Subword Tokenization and CBOW Embeddings for Financial Sentiment Classification

1 Introduction

This project implements a complete Natural Language Processing (NLP) pipeline from scratch, including:

- Subword tokenization using Byte Pair Encoding (BPE)
- Continuous Bag-of-Words (CBOW) embedding training
- Sentiment classification using Logistic Regression
- Comparison against the VADER rule-based baseline

Word embeddings were trained on 50,000 sentences from Simple English Wikipedia. The learned representations were evaluated on the Financial PhraseBank (AllAgree subset) using Macro F1 score.

2 Tokenizer Selection: BPE vs WordPiece

2.1 Motivation for Subword Tokenization

Word-level tokenization suffers from large vocabulary size, out-of-vocabulary (OOV) issues, and poor handling of morphological variations such as:

- profit
- profitable
- profitability

Subword tokenization allows sharing of morphological components and improves generalization.

2.2 Why BPE?

Byte Pair Encoding (BPE) iteratively merges the most frequent adjacent character pairs to construct a subword vocabulary.

Advantages of BPE in this project:

- Deterministic and simple to implement
- Frequency-based merging suitable for medium-sized corpora
- Efficient vocabulary construction
- Reduced OOV problem

2.3 Why Not WordPiece?

WordPiece selects merges based on likelihood improvement rather than raw frequency.

It was not selected due to:

- Greater implementation complexity
- Additional probabilistic scoring overhead
- Limited practical gain for the dataset size used

BPE provides a balanced trade-off between simplicity and representational power.

3 Embedding Model

3.1 CBOW Architecture

The Continuous Bag-of-Words (CBOW) model predicts a target token from its surrounding context tokens.

Key hyperparameters:

- Embedding dimension: 384
- Context window: 5
- Negative samples (K): 5
- Batch size: 512
- Unigram smoothing exponent: 0.75

Negative sampling approximates full softmax and significantly reduces computational cost. The unigram distribution raised to the 0.75 power reduces dominance of extremely frequent tokens while preserving corpus frequency structure.

Embeddings were trained exclusively on Wikipedia and were not exposed to financial sentiment labels.

4 Sentence Representation

Each financial headline was converted to a fixed-length vector using:

4.1 Mean Pooling

The sentence embedding was computed as the average of its token embeddings:

$$\mathbf{s} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{w}_i$$

4.2 TF-IDF Weighted Mean

Tokens were weighted by inverse document frequency (IDF) to reduce the influence of highly frequent tokens:

$$\mathbf{s} = \frac{\sum_{i=1}^{n} \mathrm{IDF}(w_i)\mathbf{w}_i}{\sum_{i=1}^{n} \mathrm{IDF}(w_i)}$$

5 Sentiment Classification

A Logistic Regression classifier with `class_weight="balanced"` was trained on the Financial PhraseBank dataset using a stratified 80/20 train-test split.

6 Confusion Matrix Analysis

Confusion matrices reveal structural differences between the models.

6.1 VADER

VADER shows a strong bias toward positive predictions and struggles with contextual polarity shifts.

6.2 CBOW + Mean

The embedding-based model reduces negative-to-positive misclassifications and better captures contextual semantics.

6.3 CBOW + TF-IDF

TF-IDF weighting improves minority class recall and reduces neutral-positive confusion.

7 Error Analysis

Several cases highlight where embeddings outperform VADER:

7.1 Contextual Polarity Shift

Example: "Profits declined sharply."

VADER predicts positive due to the presence of "profits," while the embedding model correctly captures the negative sentiment introduced by "declined."

7.2 Expectation-Based Sentiment

Example: "Revenue increased but fell short of expectations."

The embedding model accounts for contextual modifiers, whereas VADER focuses on isolated positive tokens.

8 Conclusion

This project demonstrates that:

- BPE provides an efficient and robust subword tokenization strategy.
- CBOW embeddings trained with negative sampling capture meaningful semantic relationships.
- Embedding-based sentence representations significantly outperform rule-based sentiment analysis on financial text.
- TF-IDF weighted pooling further improves classification performance.

The results validate the effectiveness of learned distributional representations for domain-specific sentiment analysis.