Linear Text Classification: A Comparative Analysis of Semantic and Spelling-Based Tasks

Nusaibah Binte Rawnak (261090557)
COMP 550, Winter 2026

1 Problem Setup

This study investigates the comparative performance of linear classifiers on two distinct binary text classification tasks: one based on semantics and one based on spelling patterns.

1.1 Task A: Sentiment Classification

We formulated a binary sentiment classification problem where Class 0 consists of sentences expressing clearly negative sentiment, including dissatisfaction, complaints, or criticism (e.g., "The product broke after just two days of normal use"). Class 1 consists of sentences expressing clearly positive sentiment, including praise, satisfaction, or enjoyment (e.g., "The concert last night was absolutely spectacular and exceeded all my expectations"). Sentences with mixed or neutral sentiment were excluded to ensure unambiguous labeling.

1.2 Task B: Alliteration Classification

We formulated a binary classification problem based on phonological patterns. Class 1 consists of sentences containing at least three consecutive words that begin with the same letter (e.g., "Seven silly swans swam silently across the lake"). Class 0 consists of sentences with no such alliteration patterns (e.g., "The library was quiet during the afternoon hours").

1.3 Hypotheses

We hypothesized that linear classifiers would perform well ($\geq$75% accuracy) on Task A because sentiment polarity correlates strongly with emotionally charged words (e.g., "excellent" vs. "terrible"), which are directly captured by bag-of-words features. We further hypothesized that linear classifiers would perform very well ($\geq$85% accuracy) on Task B because alliteration creates explicit, consistent surface-form patterns that are more mechanically detectable than semantic distinctions.

2 Dataset Generation and Experimental Procedure

2.1 Dataset Creation

We generated 160 sentences per task (80 per class) using a combination of manual writing and AI-assisted generation with Claude AI. All generated sentences were manually inspected to ensure they satisfied the labeling criteria. Sentences were stored in plain text files (UTF-8 encoding) with one sentence per line: `synsem0.txt` and `synsem1.txt` for Task A; `morphphon0.txt` and `morphphon1.txt` for Task B.

2.2 Evaluation Strategy

For each task, we used a 70/30 stratified train-test split to preserve class balance while providing a reasonably sized test set (48 samples). The test set was held out entirely and used only for final performance evaluation. All preprocessing comparisons and model selection were performed using 5-fold cross-validation on the training set only, ensuring no test-set leakage.

3 Parameter Settings Explored

We explored three preprocessing decisions: (1) **Lowercasing** (with vs. without), which reduces vocabulary size and sparsity; (2) **Stopword removal** (with vs. without English stopwords), particularly relevant for sentiment where function words may carry limited signal; and (3) **Feature weighting** (raw term counts via `CountVectorizer` vs. TF-IDF via `TfidfVectorizer`), which downweights common words.

We tested two linear classifiers: **Logistic Regression**, a probabilistic linear baseline widely used for text classification, and **Linear SVM**, a margin-based classifier effective in high-dimensional sparse spaces. For Task B, we additionally tested **bigram features** (unigrams + bigrams) to capture consecutive word patterns critical for alliteration detection. All models were implemented using scikit-learn with default hyperparameters.

4 Results and Conclusions

Table 1 summarizes the best performing configurations for each task.

Task	Best Config	Test Acc.
Task A	Baseline + LR/SVM	70.83%
Task B	Baseline + LR/SVM	93.75%

Table 1: Best test accuracy for each task.

4.1 Task A: Sentiment Classification

Task A achieved a best test accuracy of 70.83% using Logistic Regression or Linear SVM with baseline preprocessing (no lowercasing, no stopword removal, raw counts). Cross-validation scores ranged from 54-62% across configurations, with stopword removal yielding higher CV scores but lower test performance, suggesting potential overfitting on the small training set. While below the hypothesized 75% threshold, the results substantially exceed the 50% random baseline, demonstrating that linear classifiers can learn sentiment patterns from limited data.

4.2 Task B: Alliteration Classification

Task B achieved a best test accuracy of 93.75% using either Logistic Regression or Linear SVM with baseline preprocessing and unigram features. Cross-validation scores were consistently high (82-89%), indicating stable performance. Surprisingly, adding bigram features did not improve performance, likely because unigrams already capture the repeated initial letters. Lowercasing and stopword removal generally degraded performance, as capitalization and function words help preserve alliteration patterns.

4.3 Comparison and Hypothesis Validation

The results confirm our hypothesis that spelling-based patterns are more explicitly detectable by linear models than semantic distinctions. Task B's 93.75% accuracy validates that alliteration creates strong, mechanically separable features. Task A's more modest 70.83% reflects the inherent complexity of semantic understanding, where sentiment requires interpreting word meaning rather than surface patterns. The 23-percentage-point gap demonstrates that linear classifiers excel at explicit spelling tasks but face greater challenges with semantics.

5 Limitations

This study has several limitations. First, the small dataset size (160 sentences per task) limits statistical power and may introduce high variance in test performance. Second, the experiments are specific to English and manually curated data; results may not generalize to other languages or naturally occurring text. Third, we used default hyperparameters without tuning, which may not reflect optimal performance. Future work could explore larger datasets, character n-gram features for phonological tasks, and more sophisticated preprocessing pipelines. Despite these limitations, the experiments successfully demonstrate the differential performance of linear classifiers on semantic versus spelling-based classification.

References

[1] 42Signals. Positive vs negative sentiment analysis. https://www.42signals.com/blog/positive-vs-negative-sentiment-analysis/, 2024. Accessed: January 2026.

[2] MAXQDA. Sentiment analysis: A complete guide. https://www.maxqda.com/research-guides/sentiment-analysis, 2024. Accessed: January 2026.

[3] SHAP Documentation. Positive vs. negative sentiment classification. https://shap.readthedocs.io/en/latest/example_notebooks/text_examples/sentiment_analysis/, 2024. Accessed: January 2026.

[4] Wikipedia. Morphology (linguistics). https://en.wikipedia.org/wiki/Morphology_(linguistics), 2024. Accessed: January 2026.