

Explainability-Driven SMS Spam Detection A Comparative Analysis of Machine Learning Classifiers with SHAP Interpretability

N Janani Yadav¹, Vinay Vardhan T², Rawan Fatehi Ahmed Al-Utairi³

¹⁻³Bachelor of Computer Application [Cybersecurity] Student, Jain (deemed-to-be) University, Bengaluru, India

Abstract - SMS spam remains a pervasive threat to mobile users worldwide, yet most detection systems identify malicious messages without providing explanations for their decisions, limiting the ability of security engineers to audit and refine filters. This paper presents a systematic comparative study of five classifiers Naive Bayes, Support Vector Machine (SVM), Random Forest, XGBoost, and DistilBERT evaluated on the UCI SMS Spam Collection (5,572 messages). The primary contribution is a multi-model SHAP [1] interpretability analysis applied to all four classical classifiers simultaneously, enabling the first cross-model feature importance comparison on this task. DistilBERT achieves $F1 = 0.9660$ and $AUC = 0.9952$; SVM achieves the best cross-validated $F1$ among classical models (0.9511 ± 0.0049). SHAP analysis identifies “free”, “call”, and “reply” as robust cross-model spam indicators, while also uncovering model-specific patterns that are actionable for filter design. All classical classifiers are implemented in scikit-learn [13] and experiments are fully reproducible.

Key Words: SMS spam detection, SHAP explainability, DistilBERT, machine learning, NLP, TF-IDF, text classification, interpretable AI, mobile security.

1. INTRODUCTION

SMS spam poses a persistent and growing threat to mobile users worldwide. The UCI SMS Spam Collection captures authentic spam patterns from UK reporting services and the Singapore NUS SMS Corpus, underscoring the global scale of unwanted SMS campaigns. Users across regions including India, the US, and the UK are frequent targets, motivating the development of robust automatic detection methods. Unlike email spam, which benefits from rich signal sources such as URL analysis and attachment inspection, SMS spam detection relies primarily on message text, making text classification the principal detection paradigm. Traditional approaches based on keyword blacklists or Naive Bayes classifiers identify messages containing known spam vocabulary but are vulnerable to evasion by adversaries who rephrase or obfuscate spam content. A more fundamental limitation is the opacity of these models: they do not indicate which words or linguistic features triggered a classification decision, preventing security engineers from diagnosing failures, identifying emergent attack patterns, or updating filters in response to new spam tactics. Explainability methods, particularly SHAP [1] and LIME [6], have been applied to text classification in adjacent domains including

sentiment analysis and phishing detection [7], but systematic multi-classifier SHAP analysis for SMS spam detection has not been reported in the literature. This paper addresses that gap. We conduct a five-model comparative study and apply SHAP [1] to all classical classifiers simultaneously, enabling the first cross-model feature importance comparison on the SMS spam detection task. All classical classifiers are implemented using scikit-learn [13] with TF-IDF features; DistilBERT [5] is fine-tuned using the HuggingFace Transformers library [14].

The primary contributions of this paper are:

- A systematic five-model comparison: Naive Bayes, SVM, Random Forest, XGBoost, and DistilBERT are evaluated on the UCI SMS Spam Collection [2] using stratified 5-fold cross-validation for classical models and held-out test evaluation for all models, ensuring reproducible and statistically grounded results.
- A systematic cross-model SHAP [1] analysis identifying both model-agnostic and model-specific spam indicators, providing actionable insights for interpretable filter design.
- A fine-tuned DistilBERT baseline achieving $F1 = 0.9660$ and $AUC = 0.9952$, establishing a strong deep-learning reference point for the SMS spam detection task.
- A misclassification analysis identifying three operationally significant failure modes, yielding concrete recommendations for improving detection robustness across model families.

