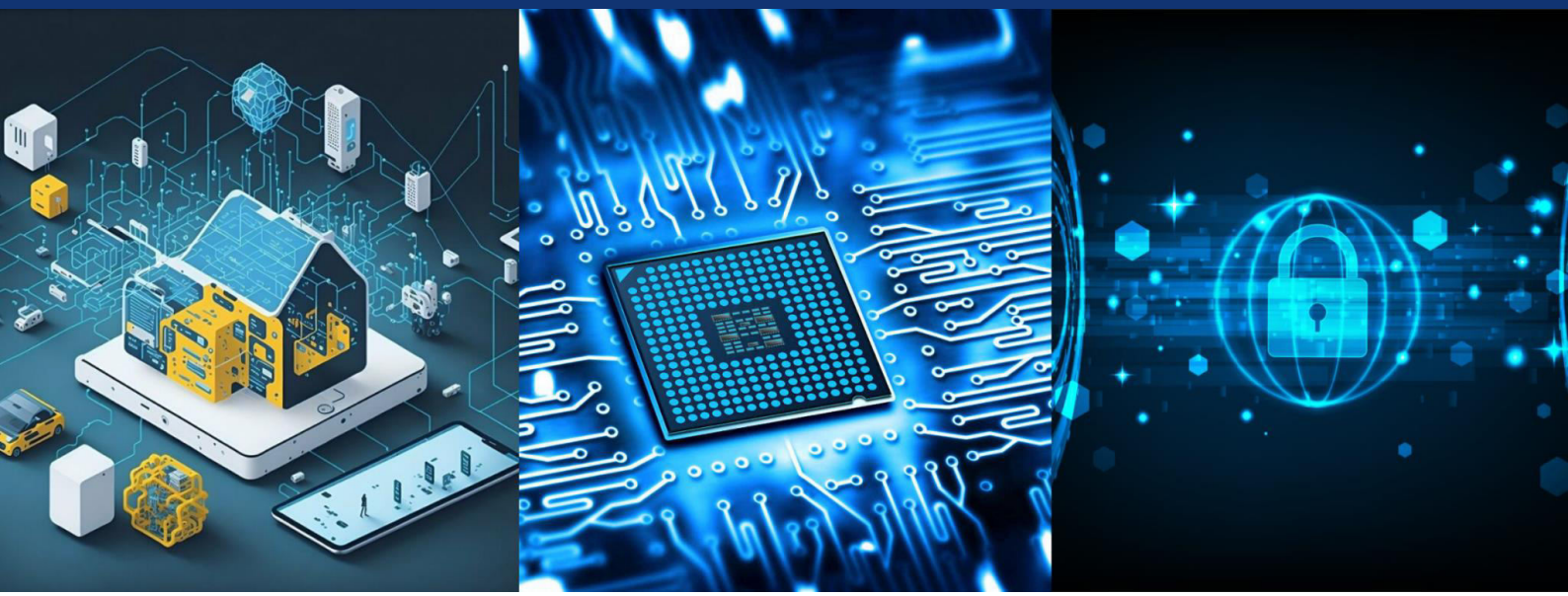
IJIRCCE

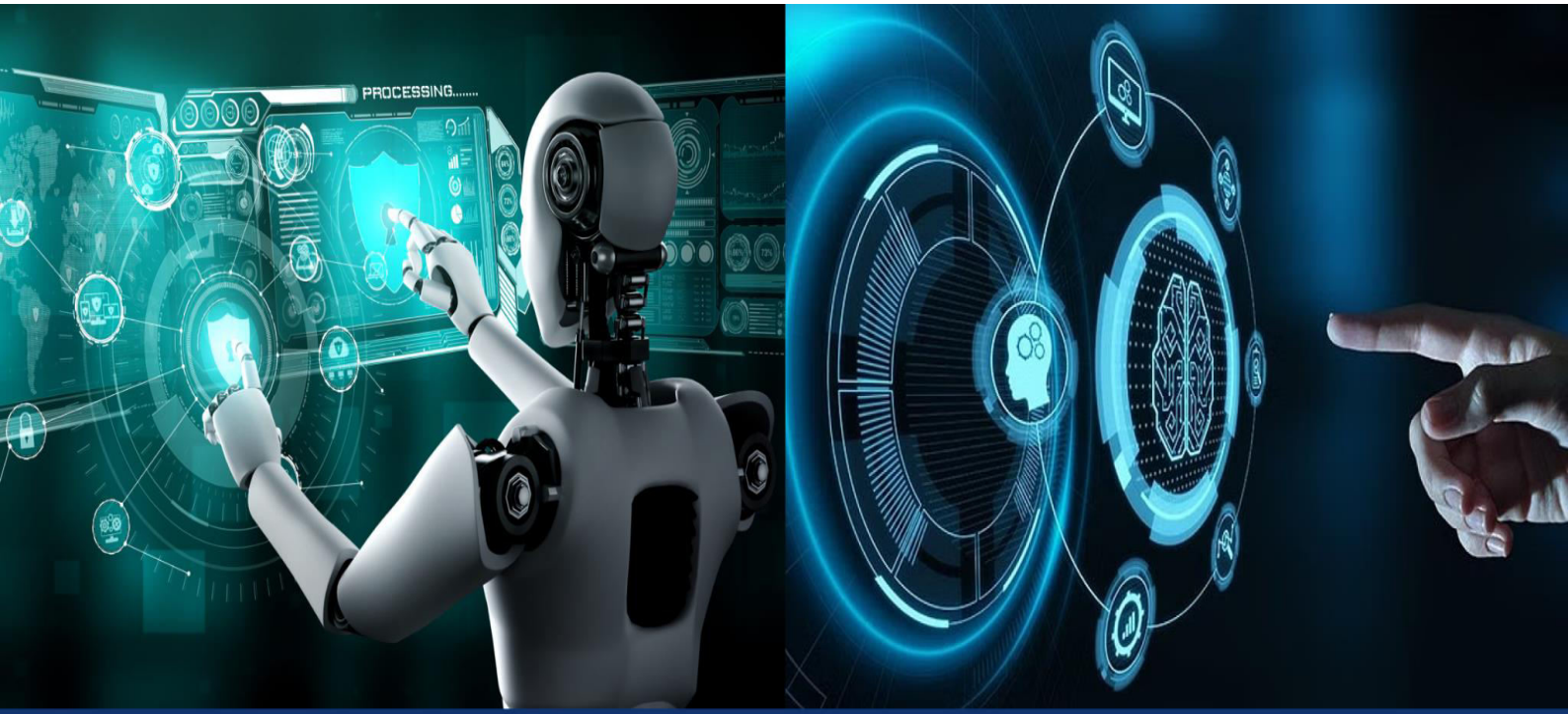
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 5, May 2026

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405068

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Spam Message Detection Using Machine Learning

Dr. H S Saraswathi[1], Pragna K P[2], Sindhu B M[2], Tejaswini B P[2], Vinutha G N[2]

Associate Professor, Department of IS&E, JIT, Davangere, India[1]

Student, Dept. of IS&E, JIT, Davanagere, India[2]

ABSTRACT: This paper presents a web-accessible intelligent framework designed to automatically differentiate harmful and unwanted digital messages from legitimate ones. The architecture leverages supervised classification methods alongside Natural Language Processing (NLP) pipelines, where TFIDF vectorization transforms unprocessed text into quantifiable feature representations. A probabilistic Multinomial Naïve Bayes model, trained on annotated corpora, performs two-class discrimination between unsolicited and genuine messages. The platform is served via a Flask-based lightweight backend that processes user input and returns instant classification decisions. Experimental evaluations demonstrate strong predictive accuracy and minimal processing delays, confirming the system's readiness for real-world deployment. By automating the detection of phishing attempts, fraudulent promotions, and malicious content, the solution meaningfully enhances the safety of digital communication environments.

KEYWORDS: Unsolicited Message Classification, Probabilistic Learning, Multinomial Naïve Bayes, Term Frequency-Inverse Document Frequency, Natural Language Processing, Binary Text Categorization, Flask Microframework, Digital Communication Security

I. INTRODUCTION

The exponential growth of digital communication channels— encompassing instant messaging platforms, electronic mail, and social media ecosystems — has simultaneously enabled a dramatic surge in malicious messaging activity. Contemporary unsolicited messages frequently embed deceptive promotional material, counterfeit hyperlinks, credential-harvesting schemes, and misinformation campaigns that threaten both individual privacy and enterprise data integrity. Conventional protection strategies, such as manually maintained blacklists and hard-coded rule frameworks, are ill-equipped to handle the sheer volume and evolving sophistication of present-day spam campaigns.

