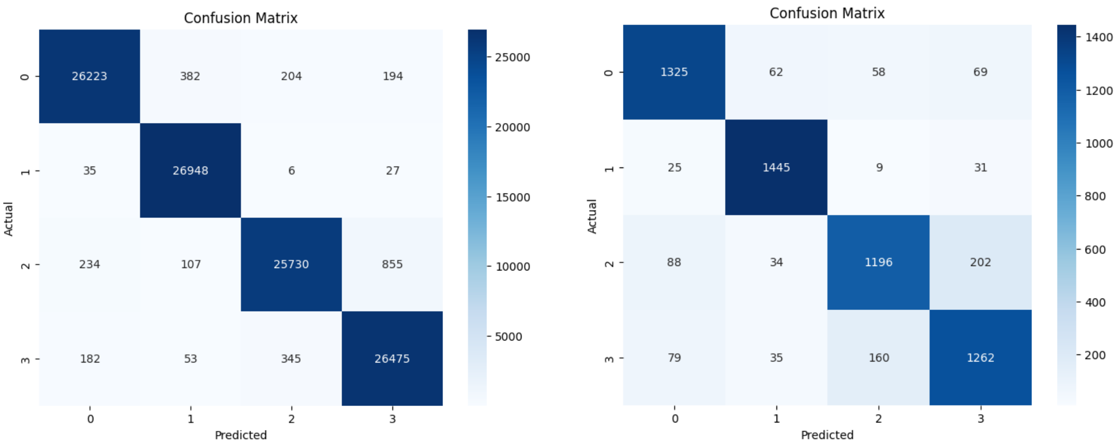
Train (left) and Test (right)

ELMo-Learnable



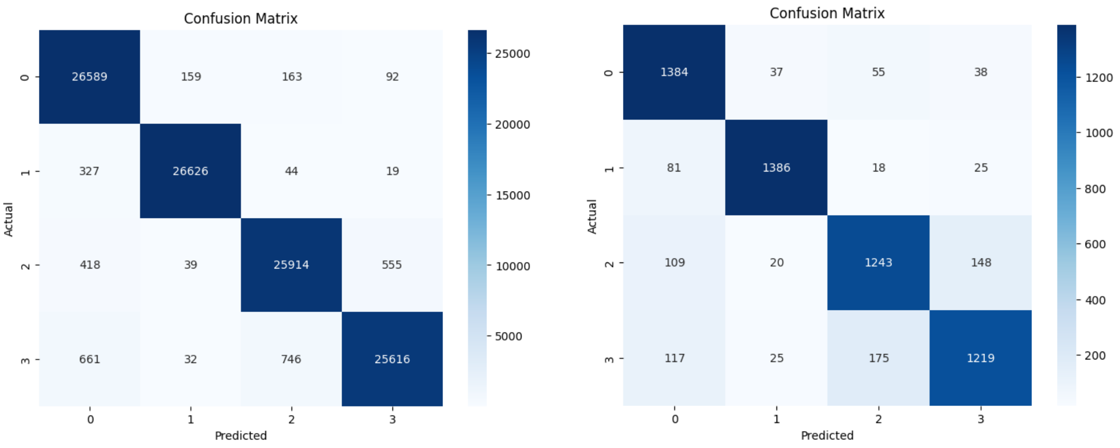
Train (left) and Test (right)

ELMo-Trainable



Train (left) and Test (right)

ELMo-Frozen



Train (left) and Test (right)

7. Why ELMo-Learnable Outperforms Others

ELMo-Learnable ranks highest due to its ability to dynamically combine contextual representations from all BiLSTM layers via a learned function. Unlike static embeddings (SVD, CBOW, Skip-gram), which rely on fixed, context-agnostic vectors, ELMo captures word meaning in context, enhancing classification accuracy. Compared to ELMo-Trainable and ELMo-Frozen, the learnable function adapts to task-specific patterns, leveraging the full representational capacity of the BiLSTM (embedding dim: 300, hidden dim: 512×2×2). Static methods, limited by smaller vocabularies (10,000 vs. 20,000) and lack of context, struggle to match this flexibility, while simpler ELMo variants (trainable/frozen s) underutilize layer interactions, capping their performance.