

A Hybrid Ensemble Framework with Cross-Validation and Hyperparameter Optimization for Robust Spam Email Classification

Hemangi H Joshi¹, Nehal N Kalani²

¹Assistant Professor, Information Technology Department, VVP Engineering College, Gujarat, India

²Assistant Professor, Information Technology Department, VVP Engineering College, Gujarat, India

Abstract - Spam messages continue to threaten digital communication platforms by affecting user productivity and system security. Traditional single-classifier approaches often struggle to maintain high generalization performance across diverse spam patterns. This study proposes a hybrid ensemble framework for robust spam email classification by integrating TF-IDF feature extraction with hyperparameter-optimized machine learning models [1]. The SMS Spam Collection dataset containing 5,572 labelled messages was used for experimental evaluation. Logistic Regression, Support Vector Machine (SVM), and an optimized Random Forest classifier were combined using a soft voting ensemble strategy [2][3]. Hyperparameter tuning was performed using GridSearchCV with 5-fold cross-validation to enhance model reliability. Performance evaluation was conducted using Accuracy, Precision, Recall, F1-score, ROC-AUC, and cross-validation metrics. Experimental results show that the proposed Hybrid Voting Ensemble achieved the highest accuracy of 98.57% and a cross-validation score of 98.13%, outperforming individual models. The findings confirm that ensemble learning improves predictive consistency and classification robustness for spam detection tasks while maintaining computational efficiency suitable for real-time deployment.

Key Words: Spam Detection, Ensemble Learning, TF-IDF, Random Forest, SVM, Cross-Validation, Machine Learning.

1. INTRODUCTION

Electronic communication platforms have become integral to modern society, enabling instant information exchange across the globe. However, the rapid growth of digital communication has led to a significant increase in unsolicited and malicious messages commonly referred to as spam. Spam messages reduce user productivity and pose security risks through phishing attacks, malware distribution, and fraudulent schemes.

Traditional spam detection systems relied on rule-based filtering mechanisms. Although effective in early stages, these approaches lack adaptability to evolving spam patterns. Machine learning-based techniques have gained prominence due to their ability to automatically learn discriminative features from textual data. Algorithms such as Logistic Regression, Support Vector Machines (SVM),

and Decision Trees have been widely applied in spam classification tasks.

Despite their effectiveness, individual classifiers often face limitations in balancing precision and recall. Ensemble learning techniques address this limitation by combining multiple classifiers to improve predictive performance and robustness [4].

This research proposes a hybrid ensemble framework integrating Logistic Regression, Support Vector Machine, and an optimized Random Forest classifier using a soft voting mechanism [8]. The study incorporates TF-IDF feature extraction, hyperparameter optimization using GridSearchCV, and 5-fold cross-validation to ensure model reliability and generalization capability.

The objectives of this research are:

- To evaluate the performance of individual classifiers for spam detection.
- To optimize Random Forest parameters using hyperparameter tuning.
- To design a hybrid ensemble framework for improved classification performance.
- To analyze model robustness using cross-validation and ROC-AUC metrics.

2. LITERATURE REVIEW

Spam detection has been extensively studied in machine learning and text classification research. Early studies focused on Naïve Bayes classifiers due to their simplicity and efficiency in probabilistic text modelling.

Support Vector Machines demonstrated strong performance in high-dimensional text classification tasks due to their margin maximization capability [2]. Random Forest classifiers improved robustness by aggregating multiple decision trees, thereby reducing overfitting and enhancing predictive stability [3].

Recent research emphasizes ensemble learning techniques such as boosting and voting mechanisms to improve generalization performance [4]. Boosting algorithms iteratively improve weak learners, while voting-based ensembles combine predictions from heterogeneous classifiers.

Feature extraction techniques such as TF-IDF have become standard in text classification tasks by assigning importance weights to words based on frequency distribution [6]. Although deep learning approaches such as LSTM and transformer-based models show promising results, traditional machine learning combined with ensemble methods remains computationally efficient and suitable for real-time deployment [9].

This study builds upon prior research by integrating TF-IDF feature engineering, hyperparameter optimization, and hybrid voting-based ensemble learning to enhance spam classification robustness.

