International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 02 | Feb 2026 www.irjet.net p-ISSN: 2395-0072

Comparative Analysis of Naive Bayes and Logistic Regression for Spam Detection: Impact of Laplace Smoothing on Classification Performance

Vikrant Singh

[1]*Department of Artificial Intelligence and Data Science Pune Institute of Computer Technology Pune, Maharashtra, India*

Abstract - *This study presents a comparative analysis of Multinomial Naive Bayes and Logistic Regression for spam classification using the SMS Spam Collection Dataset. The primary objective is to evaluate the effect of Laplace smoothing on Naïve Bayes performance and analyse the impact of the independence assumption. Experimental results show that optimal smoothing (α = 0.1) significantly improves classification performance. Naïve Bayes achieved 98.39% accuracy and 92% recall, outperforming Logistic Regression, which achieved 95.52% accuracy and 68.67% recall. The results demonstrate that proper smoothing enhances generative classifier performance and that Naive Bayes is particularly effective for high-dimensional sparse text classification tasks.*

Key Words: **Spam Detection, Naive Bayes, Logistic Regression, Laplace Smoothing, Machine Learning, Text Classification**

1. INTRODUCTION

Spam detection is a fundamental problem in text classification and machine learning. Among various classification algorithms, Multinomial Naive Bayes and Logistic Regression are widely used due to their computational efficiency and effectiveness. Naive Bayes assumes conditional independence between features, which is often violated in real-world text data. Logistic Regression, on the other hand, is a discriminative model that does not rely on independence assumptions. This research aims to experimentally evaluate the sensitivity of the Multinomial Naive Bayes classifier to Laplace smoothing and to compare its performance with Logistic Regression in spam detection tasks. Specifically, the study investigates how varying the smoothing parameter influences classification accuracy, precision, recall, and F1 score. The objective is to analyze the practical implications of the independence assumption in Naive Bayes and to determine optimal smoothing conditions for improved classification performance.

2. CONTRIBUTION

The main contributions of this study are as follows:
•Experimental sensitivity analysis of Laplace smoothing Parameter on Multinomial Naive Bayes performance.
•Comparative performance evaluation of Naive Bayes and Logistic Regression using TF-IDF features.
•Empirical demonstration that moderate smoothing significantly improves classification performance.
•Analysis showing that Naive Bayes achieves higher recall in spam detection, making it more suitable for real-world spam filtering applications.

3. RELATED WORK

Naive Bayes has been widely used in spam detection due to its simplicity, efficiency, and relatively strong performance in high-dimensional text classification problems. Logistic Regression has also demonstrated robust performance in classification tasks due to its discriminative nature and ability to model feature relationships directly. Previous research has shown that Laplace smoothing plays an important role in improving probability estimation in Naïve Bayes classifiers by preventing zero probabilities. However, excessive smoothing can negatively impact classification accuracy. While both Naive Bayes and Logistic Regression have been extensively studied, experimental analysis focusing on smoothing sensitivity and independence assumption effects remains limited. Naive Bayes is a widely used probabilistic classifier in text classification tasks [1]. Logistic Regression is a powerful discriminative model for classification problems [2], [3]. The experiments in this study use the SMS Spam Collection Dataset [4].

4. METHODOLOGY

A. Naive Bayes Classification

Naive Bayes is a probabilistic classifier based on Bayes' theorem:

$$P(Y \mid X) = (P(X \mid Y) \times P(Y)) / P(X) \quad (1)$$

where Y represents the class label and X represents the feature vector.

Laplace smoothing modifies probability estimation as follows:

$$P(xi|Y) = (count(xi, Y) + \alpha)/(count(Y) + \alpha n) \quad (2)$$

where α is the smoothing parameter and n is the number of features.

B. Logistic Regression

Logistic Regression is a discriminative classifier that models the probability of a class label using the sigmoid function:

$$P(y = 1|x) = 1/(1 + e^{-(wT x+b)}) \quad (3)$$

where:
- x is the feature vector representing the input document,
- w is the weight vector that represents the importance of each feature,
- b is the bias term,
- w^T x represents the dot product between the weight vector and feature vector.

The weight vector w and bias term b are learned during training to minimize classification error. Unlike Naive Bayes, Logistic Regression does not assume independence between features and directly learns decision boundaries from the data.

