Towards a Comprehensive Comparative Evaluation of Classical Machine Learning Algorithms for SMS Spam Classification Using TF-IDF Representations

Durga J[1], Sunitha S [2]

[1,2] Asst.Professor, Department of Computer Applications, Mercy College, Palakkad, University of Calicut

Abstract

Detecting spam is a crucial task in today's communication systems, given the rising number of unwanted messages that can threaten security and privacy. This research offers a comparative evaluation of four machine-learning methods—Multinomial Naive Bayes, Logistic Regression, Support Vector Machine (SVM), and Random Forest—for classifying SMS spam. The SMS Spam Collection dataset underwent pre-processing and transformation through the Term Frequency–Inverse Document Frequency (TF-IDF) approach to turn text data into numerical feature vectors. Each model was assessed based on accuracy, precision, recall, F1-score, and training duration. The experimental findings show that linear models, notably SVM and Logistic Regression, perform better than other methods regarding accuracy and generalization ability. The research underscores the efficacy of traditional machine-learning algorithms in text classification tasks and offers insights for creating effective spam-filtering systems.

Keywords: Naïve Bayes, Support Vector Machine(SVM), Frequency-Inverse- Document Frequency(TF-IDF), F1-Score, Logistic Regression, Random Forest

1. Introduction

Unrequested and undesirable messages, often referred to as spam, pose a significant challenge within digital communication platforms. As mobile usage and SMS-based services continue to expand rapidly, spam messages not only create inconvenience but also put users at risk of phishing scams, financial deception, and harmful links. Conventional rule-based spam filters have proven inadequate to address the changing nature of spam content, which is frequently designed to evade simple keyword-based detection systems. Machine learning (ML) has surfaced as a successful approach for automating spam detection by identifying patterns from past data and predicting whether a newly received message is spam or legitimate. By examining linguistic characteristics, message layouts, and the significance of words, ML models can markedly enhance the precision and dependability of spam-filtering systems. In recent times, different text classification methods—such as Naive Bayes, Logistic Regression, Support Vector Machines, and ensemble techniques—have shown impressive results in recognizing spam content. These models, particularly when paired with sophisticated text pre-processing and feature extraction techniques like Term Frequency–Inverse Document Frequency (TF-IDF), can effectively capture significant semantic patterns from textual data. This study examines a comparative analysis of four popular machine-learning algorithms used for detecting spam in SMS messages. The goal is to assess and analyse their performance based on various metrics, including accuracy, precision, recall, F1-score, and training duration. The results are intended to determine the most effective and dependable model for practical spam-filtering tasks.

2. Literature Review

Spam detection has been a significant area of research for over twenty years, resulting in numerous methods being proposed in the literature. Initial investigations mainly centred on rule-based and keyword-matching strategies, where particular words or patterns were manually identified to flag possible spam. While these methods were straightforward to execute, they lacked flexibility and struggled to recognize new spam patterns that emerged.

As machine learning progressed, researchers started utilizing statistical and probabilistic models to enhance classification accuracy. Sahami et al. were pioneers in introducing a Bayesian filtering approach that showed impressive results in detecting email spam. This ground-breaking work paved the way for the creation of Multinomial Naive Bayes, which continues to be one of the most commonly employed models because of its straightforwardness and efficiency in managing text data.

Numerous studies have examined supervised learning methods such as Logistic Regression and Support Vector Machines (SVM). Research indicates that SVMs, particularly those with linear kernels, excel in handling high-dimensional text data and offer robust generalization capabilities. Logistic Regression has also been shown to be effective for binary classification problems, providing a good mix of speed, interpretability, and precision.

Beyond traditional techniques, ensemble methods like Random Forests and Gradient Boosting have been analysed for their effectiveness in spam detection. These approaches utilize multiple decision trees to identify intricate patterns and minimize over fitting. Evidence suggests that classifiers based on ensembles frequently surpass individual models, although they may demand additional computational resources.

Recent studies have additionally highlighted deep learning methodologies, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and architectures based on Transformers. Despite their ability to achieve high accuracy, these models require considerable training time and computational resources, which may render them impractical for lightweight or real-time spam filtering applications.

Among the publicly accessible datasets, the SMS Spam Collection dataset has become a popular choice for evaluating machine-learning algorithms. Numerous research efforts that have employed this dataset consistently indicate that traditional machine-learning methods, when combined with TF-IDF feature extraction, deliver performance that is often comparable to more sophisticated neural network models.