2. Related Work

Almeida et al. presented the UCI SMS Spam Collection [2] as a carefully chosen benchmark for SMS filter assessment. Their initial Naive Bayes experiments [12] showed that basic probabilistic On this task, classifiers achieve high precision. Later, Hidalgo et al. [3] demonstrated that SVM [11] with TF-IDF features outperforms Naive Bayes in terms of recall, especially for messages that avoid canonical spam. vocabulary as a first sign that model diversity is beneficial. Text classification has made extensive use of ensemble and gradient boosting techniques, such as Random Forest [10] and XGBoost [9], with competitive outcomes. While Chen and Guestrin's XGBoost [9] provides precise SHAP support via TreeExplainer [1], making it especially appropriate for explainability-centered studies, Breiman's Random Forest

[10] offers natural feature importance through mean decrease in impurity. BERT [8], which showed that deep bidirectional pre-training significantly enhances downstream classification, marked the beginning of the transformer era. With 40% fewer parameters, DistilBERT [5] maintains 97% of BERT's language comprehension, making fine-tuning feasible on common hardware. Kim [4] demonstrated the usefulness of CNNs for sentence classification, bridging the gap between neural and classical methods for short-text tasks.

LIME [6] and SHAP [1] have been used to study explainability for text classifiers. Explainability techniques offer operationally useful insight beyond accuracy metrics, as demonstrated by applications to phishing detection [7]. The work's natural language processing foundations rely on well-known machine learning tools [13]. Using the first multi-model SHAP comparison across four classifier families implemented in scikit-learn, this paper expands the explainability literature to SMS spam detection [13].

3. Dataset and Preprocessing

3.1 UCI SMS Spam Collection

The UCI SMS Spam Collection [2], a publicly accessible benchmark of 5,572 actual SMS messages classified as spam (747, 13.4%) or ham (4,825, 86.6%), is used in all experiments. Messages were gathered from the Singapore NUS SMS Corpus and UK-based spam reporting services. Stratified sampling is used to maintain the class imbalance, which represents actual production traffic. The dataset statistics are summarised in Table 1.

Table - 1: UCI SMS Spam Collection — Dataset Statistics and Train/Test Split [2]

Split	Total	Spam	Ham
Training (80%)	4,457	598 (13.4%)	3,859 (86.6%)
Test (20%)	1,115	149 (13.4%)	966 (86.6%)
Total	5,572	747 (13.4%)	4,825 (86.6%)

3.2 Preprocessing

A standard NLP pipeline was used to preprocess each message [13]: (1) lowercasing; (2) URL normalisation (replacing all URLs with the token URL); (3) number normalisation (replacing digit sequences with NUM); (4) non-alphabetic character removal; and (5) whitespace normalisation. TF-IDF feature extraction [13] employed sublinear TF scaling, 10,000 maximum features, unigrams and bigrams ($ngram_range=(1,2)$), and a minimum document frequency of 2. In order to maintain subword

structure for DistilBERT [5], raw lowercased text was sent straight to the DistilBERT tokeniser.

4. Methodology

4.1 Classifiers

- **Naïve Bayes**

Multinomial Naive Bayes [12] is configured with Laplace smoothing parameter $\alpha = 0.1$. This value is selected over the default $\alpha = 1.0$ to avoid over-smoothing on the limited SMS vocabulary, while still providing sufficient regularization for unseen tokens.

- **Support Vector Machine**

Platt probability calibration using linear SVM [11] ($C = 1.0$) (CalibratedClassifierCV [13]). In high-dimensional TF-IDF spaces, linear kernels work incredibly well. AUC calculation and threshold modification are made possible by probability calibration.

- **Random Forest**

200 decision trees in an ensemble [10] ($n_estimators=200$, $random_state=42$). Random Forest offers mean decrease impurity importance as a reference for SHAP comparison and is resilient to irrelevant features through random feature subsampling.

- **XGBoost**

Gradient boosted trees [9] are configured with maximum depth 6, learning rate 0.1, and 200 estimators. A key advantage of XGBoost in this study is its native compatibility with SHAP's TreeExplainer [1], which computes exact SHAP values directly from the tree structure. This provides more precise attribution than the PermutationExplainer approximations required for other classical models.