5. EXPERIMENTAL SETUP

The dataset was divided into training and testing sets using an 80:20 split. Text messages were transformed using TF-IDF Vectorization, which assigns weights to words based on their importance across documents. This helps reduce the influence of commonly occurring words and improves classification performance.

The Multinomial Naive Bayes classifier was trained using multiple smoothing parameter values ranging from 0.001 to 100. Logistic Regression was trained using default parameters.

All experiments were conducted using Python 3.11 and scikit-learn on a standard computing environment.

5.1 Dataset

The spam dataset consists of labelled text messages categorized as spam or non-spam. The text data was pre-processed using standard techniques including tokenization and vectorization.

The experiments were conducted using the SMS Spam Collection Dataset, which contains 5,572 labeled SMS messages categorized as spam or ham (non-spam). The dataset consists of 747 spam messages and 4,825 ham messages and is widely used as a benchmark for spam classification research.

5.2 Feature Extraction

Text messages were transformed using TF-IDF Vectorization. TF-IDF assigns weights to words based on their frequency within a document and their inverse frequency across all documents. This representation improves classification performance by emphasizing informative words and reducing the impact of common words.

5.3 Models Used

- Multinomial Naive Bayes
- Logistic Regression.

5.4 Laplace Smoothing Analysis

The smoothing parameter α was varied across multiple values ranging from 0.001 to 100 to analyze its impact on classification performance.

5.5 Evaluation Metrics

The following evaluation metrics were used:
- Accuracy
- Precision
- Recall
- F1 Score

6. IMPLEMENTATION DETAILS

The models were implemented using the scikit-learn machine learning library in Python. TF-IDF vectorization was performed using TfidfVectorizer. Multinomial Naive Bayes was implemented using MultinomialNB, and Logistic Regression was implemented using LogisticRegression.

Model performance was evaluated on the test dataset using accuracy, precision, recall, and F1 score metrics.

7. RESULTS AND DISCUSSION

Table -1: PERFORMANCE COMPARISON OF MODELS

Model	Accuracy	Precision	Recall	F1 Score
Naive Bayes (optimal α)	98.39%	95.83%	92%	93.88%
Logistic Regression	95.52%	97.17%	68.67%	80.47%

The optimal smoothing parameter was found to be α = 0.1, which produced the highest F1 score.

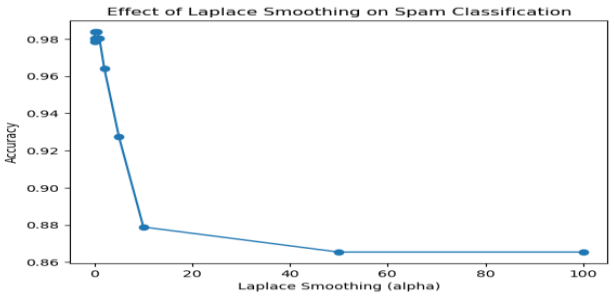

Chart -1: Effect of Laplace smoothing on Accuracy

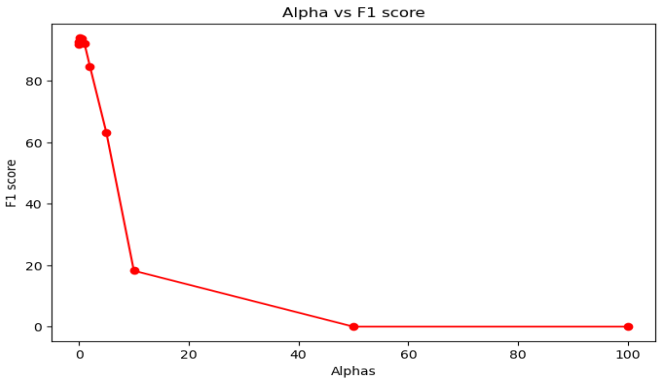

Chart -2: Effect of Laplace smoothing on F1 Score

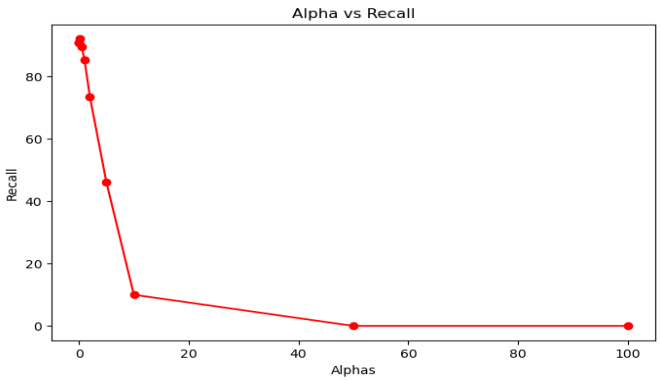

Chart -3: Effect of Laplace smoothing on Recall(Sensitivity)