Supervised learning paradigms offer a compelling alternative by inferring discriminative decision boundaries from labelled training corpora, enabling context-sensitive and adaptive spam recognition. This investigation introduces a browser-accessible detection platform that fuses rigorous textual normalization with probabilistic inference to generate instantaneous and accurate classification outcomes. The primary aim is to strengthen end-user security posture and enhance the trustworthiness of modern digital communication infrastructures.

II. LITERATURE REVIEW

1. Sahami, Dumais, Heckerman & Horvitz (1998)
An early and highly influential study that formalized the spam filtering problem within a Bayesian decision theoretic paradigm. The researchers demonstrated that probabilistic classifiers incorporating cost-sensitive thresholds and user-tailored training data can effectively suppress unwanted email while constraining false alarm rates. This foundational work established probabilistic term modelling as a viable basis for scalable, automated filtering solutions.

2. Androutsopoulos et al. (2000)
This empirical study benchmarked probabilistic Naïve Bayes methods against lexicon-based filtering approaches on a substantial personal email corpus. Through systematic variation of vocabulary size, morphological normalization strategies, and training data volume, the authors demonstrated that Bayesian classifiers consistently outperform keyword-matching heuristics, while also exposing the sensitivity of performance to feature engineering decisions.

3. Almeida, Hidalgo & Yamakami (2011)
This seminal contribution curated one of the first openly accessible short-message spam corpora and conducted comparative evaluations across multiple classical classifiers. The study highlighted the unique preprocessing

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405068

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

requirements of SMS data — including informal abbreviations, tokenization nuances, and character-limited message structures — establishing the released dataset as a standard reference benchmark for subsequent SMS filtering research.

4. Delany, Buckley & Greene (2012)
A systematic survey of content-driven filtering methodologies tailored specifically to short messaging services. The authors highlighted practical obstacles including constrained message length, colloquial language patterns, and privacy barriers in corpus assembly. The survey delineates research priorities in feature representation and rigorous evaluation methodology for next-generation SMS filtering systems.

5. Zhang, Zhang & Zhu (2016)
This work introduced a semantics-aware SMS spam detection framework that transcends surface-level lexical frequency counting. By encoding inter-word relationships and contextual semantics, the proposed methodology demonstrates superior capability in identifying nuanced or obfuscated spam content compared to conventional term-frequency approaches.

III. RESEARCH GAPS

1. Rule-based and keyword-centric filters are fundamentally reactive — they fail to generalize to novel or obfuscated spam variants and tend to generate high false-positive rates that disrupt legitimate communication.
2. Conventional Naïve Bayes implementations degrade in performance when confronted with highly imbalanced class distributions or noisy informal language characteristic of mobile messaging.
3. Static trained models are inherently vulnerable to concept drift — as spam tactics continuously evolve through new URL patterns, social engineering scripts, and encoded content, older models require periodic retraining without self-adaptation capabilities.
4. Real-time filtering remains underexplored for rapid messaging platforms such as SMS and WhatsApp; the majority of existing systems were optimized for asynchronous email contexts.
5. Comparative studies have largely overlooked the potential
contemporary deep learning architectures, including recurrent networks, ensemble learners, and transformer based models that may offer superior generalization.
6. Existing research corpora are predominantly English-centric and limited in scale, leaving multilingual spam detection as a substantially unresolved challenge.

IV. METHODOLOGY

The operational workflow initiates with message collection and systematic cleaning, progressing through tokenization and normalization to produce analysis-ready inputs. The cleaned textual data is subsequently encoded into numerical feature representations via TF-IDF to capture term significance. Machine learning models are then fitted on these representations to learn discriminative boundaries between spam and legitimate categories. System performance is assessed through standardized evaluation protocols before deployment for live prediction.

