A Comparative Analysis of Statistical and Rule-Based Sentiment Analysis for In-Domain and Out-of-Domain Text

Ramiz Aliguliyev
COM6115: Text Processing
The University of Sheffield

November 2025

Contents

1 Introduction

The main objective of this assignment is to compare a **domain-agnostic model** (designed for general use) with a **domain-specific model** (trained for high performance in a single narrow field) in the task of recognizing the sentiment tone of text.

This report compares two primary classifiers:

- A statistical (probabilistic) classifier: **Naïve Bayes**, which learns features from input data. Naïve Bayes classifiers assume that features are conditionally independent, simplifying the probabilistic calculation.
- A lexicon-based (rule-based) classifier: This method uses a pre-built dictionary (lexicon) of words and phrases associated with positive, negative, or neutral sentiment to score the text.

Sentiment analysis is straightforward for humans but challenging for computers, particularly when dealing with domain shifts or linguistic subtleties like **sarcasm** or negation (where negative words convey a positive meaning, and vice versa). This report implements and evaluates both models on in-domain (film reviews) and out-of-domain (Nokia product reviews) data to analyze their performance, investigate their learned features, and propose improvements to the rule-based system.

2 Methodology: Datasets and Lexicon Parsing

2.1 Datasets (Step 1)

The evaluation utilized the following datasets:

- **In-Domain (Film Reviews):** Fragments of movie feedback from Rotten Tomatoes (`rt-polarity.pos` and `rt-polarity.neg`).
- **Out-of-Domain (Product Reviews):** Fragments of feedback for Nokia phones (`nokia-pos.txt` and `nokia-neg.txt`).
- **Sentiment Dictionary (Lexicon):** `negative-words.txt` (4783 words) and `positive-words.txt` (2006 words).

2.2 Lexicon Processing

To correctly populate the `posWordList` and `negWordList`, the lexicon files were processed using list comprehension:

```
[line.strip() for line in posDictionary.readlines()
 if line.strip() and not line.startswith(';')]
```

This code performs three crucial cleaning operations: it uses `line.strip()` to remove whitespace, `if line.strip()` to filter out blank lines, and `not line.startswith(';')` to ignore all comment lines, ensuring only valid sentiment words are included.

3 Evaluation Metrics

The models were evaluated using the following metrics:

- **Accuracy:** The total percentage of correct predictions ($\frac{TP+TN}{\text{Total}}$).

- **Precision (for a class):** The proportion of true positives among all instances the model predicted as positive ($\frac{TP}{TP+FP}$). This measures the quality of the model's positive prediction.
- **Recall (for a class):** The proportion of true positives that the model successfully identified ($\frac{TP}{TP+FN}$). This measures the model's ability to find all positive instances.
- **F1-Score:** The harmonic mean of Precision and Recall ($\frac{2\times(\text{Precision}\times\text{Recall})}{\text{Precision}+\text{Recall}}$), providing a balanced measure that penalizes models favoring one metric heavily.

4 Naïve Bayes Results and Analysis

4.1 Performance on Rotten Tomatoes (In-Domain)

The Naïve Bayes classifier showed strong performance on the in-domain film review data.

Table 1: **Table 4.1: Naïve Bayes Classification for Films (In-Domain)**

Data Set	Metric	Positive Class	Negative Class
4*Train Data	Accuracy	88.83%	88.83%
	Precision	89.29%	88.39%
	Recall	88.15%	89.50%
	F1-score	88.71%	88.94%
4*Test Data	Accuracy	77.31%	77.31%
	Precision	78.71%	75.88%
	Recall	76.95%	77.69%
	F1-score	77.82%	76.77%

4.2 Performance on Nokia (Out-of-Domain)

The model failed to generalize to the out-of-domain data, demonstrating poor performance.

Table 2: **Table 4.2: Naïve Bayes Classification for Nokia (Out-of-Domain)**

Data Set	Metric	Positive Class	Negative Class
4*All Data	Accuracy	57.89%	57.89%
	Precision	77.21%	37.69%
	Recall	56.45%	61.25%
	F1-score	65.22%	46.67%

The accuracy of 57.89% is only slightly better than a 50/50 random guess. This severe drop is a clear example of **domain mismatch**, where the film-specific vocabulary learned by the model is ineffective for analyzing product reviews.

