Comparative Analysis of Text Representations on BBC News Classification and Retrieval

Name: Aryan Pawar
Student ID: SE22UARI195
Course: Computational Sequence Modeling

Abstract

This report presents a systematic comparison of text representation methods for news article classification and information retrieval. We evaluate seven representation techniques across two families: sparse methods (One-Hot Encoding, Bag-of-Words, N-grams, TF-IDF) and dense embeddings (Word2Vec variants, GloVe). Using the BBC News dataset with deterministic train-dev-test splits, we trained logistic regression classifiers and evaluated retrieval performance via cosine similarity ranking. Results indicate that sparse representations achieve superior classification performance (Bag-of-Words: 0.9683 Macro-F1), while dense embeddings demonstrate stronger retrieval capabilities (GloVe: 0.867 MAP@5). These findings highlight the importance of matching representation choice to the downstream task.

1. Introduction

Text representation is fundamental to natural language processing applications. The challenge lies in converting unstructured text into numerical features that preserve semantic and syntactic information while remaining computationally tractable. Traditional sparse representations rely on explicit word counting and weighting, while modern dense embeddings learn distributed representations that capture semantic relationships.

This study addresses two questions: (1) How do sparse and dense representations compare for supervised text classification? (2) Which representation family better supports semantic retrieval tasks? We evaluate these questions using the BBC News dataset, which contains 2,225 news articles across five balanced categories.

Our experimental design uses a deterministic splitting protocol based on the student roll number to ensure reproducibility. We report comprehensive metrics across both representation families, including training efficiency, memory footprint, classification accuracy, and retrieval effectiveness.

2. Methodology

2.1 Dataset and Preprocessing

The BBC News dataset comprises 2,225 articles distributed across business, entertainment, politics, sport, and technology categories. We applied deterministic stratified splitting using

CRC32 hash of the roll number, yielding:

- TRAIN: 1,335 documents (60%)
- DEV: 445 documents (20%)
- TEST: 445 documents (20%)

All text underwent standardized preprocessing: lowercasing, tokenization via NLTK, stopword removal using the English stoplist, and lemmatization using WordNet. Vocabulary construction was restricted to the training set to prevent test set leakage.

2.2 Sparse Representations

One-Hot Encoding (OHE): Binary features for the 2,000 most frequent training tokens.

Bag-of-Words (BoW): Unigram term frequencies with minimum document frequency threshold of 2.

N-grams: Combined unigram and bigram features with minimum document frequency of 3 to reduce dimensionality.

TF-IDF: Term frequency weighted by smoothed inverse document frequency: idf(t) = log((N+1)/(df(t)+1)) + 1. We verified manual TF-IDF calculations matched sklearn output within numerical precision (1e-6).

2.3 Dense Representations

Word2Vec: We trained four Word2Vec variants using the gensim library: Skip-gram and CBOW architectures, each with both Negative Sampling (negative=5) and Hierarchical Softmax. All models used 100-dimensional vectors, context window of 5, minimum word count of 3, and trained for 10 epochs.

GloVe: We used pretrained 100-dimensional GloVe embeddings (glove.6B.100d) trained on 6 billion tokens.

For document-level representations, we pooled word embeddings via TF-IDF-weighted averaging over in-vocabulary tokens, discarding out-of-vocabulary words.

2.4 Evaluation Tasks

Classification: Logistic regression with L2 regularization, hyperparameter C selected from {0.01, 0.1, 1.0, 10.0, 100.0} based on DEV set Macro-F1. Final evaluation used TEST set.

Retrieval: Generated 20 deterministic queries (15 TF-IDF-based queries from evenly-spaced training documents, 5 negation queries from shortest documents). Ranked TEST documents via cosine similarity and computed Mean Average Precision at 5 (MAP@5), Recall at 10, and negation top-1 accuracy.

3. Results

3.1 Representation Efficiency

Table 1: Sparse Representation Statistics

Method	Vocab Size	Sparsity	OOV Rate (TEST)	Fit Time (s)	Memory (MB)
OHE	2,000	0.9498	0.2863	0.224	1.54
BoW	11,515	0.9879	0.0706	0.221	2.14
N-gram	18,625	0.9907	0.0935	0.724	2.65
TF-IDF	11,515	0.9879	0.0706	0.200	2.14

Sparse methods exhibit high sparsity (>94%) as expected. One-Hot Encoding shows the highest out-of-vocabulary rate (28.6%) due to its restricted 2,000-token vocabulary. N-grams require the most training time (0.724s) due to combinatorial feature generation.