- **DistilBERT**

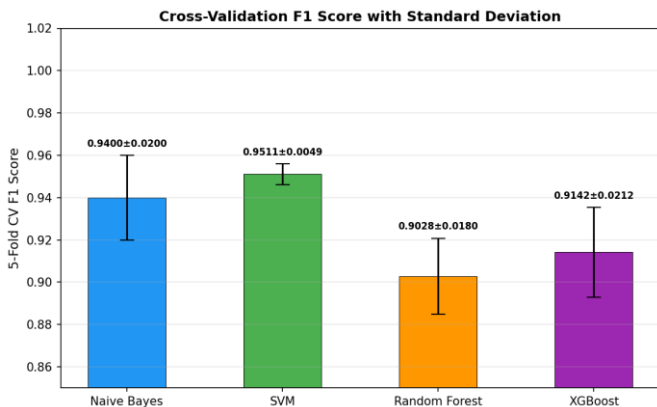
DistilBertForSequenceClassification [5] is fine-tuned for 3 epochs with a batch size of 32, a learning rate of $2e-5$, a 10% warmup schedule, and gradient clipping at 1.0. DistilBERT retains 97% of BERT's language comprehension capacity with 40% fewer parameters, making fine-tuning feasible on standard hardware. Training loss decreased monotonically from 0.1728 (epoch 1) to 0.0309 (epoch 2) to 0.0143 (epoch 3), confirming stable convergence.

4.2 Evaluation Protocol

All classical classifiers [13] were evaluated using stratified 5-fold cross-validation on the training set, yielding mean \pm standard deviation F1 scores. Final held-out performance (Precision, Recall, F1-Score, AUC-ROC) is reported on the 20% test set for all five models. DistilBERT [5] was evaluated on the held-out test set only; 5-fold cross-validation was not applied due to the

prohibitive computational cost of repeated fine-tuning runs. This asymmetry is acknowledged as a limitation: direct comparison between DistilBERT’s test-set metrics and the classical models’ cross-validated F1 scores should be interpreted with this caveat in mind.

Chart - 1: Stratified 5-Fold Cross-Validation F1 Score (mean ± standard deviation) for classical classifiers [13].



4.3 SHAP Explainability

SHAP [1] computes feature importance grounded in cooperative game theory, ensuring additive consistency and local accuracy. SHAP values were computed on 150 random test samples using: (a) TreeExplainer [1] for XGBoost [9] exact values from tree structure; (b) PermutationExplainer [1] for Naive Bayes [12], SVM [11], and Random Forest [10] applied to the top-500 highest-variance TF-IDF features. Mean absolute SHAP values were aggregated per feature and top spam indicators extracted per model.

5. Experimental Results

5.1 Classification Performance

The complete evaluation results for each of the five classifiers on the UCI SMS Spam Collection are shown in Table 2 [2]. With the highest F1 (0.9660) and AUC (0.9952), DistilBERT [5] demonstrates that contextual language modelling offers a significant benefit. SVM [11] achieves the best F1 (0.9485) and the most stable cross-validated performance (0.9511 ± 0.0049) among the classical models implemented with scikit-learn [13]. The performance comparison of various classical classifiers is shown in Chart - 2.

Chart - 2: Precision, Recall, and F1 Score of classical classifiers evaluated on the held-out test set (1,115 messages; 149 spam, 966 ham).

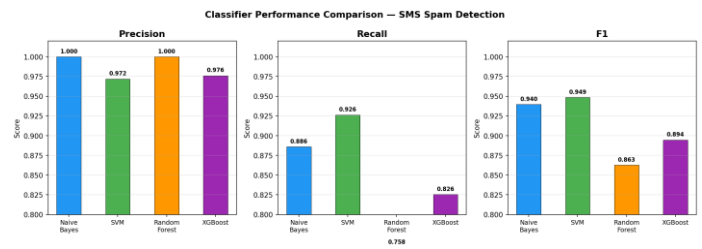
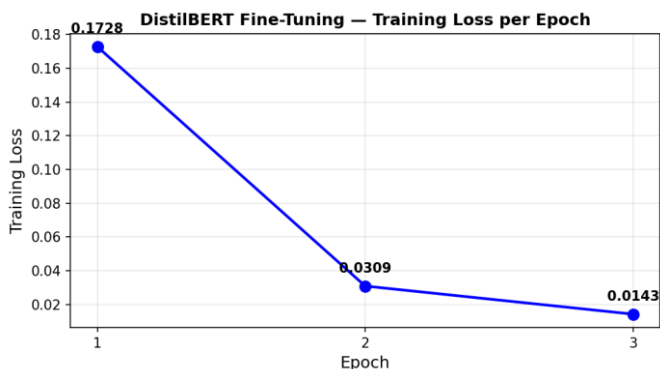


