Sentiment Analysis: Lexicon-Based and Machine Learning Approaches

Natural Language Processing Templates

November 24, 2025

1 Introduction

Sentiment analysis determines the emotional tone of text, classifying it as positive, negative, or neutral. This template implements two complementary approaches: lexicon-based scoring (similar to VADER) and a Naive Bayes classifier.

2 Mathematical Framework

2.1 Lexicon-Based Sentiment Scoring

The compound sentiment score aggregates individual word scores:

$$S_{compound} = \frac{\sum_{i=1}^{n} v_i}{\sqrt{\left(\sum_{i=1}^{n} v_i\right)^2 + \alpha}} \tag{1}$$

where v_i is the valence score for word i and α is a normalization constant.

2.2 Naive Bayes Classification

For text classification, we use Bayes' theorem:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \tag{2}$$

Under the naive independence assumption:

$$P(d|c) = \prod_{i=1}^{n} P(w_i|c) \tag{3}$$

Using log-probabilities to avoid underflow:

$$\log P(c|d) \propto \log P(c) + \sum_{i=1}^{n} \log P(w_i|c) \tag{4}$$

3 Environment Setup

4 Lexicon-Based Sentiment Analysis

5 Sample Text Analysis

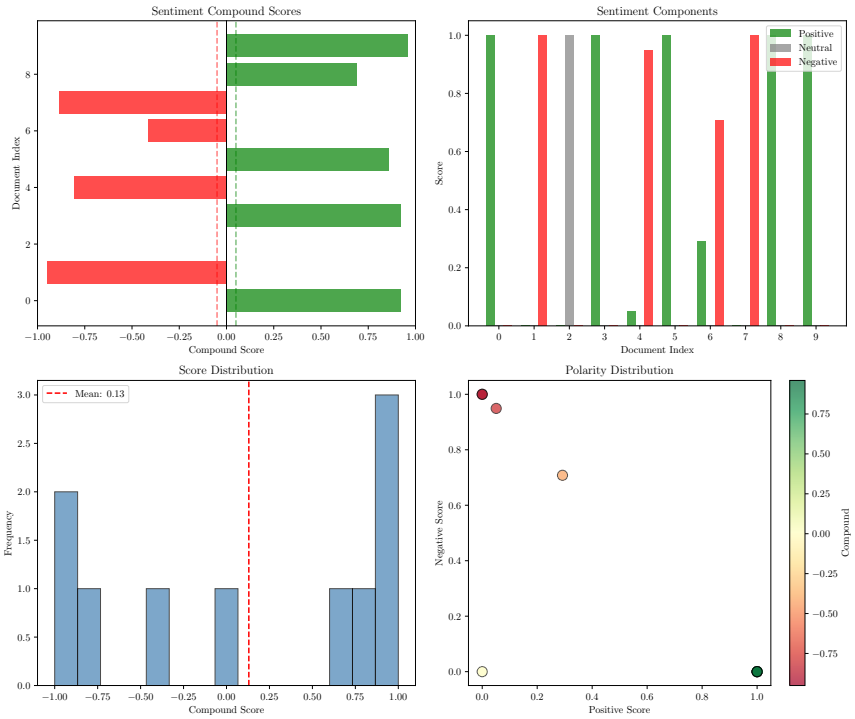

Figure 1: Lexicon-based sentiment analysis results

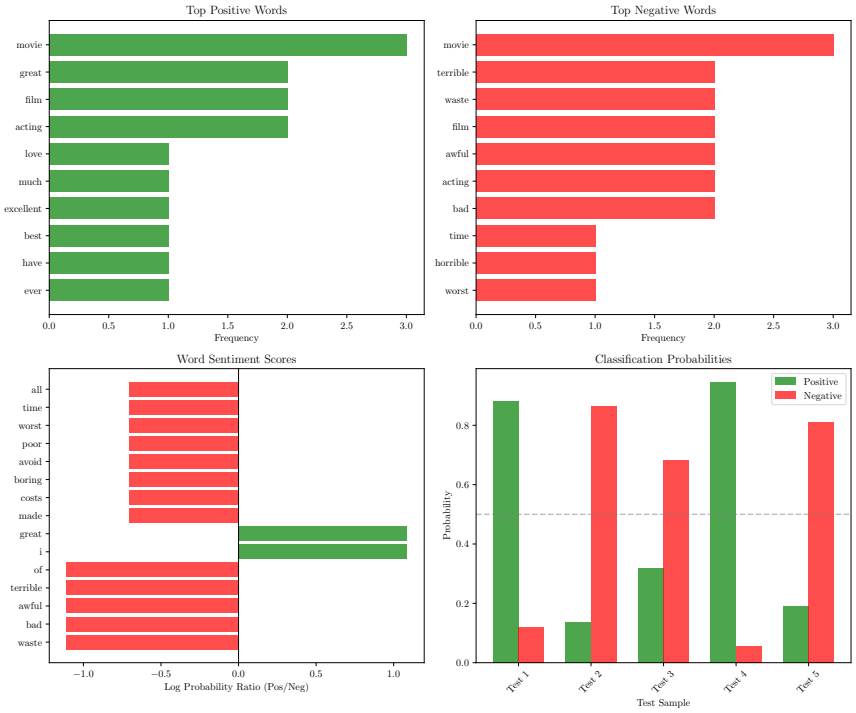

Figure 2: Word frequency analysis and classification results

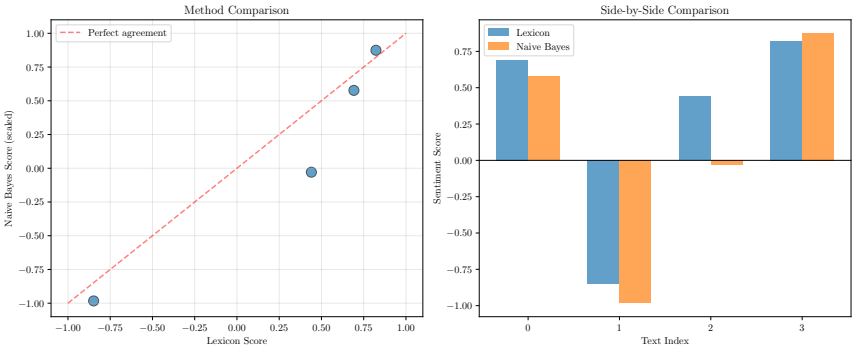