4.3 Feature Analysis (Step 4)

The function used to determine the **most predictive words** for the positive class yielded the following:

['stylistic', 'delicious', 'smartly', 'gradually', 'melancholy', 'marvel', 'resist', 'ramsay', 'portrayal']

The learned vocabulary contains words highly specific to film critique ("marvel," "portrayal," "stylistic"), confirming why the model's expertise is entirely useless on the Nokia dataset.

5 Rule-Based System Comparison and Improvement

5.1 Baseline Rule-Based Performance

The lexicon-based model's performance on both domains is summarized below.

Table 3: **Table 5.1: Baseline Dictionary-based Classification**

Data Set	Metric	Positive Class	Negative Class
4*Films (Test Data)	Accuracy	64.31%	64.31%
	Precision	65.58%	63.47%
	Recall	54.44%	73.48%
	F1-score	59.49%	68.11%
4*Nokia (All Data)	Accuracy	79.70%	64.31%
	Precision	88.37%	63.83%
	Recall	81.72%	75.00%
	F1-score	84.92%	68.97%

Model Comparison (Baseline): The Naïve Bayes model acts as a **specialist** (high in-domain accuracy, low out-of-domain), whereas the dictionary-based model acts as a **generalist**. While less accurate on films, its performance on the out-of-domain Nokia data (79.70% accuracy) is significantly higher than Naïve Bayes, demonstrating superior portability because its lexicon is universal.

5.2 Improved Model Implementation (Negation Only)

The chosen improvement was adding a **negation window** (a window of 3 words following a negative word) to invert the sentiment score. This was implemented to correct misclassifications caused by phrases like "not good."

Table 4: **Table 5.2: IMPROVED Dictionary-based Classification (Negation Only)**

Data Set	Metric	Positive Class	Negative Class
4*Films (Test Data)	Accuracy	64.41%	64.41%
	Precision	65.73%	63.52%
	Recall	54.44%	73.66%
	F1-score	59.56%	68.22%
4*Nokia (All Data)	Accuracy	82.71%	82.71%
	Precision	92.17%	67.00%
	Recall	82.26%	83.75%
	F1-score	86.93%	74.44%

The negation-only model slightly improved accuracy on films and provided a substantial boost to the Nokia dataset ($\approx +3.01\%$), confirming the utility of basic linguistic rules across domains.

5.3 Intensifiers and Diminishers Experiment (Discarded)

An experimental model was also tested, incorporating intensifiers and diminishers (words like "very" or "slightly"). The results showed a mixed impact, indicating the risk of **over-engineering** in rule-based systems.

Table 5: **Table 5.3: Classification with Intensifiers and Diminishers (DISCARDED)**

Data Set	Metric	Positive Class	Negative Class
4*Films (Test Data)	Accuracy	63.48%	63.48%
	Precision	65.63%	61.88%
	Recall	56.17%	70.75%
	F1-score	60.53%	66.02%
4*Nokia (All Data)	Accuracy	80.45%	80.45%
	Precision	91.88%	63.21%
	Recall	79.03%	83.75%
	F1-score	84.97%	72.04%

The comparison showed that the complex model's accuracy degraded on the Nokia data compared to the simpler negation-only model. This suggests that the simpler 'Improved' model (negation only) was superior due to its better **generalizability**.

6 Error Analysis (Step 6)

6.1 Errors in Naïve Bayes

> **ERROR (Negative classed as Positive, 0.79):** while the production details are lavish , film has little insight into the historical period and its artists , particularly in how sand developed a notorious reputation .

The Naïve Bayes model is easily misled by mixed-sentiment text and struggles to understand sentence structure. The high positive probability is likely driven by strongly positive words like "lavish" and "developed."

6.2 Errors in Improved Dictionary Model

> **ERROR (Negative classed as Positive, score 3):** a film that plays things so nice 'n safe as to often play like a milquetoast movie of the week blown up for the big screen .