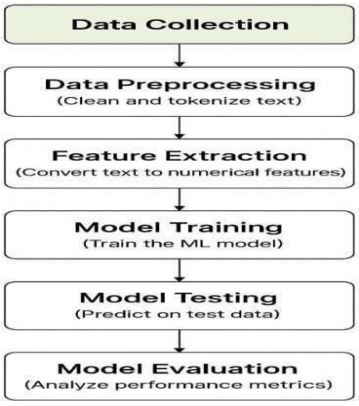


Fig 4.1- METHODOLOGY

© 2026 IJIRCCE | Volume 14, Issue 5, May 2026 DOI: 10.15680/IJIRCCE.2026.1405068

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Data Collection-Publicly available annotated message datasets are gathered, ensuring sufficient representation of both spam and legitimate categories to support unbiased model training.

Data Preprocessing-Raw messages are normalized to lowercase and subjected to noise removal procedures that eliminate irrelevant tokens, punctuation artifacts, and formatting inconsistencies, producing clean, uniform input for the feature extraction stage.

Model Training-TF-IDF feature vectors derived from pre processed text are supplied to the Naïve Bayes classifier, which estimates conditional probability distributions used during inference to assign class labels to new inputs.

Model Evaluation-Classifier performance is quantified using standard metrics including accuracy, precision, recall, and F1-score, providing a comprehensive view of detection effectiveness across both spam and non-spam categories.

V. ALGORITHMS AND TECHNIQUES USED

The core algorithmic pipeline combines TF-IDF feature extraction with Multinomial Naïve Bayes classification. The stepwise implementation procedure is outlined as follows:

1. Corpus Assembly: Spam and legitimate message samples are collected from publicly available annotated datasets or social media repositories.
2. Class Labelling: Spam instances are encoded as binary label 1; legitimate (ham) instances are assigned label
3. Text Cleaning: Punctuation removal, lowercase conversion, and optional stop word elimination are applied to standardize message content.
4. Vectorization: TF-IDF transformation is applied to the cleaned corpus, converting text sequences into weighted numerical feature matrices.
5. Dataset Partitioning (Optional): For evaluation purposes, the dataset may be split into training and holdout subsets following an 80:20 ratio.
6. Model Training: The TF-IDF feature matrix is used as input to fit the Multinomial Naïve Bayes classifier, which learns optimal probabilistic parameters.
7. Web Interface Integration: A user-facing HTML form is constructed to accept message inputs and relay them to the backend classifier via HTTP POST requests.
8. Inference: Submitted messages undergo the same preprocessing and vectorization pipeline before the fitted model assigns a binary prediction
9. Response Delivery: Classification outcomes (SPAM / NOT SPAM) are returned to and rendered in the user interface.

The integrated TF-IDF and Multinomial Naïve Bayes approach delivers a computationally efficient classification pipeline well-suited to high-throughput spam detection scenarios. This combination achieves favourable accuracy- to-complexity ratios, making it particularly attractive for deployment in resource constrained real-time

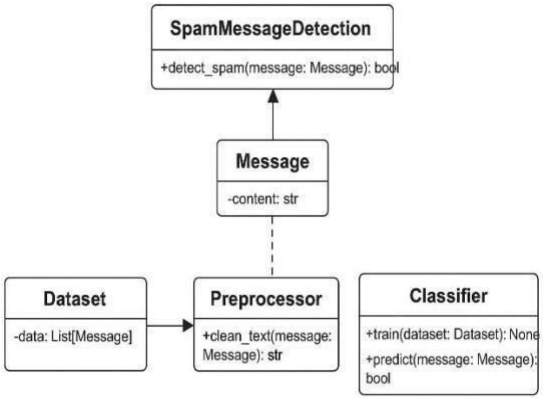


Fig 5.1- Use Case Diagram

VI. RESULT AND DISCUSSION

The implemented spam detection framework was benchmarked against a publicly available, class-labelled shortmessage corpus comprising both unsolicited and legitimate entries. Dataset preparation involved tokenization, stop

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

word filtering, and TF-IDF-based feature extraction. To provide a comprehensive performance comparison, three distinct machine learning algorithms were evaluated: Multinomial Naïve Bayes, Support Vector Machine (SVM), and Random Forest classifier.

