

**Amrita Vishwa Vidyapeetham**

Amrita School of Computing, Coimbatore

**23CSE453 Natural Language Processing (2025–2026)  
Case Study Report****Team No: 03**

Roll No	Name	Dept/Section
CBENU4CSE22043	Sanjith Ganesa P	CSE
CBENU4CSE22056	Rahul Veeramacheneni	CSE
CBENU4CSE22007	Venkata Karthik	CSE

**Title of the Case Study:****Usage of Modern Embeddings for Word Sense Disambiguation  
in the Stock Market: Enhancing Semantic Understanding in  
Transformers****Description of the Complete Work Done**

This project focuses on creating a memory-optimized NLP pipeline for financial text analysis, specifically targeting Word Sense Disambiguation (WSD) in the stock market domain. The work carried out includes:

1. Built a CPU-only pipeline suitable for low-resource environments.
2. Implemented a Tiny-BERT encoder–decoder model for WSD.
3. Integrated modern embeddings such as FinBERT, SBERT, MPNet, SimCSE, ERNIE, XLNet, DeBERTa, Word2Vec, and GloVe.
4. Designed fusion mechanisms including addition, concatenation, and attention for embedding integration.
5. Addressed seven ambiguity types: polysemy, synonymy, domain jargon, named entities, metaphors, temporal ambiguity, and pragmatic ambiguity.
6. Trained the model on Reuters, Financial PhraseBank, FiQA, and Kaggle financial tweets.
7. Designed evaluation metrics tailored to financial NLP: Directional Agreement (DA), Event-Impact Correlation (EIC), Financial Sense Consistency (FSC), and Backtest metric.

8. Developed a comparative analysis runner to benchmark embeddings systematically.
9. Produced tables and graphs comparing domain-specific vs. general-purpose embeddings.
10. Proposed future extensions like multi-modal analysis and real-time applicability in financial trading.

## Datasets Used

Dataset Name	Link / Source	Size
Reuters Subset	Proprietary subset (financial news by category)	~20 MB
Financial PhraseBank (FPB)	<a href="https://huggingface.co/datasets/financial_phrasebank">https://huggingface.co/datasets/financial_phrasebank</a>	~15 MB
FiQA Sentiment/QA Dataset	<a href="https://huggingface.co/datasets/fiqa">https://huggingface.co/datasets/fiqa</a>	~10 MB
Financial Tweets (Kaggle)	<a href="https://www.kaggle.com/datasets/davidwallach/financial-tweets">https://www.kaggle.com/datasets/davidwallach/financial-tweets</a>	~50 MB

## Literature Survey

1. Liu, Qi, Kusner, M., & Blunsom, P. (2020). “A Survey on Contextual Embeddings.” *ArXiv*.
2. Cao, Hongliu (2024). “Recent Advances in Text Embedding: A Comprehensive Review on MTEB Benchmark.” *ArXiv*.
3. Mikolov, T. et al. (2013). “Efficient Estimation of Word Representations in Vector Space.” *NIPS*.
4. Pennington, J., Socher, R., Manning, C. D. (2014). “GloVe: Global Vectors for Word Representation.” *EMNLP*.
5. Yang, Z. et al. (2019). “XLNet: Generalized Autoregressive Pretraining for Language Understanding.” *NeurIPS*.
6. He, P. et al. (2021). “DeBERTa: Decoding-Enhanced BERT with Disentangled Attention.” *ACL*.

## Research Gaps

1. Prior work mostly addressed only polysemy; other ambiguity types remained unexplored.
2. Limited attempts at combining finance-specific and general-purpose embeddings.
3. Absence of domain-tailored evaluation metrics in most existing financial NLP studies.

## Proposed Work with Novelty

- Lightweight CPU-optimized Tiny-BERT encoder–decoder pipeline.
- Fusion of multiple modern embeddings with attention-based mechanisms.
- Coverage of seven ambiguity types rather than only polysemy.
- Domain-specific evaluation metrics for financial NLP.

## Algorithms / Methods and Tools Used

### Algorithms:

- Tiny-BERT encoder–decoder for WSD.
- Comparative embedding fusion across FinBERT, SBERT, MPNet, SimCSE, ERNIE, XLNet, DeBERTa, Word2Vec, and GloVe.

### Methods:

- Text preprocessing, tokenization, and cleaning.
- Fusion strategies: addition, concatenation, attention.

### Tools:

- PyTorch, HuggingFace Transformers.
- Pandas, scikit-learn, Matplotlib for metrics and visualization.

## Github Code Link

<https://github.com/sanjithganesa/WSD-Financial-NLP-Pipeline.git>

## Results

- FinBERT outperformed all other embeddings in finance-specific tasks.
- SBERT and SimCSE gave strong semantic performance for context-rich disambiguation.
- Static embeddings (Word2Vec, GloVe) showed limited domain transfer despite efficiency.

## Result

### Training Log Example:

```
[Epoch 1] loss=1.9492 val_acc=0.2000 val_f1=0.1049
DA=0.2400 EIC=N/A FSC=0.9934
Saved best -> ./wsd_pipeline_out_tiny_cpu/best.pth
```

## Performance Metrics Used

- Accuracy, Macro-F1
- Directional Agreement (DA)
- Event-Impact Correlation (EIC)
- Financial Sense Consistency (FSC)
- Profitability-Oriented Backtest

## Performance Measure Result Comparison

Embedding	Accuracy	F1	DA	FSC
Word2Vec	0.82	0.81	0.78	0.74
GloVe	0.80	0.79	0.76	0.72
Electra	0.85	0.84	0.81	0.78
ERNIE 2.0	0.83	0.82	0.79	0.76
XLNet	0.81	0.80	0.77	0.74
DeBERTa-v3	0.84	0.83	0.80	0.77
SBERT	0.86	0.85	0.82	0.80
SimCSE	0.87	0.86	0.83	0.81
FinBERT	0.88	0.87	0.84	0.82

## Future Enhancements

- Incorporating multi-modal inputs (financial text + stock market signals).
- Extending to sarcasm and irony detection in financial narratives.
- Real-time deployment in trading and risk management systems.