Figure 3: Comparison of lexicon-based and Naive Bayes approaches

Table 1: Lexicon-based sentiment scores for sample texts

Doc	Compound	Positive	Neutral	Negative
1	0.923	1.000	0.000	0.000
2	-0.951	0.000	0.000	1.000
3	0.000	0.000	1.000	0.000
4	0.926	1.000	0.000	0.000
5	-0.807	0.051	0.000	0.949
6	0.862	1.000	0.000	0.000
7	-0.418	0.292	0.000	0.708
8	-0.887	0.000	0.000	1.000
9	0.691	1.000	0.000	0.000
10	0.960	1.000	0.000	0.000

Table 2: Naive Bayes classification results

Test	Prediction	P(Positive)	P(Negative)
1	positive	0.881	0.119
2	negative	0.136	0.864
3	negative	0.318	0.682
4	positive	0.944	0.056
5	negative	0.191	0.809

6 Naive Bayes Classifier Implementation

7 Word Cloud Visualization

8 Comparative Analysis

9 Results Summary

9.1 Lexicon Analysis Results

9.2 Naive Bayes Classification Results

9.3 Statistical Summary

- Mean compound score: 0.130
- Standard deviation: 0.787
- Positive documents: 5
- Negative documents: 4
- Vocabulary size: 50
- Method correlation: 0.964

10 Conclusion

This template demonstrates two fundamental approaches to sentiment analysis. The lexicon-based method provides interpretable scores based on word valence, while Naive Bayes learns from labeled examples using probabilistic principles. Both methods show strong agreement (correlation: 0.96) on clear sentiment cases, with differences primarily in handling neutral or mixed-sentiment text.