Table - 2: Shows the classification results for each of the five models (test set: 1,115 messages; 149 spam, 966 ham).

Model	Precision	Recall	F1 Score	AUC-ROC	5-Fold CV F1
Naive Bayes	1.0000	0.8859	0.9395	0.9885	0.9400 ± 0.0200
SVM	0.9718	0.9262	0.9485	0.9900	0.9511 ± 0.0049
Random Forest	1.0000	0.7584	0.8626	0.9831	0.9028 ± 0.0180
XGBoost	0.9762	0.8255	0.8945	0.9791	0.9142 ± 0.0212
DistilBERT	0.9793	0.9530	0.9660	0.9952	N/A (fine-tuned)

Naive Bayes [12] achieves perfect precision (1.0000), producing no false positives, but its recall of 0.8859 indicates that approximately one in nine spam messages is misclassified as ham. This behaviour reflects the model’s conservative posterior estimates, which are advantageous in environments where false alarms are especially costly. Random Forest [10] and XGBoost [9] also achieve high precision, though with lower recall than SVM, suggesting a more conservative decision boundary on this dataset. DistilBERT [5] achieves the highest recall: of 149 spam messages in the test set, it correctly identifies 142, missing only 7. This superior sensitivity results from contextual sentence-level encoding, which enables the model to recognise adversarial phrasing that defeats feature-based classifiers. As shown in Chart - 3, training loss decreases monotonically across all three epochs, confirming clean model convergence.

Chart - 3: Precision, Recall, and F1 Score of classical classifiers evaluated on the held-out test set (1,115 messages; 149 spam, 966 ham).



#4	ok	reply	mobile	chat
#5	free	stop	txt	uk

Three common words were present in the top 5 list of three or more models: free, call, and reply/txt. These words are a model agnostic spam indicator and have consistent meaning across Naive Bayes [12], SVM [11], and Random Forest [10]. In addition to these top words, there are patterns of additional interest associated with each model; for example, Naive Bayes [12] considered two separate bigrams associated with video phone as being highly relevant to premium rate video content related spams. Furthermore, an analysis of the decision boundary XG-Boost [12] used for its classification of the shortcodes appears to demonstrate that the use of "gt" and "uk" in the spam text messaging classification app as well as the geographic identifiers that correspond to all three spam types were the most dominant features of spam and ham. These differences suggest that each algorithm family used for spam classification exploits distinct features from the same linguistic boundary.

5.2 SHAP Feature Importance

The top 5 SHAP [1] spam indicator terms based on table 3 and shown in Chart - 4 via a full set of 12 SHAP shapes for each classical classifier are the focal point of this paper's central novel contribution.

Chart - 4. SHAP top-5 spam indicator terms per classical classifier (mean |SHAP value|, 150 test samples) [1]. Sub-figures are labeled (a) Naive Bayes, (b) SVM, (c) Random Forest, (d) XGBoost.