The Dictionary Model is strictly limited by its lexicon and fails to detect sarcasm, idioms, or uncommon critical words ("milquetoast"). Both examples highlight that neither model effectively captures linguistic subtleties, pointing to the need for advanced, context-perceptive models.

7 Conclusion

The comparison between the Naïve Bayes and the rule-based lexicon classifier demonstrated a clear trade-off between specialization and portability. The Naïve Bayes model proved to be a **specialist**, achieving high in-domain accuracy (77.31%) but failing dramatically out-of-domain (57.89%). The Baseline Rule-Based model, however, performed as a robust **generalist**, maintaining stable performance across domains (up to 79.70% on Nokia). By implementing a simple negation rule (Improved Model v1), the rule-based system's performance was

successfully optimized on both datasets ($\approx +0.53\%$ on film test data and $\approx +3.01\%$ on Nokia). Ultimately, while statistical classifiers excel in fixed, singular domains, the rule-based model, particularly when tuned for universal linguistic rules like negation, is superior for general-purpose, portable sentiment analysis.

Appendix

Learned Negative Words

['stupid', 'badly', 'unfunny', 'generic', 'dull', 'mediocre', 'routine', 'poorly', 'stale', 'shoot', 'bore', "wasn't", 'annoying', 'pointless', 'disguise', 'meandering', 'save', 'tiresome', 'boring', 'nowhere', 'disaster', 'cliché', 'offensive', 'inept', 'mindless', 'banal', 'mixed', 'plodding', 'chan', 'pinocchio', 'junk', 'apparently', 'horrible', 'trite', 'product', 'incoherent', 'seagal', 'lousy', 'kung', 'lifeless', 'conceived', 'unless', 'flat', 'ill', 'supposed', 'waste', 'animal', 'amateurish', 'harvard', 'wasted', 'ask', 'fatal', 'sadly', 'pile', 'ballistic', 'crap', 'leaden', 'ingredients', 'unintentionally', 'hollow', 'bother', 'lame', 'obnoxious', 'pathetic', 'generate', 'missed', 'unintentional', 'intentioned', 'comparison', 'produce', 'stiff', 'clumsily', 'pow', 'uninspired', 'stunt', 'numbingly', 'sara', 'inane', 'unnecessary', 'soggy', 'busy', 'halfway', 'stealing', 'store', 'serving', 'overlong', 'imitation', 'purpose', 'cable', 'guess', 'settles', 'superficial', 'devoid', 'witless', 'scattered', 'endless', 'writers', 'looked', 'putting', 'ludicrous']

Learned Positive Words

['smartly', 'melancholy', 'extraordinary', 'simplicity', 'spite', 'superbly', 'leigh', 'finely', 'para', 'physical', 'portrayal', 'harrowing', 'color', 'desperation', 'enjoyable', 'portrait', 'ages', 'evocative', 'helps', 'buoyant', 'lovers', 'pianist', 'exquisitely', 'nuanced', 'sobering', 'gradually', 'miller', 'delightfully', 'richly', 'hopeful', 'breathtaking', 'intoxicating', 'joyous', 'unique', 'absorbing', 'intimate', 'colorful', 'thoughtful', 'touching', 'masterful', 'undeniably', 'startling', 'powerful', 'resonant', 'wrenching', 'russian', 'spare', 'timely', 'poem', 'unflinching', 'transcends', 'frailty', 'twisted', 'tour', 'ingenious', 'deft', 'explores', 'subversive', 'answers', 'sadness', 'martha', 'sides', 'heartbreaking', 'jealousy', 'aspects', 'flawed', 'record', 'smarter', 'warm', 'unexpected', 'iranian', 'detailed', 'captivating', 'grown', 'polished', 'respect', 'wry', 'heartwarming', 'challenging', 'dazzling', 'bittersweet', 'beauty', 'playful', 'lively', 'mesmerizing', 'vividly', 'captures', 'tender', 'gem', 'wonderfully', 'chilling', 'refreshingly', 'realistic', 'haunting', 'riveting', 'refreshing', 'inventive', 'provides', 'engrossing', 'wonderful']