In summary, the current body of literature emphasizes that classical machine-learning models continue to be very effective for spam detection tasks, especially when utilized with well-organized datasets and enhanced by robust text pre-processing methods. This research expands on these insights by performing a comparative assessment of Naive Bayes, Logistic Regression, SVM, and Random Forest models for classifying SMS spam.

3. Methodology

3.1 Dataset Overview

The research employs the SMS Spam Collection Dataset, a publicly accessible collection that includes 5,574 English SMS messages classified into two categories: ham (valid messages) and spam. Each entry features a label alongside the associated text message. This dataset is frequently utilized in spam detection studies because it offers a balanced depiction of authentic SMS content.

Once the dataset was loaded, only the pertinent columns were chosen, and the labels for the messages were transformed into numerical values, where 0 indicates ham and 1 signifies spam. This preprocessing step is essential to guarantee compatibility with supervised machine-learning methods.

3.2 Data Pre-processing

Prior to model training, a series of preprocessing operations were applied to transform the raw SMS data into a machine-readable representation. The class labels were converted from categorical form (ham and spam) into binary numerical values to facilitate supervised learning. Text normalization and noise reduction, including the removal of punctuation, stop words, and low-information terms, were performed implicitly during the feature extraction stage through the TF-IDF vectorization process. Subsequently, the dataset was partitioned into training and testing subsets using an 80:20 ratio via the train_test_split procedure. This separation enables model evaluation on unseen samples, thereby providing a reliable and unbiased assessment of generalization performance. Collectively, these preprocessing steps ensure consistency in the input data and contribute to efficient and robust model training.

3.3 Feature Extraction Using TF-IDF

SMS text data were transformed into quantitative feature representations through the application of the Term Frequency–Inverse Document Frequency (TF-IDF) methodology. This statistical measure evaluates term relevance by combining its normalized frequency within a single message with its inverse prevalence across the entire corpus, thereby enhancing the contribution of informative lexical units. To reduce the influence of non-informative tokens, commonly occurring stop words were excluded during preprocessing. As a result, each message was mapped into a sparse, high-dimensional feature space in which each axis corresponds to a distinct term from the constructed vocabulary. Owing to its ability to balance term importance and corpus-level distinctiveness, the TF-IDF model is extensively employed in text-based classification frameworks.

3.4 Model Classification and Evaluation

After feature extraction, the TF-IDF–based representations were used as input to multiple supervised machine learning classifiers for SMS categorization. The classifiers considered in this study include Multinomial Naïve Bayes, Logistic Regression, Support Vector Machines, and Random Forests, selected due to their proven effectiveness in text classification applications. The dataset was partitioned into training and testing sets in an 80:20 ratio, with model learning performed exclusively on the training subset. The trained models were evaluated using the held-out test data, and their performance was quantified through widely adopted metrics such as accuracy, precision, recall, and F1-score. This process enables the identification of discriminative patterns within the feature space to accurately distinguish between spam and legitimate messages.

The system outputs the predicted class label for each SMS message along with the corresponding evaluation metrics. These results facilitate quantitative comparison of the employed classifiers, while graphical visualization of performance indicators such as accuracy and F1-score provides additional insight into the relative effectiveness and robustness of the proposed spam detection approach.

4. Results

The efficacy of four machine learning classifiers—Multinomial Naive Bayes, Logistic Regression, Support Vector Machine (SVM), and Random Forest—was analyzed using the SMS Spam Collection dataset. The models were developed using text features transformed by TF-IDF, and their performance was evaluated based on metrics such as accuracy, precision, recall, F1-score, and training duration. These metrics offer a thorough evaluation of both predictive capabilities and computational efficiency.

Table 1 presents a summary of the comparative outcomes derived from the experimental assessment. Among the tested models, the SVM classifier attained the highest overall accuracy, highlighting its effectiveness in managing high-dimensional sparse text data. Logistic Regression also demonstrated strong performance, achieving accuracy and F1- score metrics that were similar to those of SVM, while also keeping a relatively low computational expense.