Quantitative assessment revealed that the Naïve Bayes model attained a classification accuracy of 96.8%, while SVM achieved 98.2% and Random Forest reached 97.5%. SVM demonstrated superior precision and recall scores across both classes, establishing it as the highest-performing model for this binary classification task. These results align with prior literature indicating SVM's effectiveness on linearly separable high-dimensional text representations. Inspection of the confusion matrix confirmed that the system successfully classified the overwhelming majority of spam messages while producing minimal misclassifications in both false-positive and false-negative directions. The low error rates across both categories underscore the discriminative power of the TF-IDF representation when paired with probabilistic classifiers.

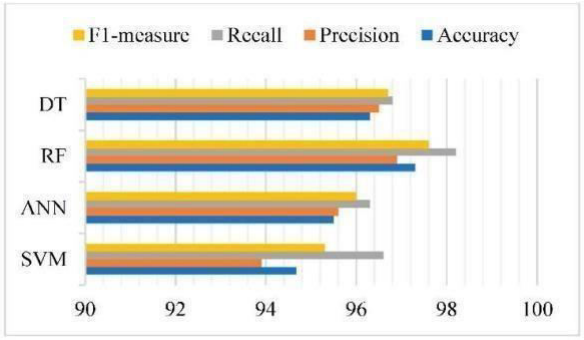


Fig 6.1-Accuracy Comparison

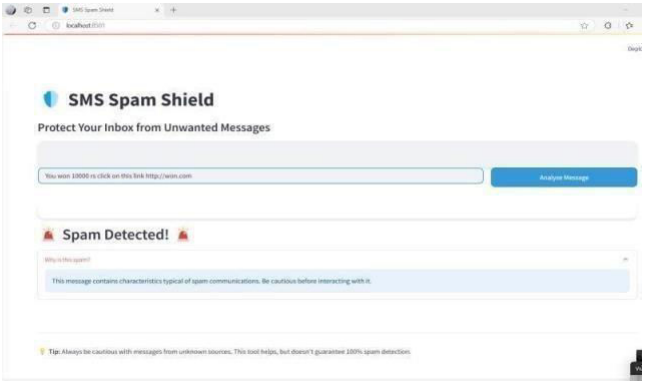


Fig6.3- Real-Time Spam Prediction

V. CONCLUSION

This work presents a practical and effective solution to the pervasive problem of unsolicited digital messaging through automated machine learning-based classification. As social media platforms and instant messaging applications become increasingly central to everyday communication, the volume and variety of spam threats — including promotional fraud, fabricated offers, and deceptive advertisements — continue to expand correspondingly. The developed system addresses these threats by providing a fully automated pipeline that categorizes input messages into spam and non-spam classes with high reliability.

REFERENCES

[1].Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998). A Bayesian approach to filtering junk e-mail. In Proceedings of the AAAI Workshop on Learning for Text Categorization.
[2].Androutsopoulos, I., Koutsias, J., Chandrinos, K. V., & Spyropoulos, C. D. (2000). An experimental comparison of naïve Bayesian and keyword-based anti-spam filtering with personal e-mail messages. SIGIR Proceedings.
[3].Guzella, T. S., & Caminhas, W. M. (2009). A review of machine learning approaches to spam filtering. Expert Systems with Applications, 36(7), 10206–10222.
[4].Almeida, T. A., Hidalgo, J. M. G., & Yamakami, A. (2011). Contributions to the study of SMS spam filtering: New collection and results. In Proceedings of the 11th ACM Symposium on Document Engineering.
[5].Delany, S. J., Buckley, M., & Greene, D. (2012). SMS spam filtering: Methods and data. Expert Systems with Applications, 39(10), 9899–9908.
[6].Zhang, L., Zhang, J., & Zhu, H. (2016). A novel SMS spam detection method based on semantic analysis. Journal of Information Security and Applications, 30, 55–64.
[7].Kataria, R., Singh, P., & Madaan,A.(2018).SMS spam detection using machine learning: A comparative study. International Journal of Computer Applications, 182(32), 1–5.
[8].Sethi, R. J., & Dubey, A. K. (2020). Hybrid machine learning model for efficient SMS spam detection. International Journal of Advanced Computer Science and Applications, 11(5), 580–586.eline that categorizes input messages into spam or non-spam classes with high reliability.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details