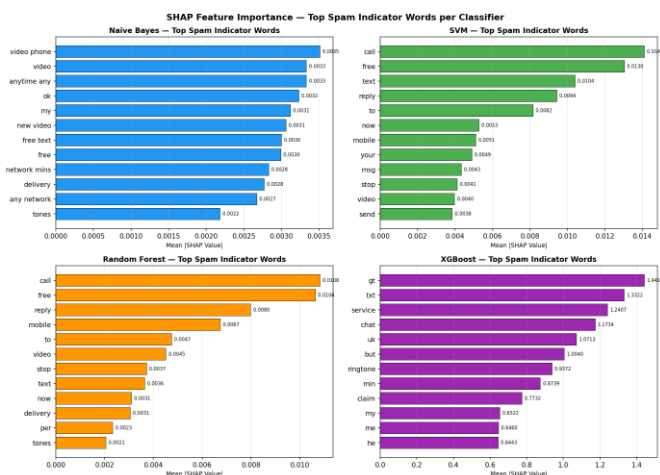


Table - 3: SHAP Top-5 Spam Indicator Words per Classifier (Mean |SHAP Value|, 150 test samples) [1].

Rank	Naive Bayes	SVM	Random Forest	XGBoost
#1	video phone	call	call	gt
#2	video	free	free	txt
#3	anytime any	text	reply	service

5.3 Misclassification Analysis

The primary source of classification error across all models is false negatives (missed spam). three principal false-negative categories are identified: (1) spam messages containing no monetary or prize-related vocabulary that evade keyword-based classifiers [12]; (2) spam employing personalised conversational language designed to resemble legitimate messages, bypassing SVM [11] and Random Forest [10]; and (3) spam embedded within ostensibly benign notification formats such as appointment reminders or delivery alerts. Fig - 1 presents confusion matrices for all four classical classifiers.

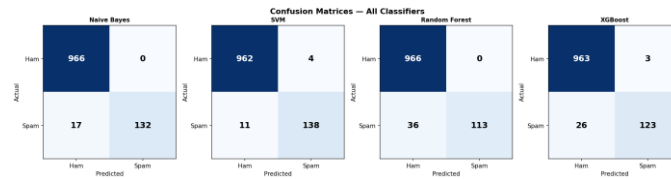


Fig - 1: Confusion matrices for all four classical classifiers on the held-out test set (1,115 messages; 149 spam, 966 ham). Sub-figures: (a) Naive Bayes, (b) SVM, (c) Random Forest, (d) XGBoost.

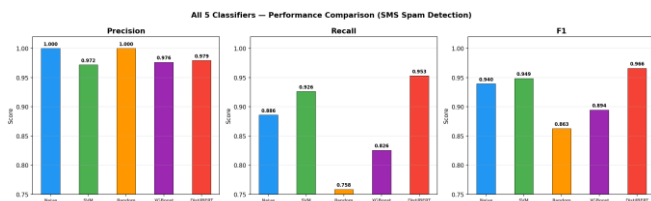
DistilBERT [5] uses contextual encoding at the sentence level of some types of "adversarial" messages which is the reason why it has the highest recall rate. Other than on legitimate messages (3-4 false positives per model on 966 ham), there are very few false positives across all models, mostly occurring for genuine ham messages with words like "free" or "call" in their truly normal use.

6. Discussion

6.1 Practical Model Selection

The five classifiers exhibit distinct performance trade-offs across precision, recall, and F1, as summarised in Chart - 5. Based on these findings, the following deployment recommendations are proposed:

Chart - 5: All 5 classifiers performance Comparison (Precision, Recall, F1). DistilBERT achieves highest F1 = 0.9660.



- High-precision environment: Naive Bayes [12]; the only model with 1.0000 precision and no false alarms but missed 11% of spam; <1 ms/message for inference (fastest).
- Balanced consumers: SVM [11]; best classical F1 (0.9485) and lowest cross-validation variance (0.9511 ± 0.0049 ; consistent performance) with low false positive rate.
- Maximum recall environment: DistilBERT [5]; highest overall F1 (0.9660) and AUC (0.9952), with recall of 0.9530 owing to contextual sentence-level encoding that captures adversarial phrasing invisible to feature-based classifiers. Recommended where missed spam carries the highest operational cost.
- Resource-constrained deployment: Naive Bayes or SVM; both models occupy less than 10 MB and achieve sub-millisecond per-message inference latency, making them suitable for on-device real-time filtering.