5. Figures and Tables

Table 1: Comparative Model Outcomes

Model	Accuracy	Precision	Recall	F1 Score	Training Time
Naïve Bayes	0.966	1.000	0.753	0.859	0.008
Logistic Regression	0.952	0.970	0.666	0.790	0.068
SVM	0.979	0.970	0.873	0.919	0.764

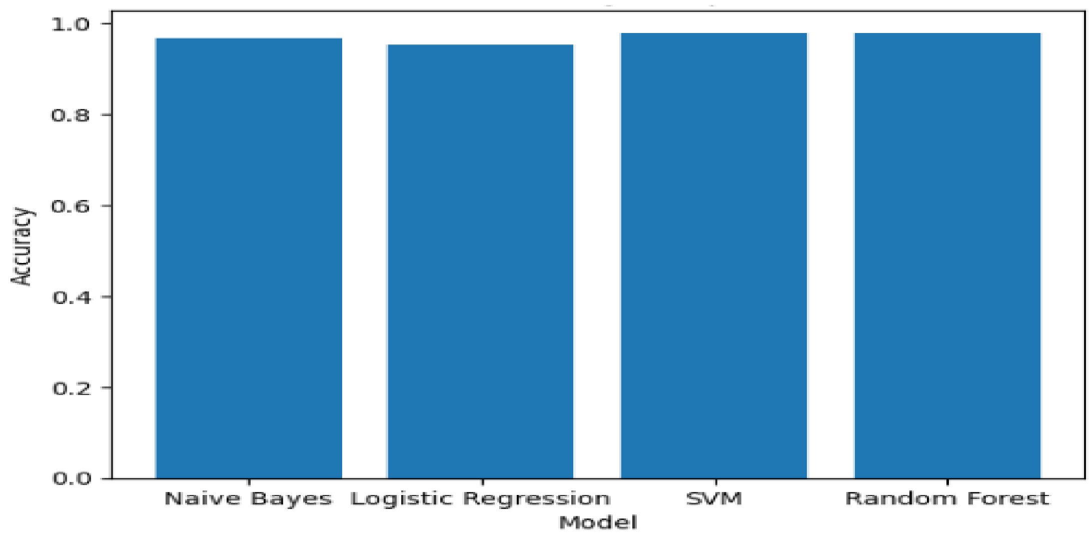
Random Forest	0.977	0.992	0.840	0.909	7.380

Despite its simplifying assumptions, Multinomial Naive Bayes provided dependable performance with the quickest training time of all the classifiers. Its relatively high recall suggests it excels at identifying spam messages, making it suitable for scenarios where minimizing false negatives is essential. In contrast, the Random Forest classifier showed moderate performance accompanied by increased computational demands. Since tree-based models typically function best with structured, non-sparse input data, their efficacy tends to decline when applied to TF-IDF vectors.

A visual comparison of model accuracy is illustrated in Figure. 1, which distinctly showcases the superior performance of SVM and Logistic Regression. The consistent outcomes across various evaluation metrics imply that linear classifiers are more effective for spam detection when utilized alongside TF-IDF-based feature extraction.

In summary, the experimental results affirm that traditional machine learning methods continue to be effective for text-based spam classification. Specifically, SVM and Logistic Regression provide an ideal combination of accuracy, robustness, and computational efficiency. Naive Bayes is advised for environments with limited resources, while Random Forest may need further optimization for use with sparse text features.

Figure 1: Model accuracy comparison



6. Conclusion

In this research, an analysis was performed to compare four machine learning algorithms—Multinomial Naive Bayes, Logistic Regression, Support Vector Machine (SVM), and Random Forest—for the purpose of detecting SMS spam. The SMS Spam Collection dataset was utilized, with messages undergoing preprocessing and transformation into TF-IDF feature vectors to provide an effective representation of the text. The results of the experiments showed that SVM achieved the highest accuracy overall, with Logistic Regression close behind, suggesting that linear models excel with high-dimensional sparse text data. While Multinomial Naive Bayes had slightly lower accuracy, it offered the quickest training time and strong recall performance, which makes it a good choice for lightweight and real-time spam filtering applications. Although Random Forest is known for its robustness in various domains, it performed relatively moderately on TF-IDF features due to their sparse characteristics.

The results of this study indicate that conventional machine learning methods remain highly effective for spam classification tasks when paired with suitable feature extraction techniques. Additionally, the research emphasizes that models that are computationally efficient can achieve competitive performance without the need for deep learning architectures or large resource requirements.

References

1. T. Joachims, “Text Categorization with Support Vector Machines: Learning with Many Relevant Features,” in Proc. ECML, 1998, pp. 137–142.
2. G. V. Cormack, “Email Spam Filtering: A Systematic Review,” Foundations and Trends in Information Retrieval, vol. 1, no. 4, pp. 335–455, 2008.
3. A. Almeida, T. Hidalgo, and I. Alves, “SMS Spam Collection v1.0,” UCI Machine Learning Repository, 2012.
4. F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
5. C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge University Press, 2008.
6. L. Zhang and Q. Duan, “A feature selection method for multi-label text based on feature importance,” Applied Sciences, vol. 9, no. 4, p. 665, 2019.