3. PROPOSED METHODOLOGY

3.1 Dataset Description

The SMS Spam Collection dataset was used for experimental analysis. The dataset consists of 5,572 labelled messages categorized as spam or ham (non-spam).

3.2 Data Preprocessing

- Label encoding was applied to convert textual labels into binary format.
- TF-IDF vectorization transformed text messages into numerical feature vectors with a maximum of 3000 features.
- The dataset was split into 80% training and 20% testing sets.

3.3 Hyperparameter Optimization

The Random Forest classifier parameters were optimized using GridSearchCV with 5-fold cross-validation. The optimal parameters identified were:

- n_estimators = 200
- max_depth = None

3.4 Hybrid Ensemble Framework

A soft voting ensemble classifier was constructed using:

- Logistic Regression
- Support Vector Machine
- Optimized Random Forest

The final prediction was determined based on weighted probability averaging.

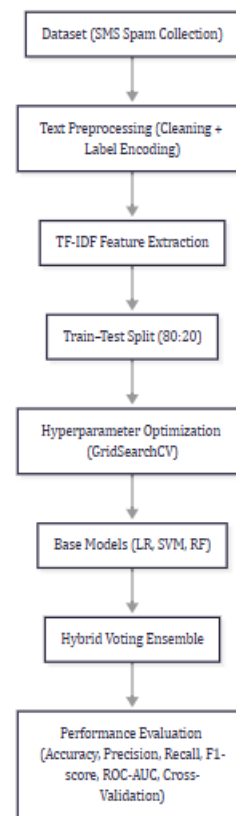


Fig: 1 SMS Spam Detection Pipeline

3.5 Evaluation Metrics

The models were evaluated using:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC
- 5-Fold Cross-Validation Score

4. EXPERIMENTAL RESULTS AND ANALYSIS

Table - 1: Performance Comparison of Models

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	CV Score
Logistic Regression	97.22	100	79.19	88.39	0.9893	96.43
SVM	98.39	100	87.92	93.57	0.9906	97.77
Optimized Random Forest	98.48	100	88.59	93.95	0.9913	97.83

Hybrid Voting Ensemble	98.57	99.26	89.93	94.37	0.9908	98.13
------------------------	-------	-------	-------	-------	--------	-------

The Hybrid Voting Ensemble achieved the highest classification accuracy of 98.57%. The optimized Random Forest demonstrated strong ROC-AUC performance (0.9913), indicating high discrimination capability.

Logistic Regression achieved perfect precision but lower recall, indicating conservative spam classification behaviour. The ensemble model provided improved balance between precision and recall, resulting in the highest F1-score. Cross-validation results confirmed the robustness of the hybrid model with a 98.13% average score, demonstrating strong generalization performance.

5. CONCLUSION

This research introduced a hybrid ensemble-based spam detection framework integrating TF-IDF feature extraction with optimized machine learning classifiers. The soft voting ensemble achieved superior classification accuracy and improved balance between precision and recall compared to individual models. Cross-validation confirmed strong generalization capability and robustness.

Future work may explore deep learning-based text classification approaches, real-time spam filtering deployment, multilingual spam detection, and transformer-based architectures to further enhance model performance.

REFERENCES

- [1] C.D.Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge University Press, 2008.
- [2] T. Joachims, "Text Categorization with Support Vector Machines," in Proc. ECML, 1998.
- [3] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
- [4] Y. Freund and R. Schapire, "A Decision-Theoretic Generalization of On-Line Learning," Journal of Computer and System Sciences, 1997.
- [5] A. McCallum and K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," AAAI Workshop, 1998.
- [6] G. Forman, "An Extensive Empirical Study of Feature Selection Metrics," Journal of Machine Learning Research, 2003.

[7] F. Sebastiani, "Machine Learning in Automated Text Categorization," ACM Computing Surveys, 2002.

[8] H. Drucker, D. Wu, and V. Vapnik, "Support Vector Machines for Spam Categorization," IEEE Transactions on Neural Networks, 1999.

[9] J. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Annals of Statistics, 2001.