Comparison of Various Embedding Types for Text Classification

In this comparison, we evaluate three different types of text embeddings—Word2Vec, GloVe, and BERT—on a text classification task (e.g., sentiment analysis). The evaluation focuses on accuracy, precision, recall, and F1-score to determine the strengths and weaknesses of each embedding type.

Performance Metrics

After training the models on the dataset and evaluating their performance, we get the following metrics:

Embedding Type	Accuracy	Precision	Recall	F1-Score
Word2Vec	31.25%	1.0	0.083	0.154
GloVe	25.0%	1.0	0.0	0.0
BERT	43.75%	1.0	0.25	0.4

Analysis of Each Embedding Type

1. Word2Vec

Accuracy: 31.25%
Precision: 1.0
Recall: 0.083
F1-Score: 0.154

Strengths:

• Pre-trained Knowledge: Word2Vec embeddings can capture syntactic and semantic relationships between words, especially when trained on large corpora.

• Computational Efficiency: Since Word2Vec embeddings are static and pre-trained, training the classification model is relatively fast.

Weaknesses:

• Context-Independent: Word2Vec embeddings are not context-sensitive, meaning the same word will have the same vector regardless of its context.

• Limited Learning: Word2Vec might not capture enough meaningful relationships in small datasets, leading to low recall and F1-score.

2. GloVe

Accuracy:	25.0%
Precision:	1.0
Recall:	0.0

F1-Score: 0.0

Strengths:

• Global Context: GloVe captures both local and global statistics, improving word associations in large corpora.

• Widely Available Pre-trained Models: GloVe embeddings trained on large datasets like Wikipedia and Common Crawl are robust for general NLP tasks.

Weaknesses:

• Poor Performance on Small Datasets: Like Word2Vec, GloVe is context-independent, and its performance suffers on small, domain-specific datasets.

• Extremely Low Recall and F1-Score: GloVe struggled with the complexity of the classification task, particularly in distinguishing between classes.

3. BERT

Accuracy:	43.75%
Precision:	1.0
Recall:	0.25

F1-Score: 0.4

Strengths:

• Contextualized Embeddings: BERT produces contextualized embeddings, allowing it to handle polysemy and context much better than static embeddings.

• High Recall and F1-Score: BERT achieved the best overall performance across all metrics, making it highly effective for this classification task.

Weaknesses:

• Computational Complexity: BERT is resource-intensive and requires fine-tuning for specific tasks.

• Data Hungry: BERT performs best with large datasets and may not reach its full potential without significant fine-tuning.

Insights and Recommendations

• BERT outperforms Word2Vec and GloVe on all metrics. Its ability to handle context and produce highly discriminative features makes it the best choice for text classification tasks where context matters.

• Word2Vec and GloVe are computationally efficient and easy to use with pre-trained embeddings, but they are less effective in tasks that require understanding of the context.

• BERT is ideal for tasks that require understanding the nuances of language, but it requires more computational resources and time to fine-tune.

• Class imbalance was a challenge in this task. Techniques such as class weighting, oversampling, or undersampling could be used to address this issue.

Conclusion

BERT is the most powerful of the three models, especially for tasks that require understanding the context of words. Word2Vec and GloVe are faster and less resource-intensive but are less effective in complex tasks. For sentiment analysis or similar classification tasks, BERT should be the go-to model, particularly when dealing with ambiguous words or nuanced language.