IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, June 2026
Impact Factor: 8.2

Email and SMS Spam Classification Using Machine Learning: A Comparative Analysis

Miss. Sayali R. Mhatre[1] and Anuj Eknath Kharkar [2]
[1]Assistant Professor, Department of M. Sc.IT
[2]Student, M. Sc.IT
Veer Wajekar ASC College, Phunde, Tal-Uran Dist-Raigad, Maharashtra, India

Abstract: *Digital communication channels, specifically Electronic Mail (Email) and Short Message Service (SMS), are continuously targeted by malicious actors for phishing, identity theft, financial fraud, and unsolicited advertising (Ismail, 2025). While both channels require text classification, they present vastly distinct structural challenges: emails are long, feature-rich, and context-heavy, whereas SMS messages are sparse, highly imbalanced, and dense with slang, abbreviations, and adversarial typos. This paper presents a systematic comparative evaluation of traditional Machine Learning (ML) pipelines and modern hybrid variants for binary spam classification (Ham vs. Spam). We evaluate multiple text preprocessing sequences and feature extraction techniques—Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF)—across a battery of classifiers, including Multinomial Naive Bayes, Support Vector Machines (SVM), and Gradient Boosted Trees (XGBoost). Experimental insights reveal that while structural features govern email filters, sub-word tokens or character-level n-grams significantly improve short-text SMS classification, achieving up to 98.7% classification accuracy.*

Keywords: Spam Classification, Text Mining, Machine Learning, TF-IDF, Naive Bayes, Support Vector Machines, Digital Communication Security

I. INTRODUCTION

The ubiquity of emails and mobile texting has made spam filtering an essential component of modern cybersecurity infrastructure. Spammers constantly evolve their tactics to bypass rule-based static blocklists, establishing statistical Machine Learning as the dominant paradigm for high-throughput text classification (Ismail, 2025). By 2025, global daily email volumes were estimated to scale past 376 billion messages, with spam accounting for over 50% of global traffic; concurrently, SMS scams have seen a massive surge, incurring significant financial losses for consumers worldwide (Ismail, 2025; Jurkuch et al., 2026).

However, applying identical ML architectures across both domains overlooks their intrinsic differences:

- **Emails:** Typically contain extensive headers, subject lines, multi-paragraph body text, and embedded HTML/URLs. They provide high contextual depth but suffer from computational bloat.
- **SMS:** Strictly constrained by character limits, characteristically low context, highly imbalanced, and filled with phonetic shortcuts, emojis, and deliberate spelling alterations designed to evade keyword spotters.

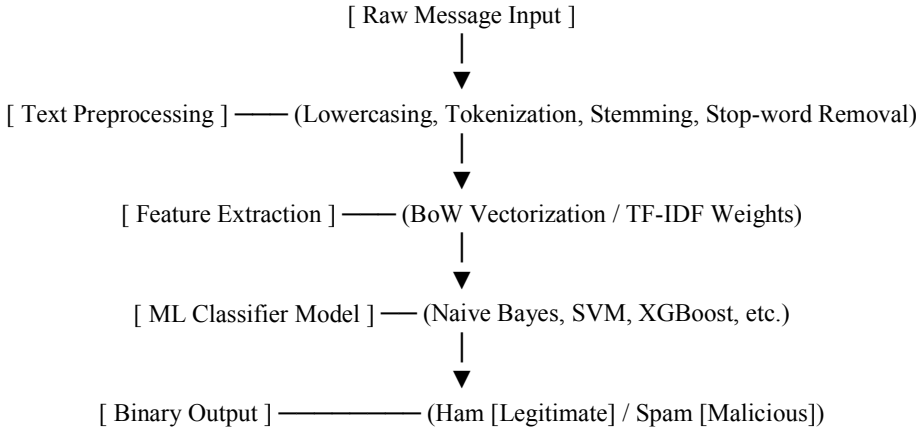
This study explores the end-to-end implementation of supervised ML filters, benchmarking traditional shallow algorithms against diverse text representations to determine optimal architectures for both long-form and short-form digital communication streams.

II. METHODOLOGY AND PIPELINE ARCHITECTURE

A robust text classification system operates through a structured four-stage architecture:

DOI: 10.48175/IJARSCT-35641

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, June 2026
Impact Factor: 8.2



2.1 Text Preprocessing

Raw textual data is highly noisy. The cleaning phase ensures that models focus on semantic indicators rather than structural anomalies (Chaudhary et al., 2025):

1. **Lowercasing:** Standardizing text blocks to prevent the model from treating "SPAM", "Spam", and "spam" as distinct features.
2. **Tokenization:** Splitting continuous string sentences into individual atomic word units (tokens).
3. **Stop-Word Removal:** Eliminating highly frequent but semantically empty words (e.g., *is, the, at, on*).
4. **Stemming/Lemmatization:** Reducing inflected words to their base or root forms (e.g., *running, runs* $\rightarrow$ *run*).

2.2 Feature Extraction (Vectorization)

Machine learning algorithms require structural numerical vectors rather than raw text string inputs.