Table 2: Dense Representation Training Efficiency

Method	Vocab Size	Training Time (s)	Tokens/sec
W2V-SG-NS	9,848	6.25	45,760
W2V-CBOW-NS	9,848	2.01	142,017
W2V-SG-HS	9,848	6.38	44,775
W2V-CBOW-HS	9,848	1.98	144,672
GloVe (pretrained)	400,000	-	-

CBOW architectures train substantially faster than Skip-gram (approximately 3x speedup). The choice between Negative Sampling and Hierarchical Softmax has minimal impact on training speed. GloVe provides immediate availability through pretraining on much larger corpora.

3.2 Classification Performance

Table 3: TEST Set Classification Results

Representation	Macro-F1	Accuracy	Best C
BoW	0.9683	0.9685	1.0
OHE	0.9654	0.9663	10.0
N-gram	0.9639	0.9640	0.1
TF-IDF	0.9639	0.9640	1.0
W2V-HS + TF-IDF	0.9327	0.9348	0.1
W2V-NS + TF-IDF	0.9306	0.9326	0.1
GloVe + TF-IDF	0.9267	0.9281	10.0

Bag-of-Words achieved the strongest classification performance at 96.83% Macro-F1. Sparse methods consistently outperformed dense embeddings by 3-4 percentage points. This suggests that explicit word frequency patterns provide stronger discriminative signals for news topic classification than semantic similarity captured by embeddings.

3.3 Retrieval Performance

Table 4: TEST Set Retrieval Results

Representation	MAP@5	Recall@10	Negation Top-1
GloVe + TF-IDF	0.867	0.094	1.00
W2V-HS + TF-IDF	0.840	0.092	1.00
W2V-NS + TF-IDF	0.832	0.091	1.00
TF-IDF	0.699	0.071	1.00

Dense embeddings substantially outperformed sparse TF-IDF for retrieval. GloVe achieved the highest MAP@5 (0.867), representing a 24% relative improvement over sparse TF-IDF. All methods achieved perfect negation detection (100%), indicating successful semantic matching for contrastive queries.

4. Discussion

4.1 Sparse vs Dense Trade-offs

The divergent performance across tasks reveals fundamental differences in representation capabilities. Sparse methods excel at classification because news categories exhibit distinctive keyword patterns (e.g., "film" for entertainment, "game" for sport). These explicit lexical features provide strong discriminative power for supervised learning.

Conversely, dense embeddings capture semantic relationships that prove valuable for retrieval. When matching queries to documents, semantic similarity (e.g., matching "victory" with "win") outperforms exact keyword overlap. GloVe's superior retrieval performance likely stems from its global co-occurrence matrix factorization, which captures broader semantic patterns than Word2Vec's local context windows.

4.2 Word2Vec Architecture Comparison

Among Word2Vec variants, Skip-gram with Hierarchical Softmax achieved slightly better retrieval performance (MAP@5: 0.840) than Skip-gram with Negative Sampling (0.832), though differences were marginal. CBOW variants showed similar patterns (results not shown in tables). The primary distinction lies in training efficiency: CBOW trains approximately 3x faster than Skip-gram, making it preferable when computational resources are limited.

4.3 Effect of TF-IDF Weighting

TF-IDF weighting proved essential for pooling word embeddings into document vectors. This approach assigns higher weight to discriminative terms while downweighting common words, creating more informative document representations than simple averaging. The consistent strong performance of TF-IDF-weighted dense methods validates this design choice.

5. Conclusion

This study demonstrates that representation choice should align with downstream task requirements. For supervised classification of news articles, traditional Bag-of-Words representations achieve excellent performance (96.83% Macro-F1) with minimal computational overhead. For semantic retrieval applications, dense embeddings—particularly pretrained GloVe—provide superior matching capability (86.7% MAP@5).

Practitioners should consider these trade-offs when designing NLP systems. Classification tasks with well-defined categories benefit from sparse representations' explicit feature encoding. Retrieval and similarity tasks requiring semantic understanding justify the computational cost of dense embeddings. Hybrid approaches combining both representation families warrant investigation for applications requiring both capabilities.