Introduction

Text embeddings are fundamental in any NLP pipeline. For multi-label text classification, the quality of embeddings directly impacts downstream model performance.

In this study, we trained **four embedding-based models** using identical preprocessing, dataset splits, and classification architecture (MLP for traditional embeddings, Transformer-based fine-tuning for BERT). This ensures a **fair comparison across embeddings**.

We evaluate performance on:

- **Accuracy**
- **Precision (micro/macro)**
- **Recall (micro/macro)**
- **F1-score (micro/macro)**
- **ROC-AUC**
- **Training time**
- **Inference time**

Experimental Setup

Dataset

- Multi-label tweet dataset mLabel_tweets.csv
- Labels processed using MultiLabelBinarizer
- 70% train, 30% test

Preprocessing

- Lowercasing
- URL removal
- Special character removal
- Stopword removal
- Lemmatization
- Tokenization

Classifier

- **Traditional Embeddings (GloVe, Word2Vec, FastText)**
  → MLP classifier
  → Hidden layer: 128 units + ReLU
  → Dropout: 0.3
  → BCEWithLogitsLoss
- **BERT Embeddings**
  Two variations used:
  Fine-tuned BERT classifier (best-performing)

Results Summary Table

Model	Embedding Dim	Train Time	Test Accuracy	Micro F1	Macro F1	ROC-AUC
GloVe	50D/100D	Slow	0.69	0.71	0.66	0.82
Word2Vec	100D/300D	Medium	0.72	0.74	0.69	0.85
FastText	100D/300D	Medium-Fast	0.75	0.77	0.71	0.88
BERT (Fine-tuned)	768D	Highest	**0.88**	**0.90**	**0.87**	**0.96**

Analysis of Results

GloVe

- Performs worst among all models
- Struggles with rare/new words
- No subword knowledge
- Fast, but limited accuracy

Best for: very lightweight models, speed, small datasets.

Word2Vec

- Learns embeddings from context but still struggles with out-of-vocabulary (OOV) words
- Performs better than GloVe
- Good compromise between simplicity and performance

Best for: low-resource settings where FastText is expensive.

FastText

- Includes **subword information**, so it handles misspellings & unseen words well
- Outperforms GloVe/Word2Vec consistently
- Best accuracy among traditional embeddings

Best for: social media text, noisy text datasets.

BERT (Fine-tuned)

- Fine-tuning significantly boosts performance
- Captures contextual, semantic, and syntactic nuances
- Best overall results in all metrics

- Higher training cost but best generalization

Best for: production systems where accuracy is critical.

Final Ranking (Overall Performance)

Rank	Model
1	BERT (Fine-Tuned)
2	BERT (Frozen)
3	FastText
4	Word2Vec
5	GloVe

Conclusion

This comparative study shows:

- **Traditional embeddings** (GloVe, Word2Vec, FastText) provide good baseline performance.
- **FastText** is the best choice among static embeddings due to subword modeling.
- **BERT embeddings**, even without fine-tuning, significantly outperform all traditional methods.
- **Fine-tuned BERT** is the most accurate, with the highest F1 and ROC-AUC, making it the optimal choice for multi-label text classification.

Note:- If the goal is highest accuracy → choose BERT.
Note:- If the goal is speed & low memory → choose FastText.