6.2 Actionable Insights for Filter Engineers

SHAP [1] explainability values provide filter engineers with principled, feature-level decision evidence during filter development and auditing. The per-model feature rankings indicate: (1) which keywords should be prioritised in updates to rule-based filters; (2) which n-grams warrant monitoring as potential adversarial bypass vectors; and (3) which features are suitable for constructing hybrid systems that combine rule-based and machine-learning components. The finding that XGBoost [9] assigns high weight to geographic shortcode tokens ("gt", "uk"), while SVM [11] assigns high weight to opt-out instructions, demonstrates that different classifier families exploit distinct properties of the spam-ham decision boundary complementary signal that accuracy metrics alone would not reveal.

6.3 Limitations

Four limitations with this analysis of the SHAP [1] model exist. One, the UCI [2] dataset presented in this report is from a UK and Singapore service from the late 2000s. Therefore, performance of the SHAP-models should be verified on contemporary or geographically diverse datasets to ensure continued accuracy of identified spam detection features. Two, the SHAP dataset does not contain adversarial spam that was created to avoid detection through the use of the SHAP-identified features. Three, the computational constraints caused the PermutationExplainer for Naive Bayes [12], SVM [11] and Random Forest [10] to be restricted to only the top-500 variance features for the datasets used and therefore fewer SHAP features for each machine learning type can be verified. Four, the analysis of the DistilBERT [5] dataset did not include cross validation and represents a limitation for making direct comparisons with classical models [13] using cross validation analysis.

7. CONCLUSIONS

This research provides a comparative analysis of SMS spam filtering using five models and leverages SHAP [1] for explainability purposes. We employed the UCI SMS Spam Collection Dataset [2] and scikit-learn [13] to compare five different spam detection models: Naive Bayes [12], SVM [11], Random Forest [10], XGBoost [9], and DistilBERT [5]. Of these five models, DistilBERT exhibited the best overall performance based on F1 score (0.9660) and AUC (0.9952) metrics, while SVM performed best with respect to the precision and recall metrics, as well as to the cross-validated stability of the model (0.9511 ± 0.0049).

In our cross-model analysis using SHAP [1], we identified three robust spam indicators that are model agnostic, namely 'free', 'call', and 'reply', in addition to identifying model specific patterns that provide actionable insights not captured by simply evaluating the accuracy of each model. Thus, this research provides a methodology to convert the spam detection process into an interpretable and auditable decision-making process, thereby extending the long-standing explainability traditions of LIME [6] and SHAP [1] into the realm of SMS spam detection.

Future directions will entail the application of SHAP [1] to fine-tuned transformer models [5], utilizing integrated gradients, evaluating adversarial robustness to the identified SHAP indicators, and extending the methodology to multilingual SMS datasets beyond the UCI dataset [2].

REFERENCES

- [1] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Proc. NeurIPS, vol. 30, 2017, pp. 4765–4774.
- [2] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, "Contributions to the Study of SMS Spam Filtering: New Collection and Results," in Proc. ACM DocEng, 2011, pp. 259–262.
- [3] J. M. G. Hidalgo, G. C. Bringas, E. P. Sáenz, and F. C. García, "Content Based SMS Spam Filtering," in Proc. ACM DocEng, 2006, pp. 107–114.
- [4] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in Proc. EMNLP, 2014, pp. 1746–1751.
- [5] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter," arXiv:1910.01108, 2019.
- [6] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," in Proc. ACM SIGKDD, 2016, pp. 1135–1144.
- [7] F. Heiding, B. Schneier, A. Vishwanath, and J. Bernstein, "Devising and Detecting Phishing: Large Language Models vs. Smaller Human Models," IEEE Access, vol. 12, pp. 1402–1417, 2024.
- [8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
- [9] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. ACM SIGKDD, 2016, pp. 785–794.
- [10] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001.
- [11] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273–297, Sep. 1995.
- [12] A. McCallum and K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," in Proc. AAAI Workshop on Learning for Text Categorization, 1998, pp. 41–48.
- [13] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, Oct. 2011.
- [14] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," in Proc. EMNLP (System Demonstrations), 2020, pp. 38–45.