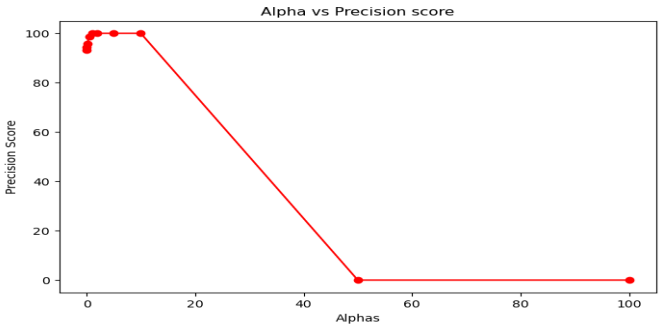

Chart -4: Effect of Laplace smoothing on Precision

The experimental results reveal several important observations regarding classifier behavior and smoothing sensitivity.

First, Multinomial Naive Bayes demonstrated strong performance when moderate smoothing values were used. Optimal smoothing improved probability estimation and prevented zero-probability issues without excessively distorting feature likelihoods. However, excessive smoothing resulted in performance degradation. Large smoothing values reduced the influence of discriminative words, causing the model to assign overly uniform probabilities across features.

Second, Logistic Regression exhibited lower recall compared to Naive Bayes. This indicates that Logistic Regression failed to identify a significant portion of spam messages. This behavior may be attributed to its linear decision boundary and sensitivity to class imbalance. Naive Bayes achieved higher recall, making it more effective in spam detection scenarios where identifying spam messages is more critical than avoiding false positives. The superior recall achieved by Naive Bayes may also be attributed to its probabilistic nature, which allows it to better capture rare word occurrences commonly associated with spam messages. Logistic Regression, as a discriminative model, may be more conservative in classification, resulting in lower recall. The dataset exhibits class imbalance, with spam messages representing only 13.41% of the total dataset. Class imbalance can significantly affect classifier performance, particularly recall. Generative classifiers such as Naive Bayes may handle such imbalance more effectively due to probabilistic modeling of feature distributions, which may explain the superior recall observed in this study.

These results suggest that generative models such asNaive Bayes can outperform discriminative models in highdimensional sparse text classification tasks, particularly when proper smoothing is applied. The findings confirm that Laplace smoothing plays a crucial role in improving Naive Bayes performance and that careful parameter tuning is essential for optimal classification.

8. LIMITATIONS

This study was conducted using a single spam dataset and default Logistic Regression parameters. Further research may include cross-validation, hyperparameter tuning for Logistic Regression, and evaluation on additional datasets to improve generalizability.

9. CONCLUSION

This study conducted a detailed experimental analysis of Laplace smoothing effects on Multinomial Naive Bayes and compared its performance with Logistic Regression for spam classification.

The results demonstrate that Laplace smoothing significantly influences Naive Bayes performance. Moderate smoothing values improve probability estimation and enhance classification accuracy, while excessive smoothing reduces model effectiveness.

Furthermore, Naive Bayes achieved higher recall compared to Logistic Regression, making it more suitable for spam detection tasks where identifying spam messages is critical. These findings provide empirical evidence that proper smoothing parameter selection is essential for optimal Naive

Bayes performance and confirm its effectiveness in highdimensional text classification problems.
Future work may extend this research to larger datasets, cross-domain classification tasks, and deep learning-based approaches.

ACKNOWLEDGMENT

The author would like to thank the academic mentors and peers who provided guidance during this research.

REFERENCES

[1] A. McCallum and K. Nigam,"A comparison of event models for naive bayes text classification," in AAAI Workshop on Learning for Text Categorization, 1998.

[2] C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.

[3] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. Springer, 2009.

[4] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, "Sms spam collection dataset," UCI Machine Learning Repository, 2011.