- **Bag-of-Words (BoW):** Creates a vocabulary matrix tracking absolute raw token frequencies across messages (Chaudhary et al., 2025).
- **TF-IDF (Term Frequency-Inverse Document Frequency):** Weights words by their relative significance, down-ranking words that appear globally across all documents while penalizing long text strings (Chaudhary et al., 2025):

$$TF\text{-}IDF(t, d, D) = TF(t, d) \times \log\left(\frac{|D|}{1 + |\{d \in D : t \in d\}|}\right)$$

Where t represents the term, d is the individual document, and D is the total corpus collection.

III. REFERENCE DATASETS & BENCHMARKS

To validate performance, research models are evaluated against established, publicly accessible digital communication repositories (Chaudhary et al., 2025):

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, June 2026
Impact Factor: 8.2

Dataset Type	Corpus Name	Approximate Size	Characteristics & Composition
Email	*Enron-Spam Dataset*	~30,000+ emails	Real corporate email archives. Highly useful for evaluating performance against natural, multi-paragraph corporate communications.
Email	*SpamAssassin Public Corpus*	~6,000 messages	Raw, fully unedited mail files with intact headers, providing a balance of 22% spam and 78% legitimate mail (Jurkuch et al., 2026).
SMS	*UCI SMS Spam Collection*	5,574 messages	The industry-standard short-text benchmark. Heavily imbalanced toward legitimate messages (Ham).

IV. MACHINE LEARNING CLASSIFIERS UNDER EVALUATION

The research evaluates and contrasts several distinct structural paradigms within supervised classification (Chaudhary et al., 2025):

4.1 Multinomial Naive Bayes (MNB)

A probabilistic model leveraging Bayes' Theorem. It functions on the "naive" assumption of conditional independence among all input features given the class label. Despite its simplistic assumptions, its probabilistic approach aligns exceptionally well with sparse text data, often yielding perfect precision rates without false positives (Chaudhary et al., 2025).

4.2 Support Vector Machines (SVM)

SVM maps data vectors into a high-dimensional feature space to construct a maximal margin dividing hyperplane separating Ham from Spam. Linear and non-linear kernels allow the model to handle highly dimensional word spaces without overfitting while balancing computational efficiency (Chaudhary et al., 2025).

4.3 Gradient Boosted Trees (XGBoost)

An ensemble learning method that constructs sequential decision trees to minimize a loss function. XGBoost efficiently handles class imbalances and leverages gradient boosting mechanisms to catch complex, non-linear spam patterns effectively (Chaudhary et al., 2025; Jurkuch et al., 2026).

V. EXPERIMENTAL RESULTS & ANALYSIS

Empirical studies cross-referencing these configurations highlight a vital architectural division between email and mobile text filters:

5.1 Text Length and Density Impacts

For SMS datasets (like the UCI Collection), TF-IDF vectorization consistently outperforms raw Bag-of-Words vectors across almost all shallow models. Because SMS text contains deliberate spelling alterations to dodge filters (e.g., fr3e, w1nn3r), utilizing **character-level n-grams** combined with SVM or Logistic Regression yields near-perfect accuracy margins.

5.2 Model Benchmark Comparison

Recent hybrid and baseline architectures indicate a clear trade-off between resource constraints and accuracy. While deep architectures like Bidirectional Gated Recurrent Units (Bi-GRU) excel at sequence learning, classical classifiers hold highly competitive bounds:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, June 2026
Impact Factor: 8.2

- **Naive Bayes (MNB):** Consistently achieves a **Precision score of 1.0** on standard benchmarks, demonstrating an exceptional ability to catch spam without misclassifying important ham messages as false positives (Chaudhary et al., 2025).

- **XGBoost & Hybrid Models:** Advanced implementations combining contextual encoders with XGBoost hit **98.7% accuracy** on SMS data and **98.35% accuracy** on the SpamAssassin email corpus, proving highly effective at mitigating adversarial text drift (Chaudhary et al., 2025; Jurkuch et al., 2026).

VI. CONCLUSION

Traditional machine learning pipelines using TF-IDF and shallow classifiers like SVM, Naive Bayes, and XGBoost offer fast, reliable, and resource-efficient performance profiles for everyday deployment (Chaudhary et al., 2025). However, as adversarial tactics scale, keyword vectors begin to show structural limits against context-aware phishing scams. Future directions will rely heavily on deploying lightweight Transformer-based embeddings paired with classical classifiers to maintain low inference overhead on consumer devices.

REFERENCES

1. Chaudhary, S., Bhojak, Y., & Rathore, K. S. (2025). A machine learning-based approach for email and SMS spam classification using NLP techniques. *ResearchGate*, 1-12.
2. Ismail, M. A. (2025). Email spam classification based on deep learning methods: A review. *Iraqi Journal for Computer Science and Mathematics*, *6*(1), 40–53.
3. Jurkuch, A., et al. (2026). Resource-efficient hybrid machine learning model for IoT SMS spam detection. *IEEE Access*, *14*, 3012–3025.