Supervised Machine Learning for SMS Spam Detection

Fernando Acúrcio Silva
Instituto Politécnico do Cávado e do Ave (IPCA), Portugal
Barcelos, Portugal

***Abstract*—Short Message Service (SMS) spam remains a relevant cybersecurity problem, enabling unsolicited advertising, phishing attempts, and fraudulent activities. Due to the short and informal nature of SMS messages, effective automatic detection poses specific challenges when compared to other text-based communication channels. This paper investigates the application of supervised machine learning techniques for SMS spam detection using a publicly available dataset.**

A reproducible classification pipeline is adopted, including text preprocessing, Term Frequency–Inverse Document Frequency (TF-IDF) feature representation, and the evaluation of multiple supervised classifiers. The performance of the models is assessed using accuracy, precision, recall, F1-score, and ROC-AUC, with particular attention given to class imbalance.

The experimental results show that all evaluated models achieve high classification accuracy, confirming that SMS spam detection is a well-defined supervised learning problem. Linear Support Vector Machines achieve the best overall performance, reaching an accuracy of 0.982 and an F1-score of 0.925, while Logistic Regression and Random Forest classifiers also demonstrate competitive results with different precision–recall trade-offs. The findings highlight the importance of using multiple evaluation metrics, as accuracy alone does not fully capture classifier effectiveness in imbalanced datasets.

Overall, the results confirm that supervised learning approaches provide a robust and practical baseline for SMS spam detection and can serve as a foundation for more advanced filtering systems.

***Index Terms*—SMS Spam Detection, Supervised Learning, Text Classification, Cybersecurity, Machine Learning**

I. Introduction

The widespread use of mobile communication has made Short Message Service (SMS) an attractive channel for unsolicited and malicious content. SMS spam messages are commonly used for advertising, phishing attempts, and fraudulent activities, representing a persistent cybersecurity concern for both users and service providers. Unlike email messages, SMS texts are typically short, informal, and highly variable, which increases the difficulty of applying traditional text filtering techniques effectively.

From a machine learning perspective, SMS spam detection can be formulated as a text categorisation problem, where messages are automatically classified into predefined categories, such as legitimate (ham) or spam. This formulation follows the general definition of automated text categorisation, where documents are assigned to one or more classes based on their content using statistical or machine learning approaches [1].

Several studies have explored the application of supervised learning techniques to SMS spam detection, showing that machine learning models can effectively learn patterns from labelled SMS datasets. Almeida et al. introduced a publicly available SMS spam dataset and evaluated multiple supervised classifiers, demonstrating that data-driven approaches are well suited for this problem [2].

In addition to establishing benchmark datasets, further research has focused on improving the evaluation and comparison of SMS spam detection models. Mohasseb et al. analysed supervised classification methods for SMS spam identification and emphasised the importance of appropriate evaluation metrics, particularly in the presence of class imbalance [3].

Earlier work by Almeida et al. also contributed to the field by analysing new SMS spam collections and validating the effectiveness of supervised classifiers on real-world data [4].

Motivated by these studies, this paper investigates the application of supervised machine learning techniques for SMS spam detection using a publicly available dataset. Multiple classifiers are trained and evaluated under consistent conditions in order to provide a clear and reproducible comparison. The objective is not only to assess classification performance, but also to analyse the strengths and limitations of supervised approaches when applied to short and imbalanced text data, such as SMS messages.

II. Related Work

Automated text categorisation has been extensively studied as a supervised learning problem, where documents are represented through textual features and assigned to predefined classes. Sebastiani provides a comprehensive survey of machine learning techniques for text categorisation, describing the typical pipeline involving text representation, feature extraction, model training, and evaluation [1].

Within this broader context, SMS spam detection can be seen as a specialised case of text categorisation, characterised by very short documents, informal language, and limited contextual information. Almeida et al. addressed these challenges by introducing a publicly available SMS spam dataset and evaluating multiple supervised classifiers on real SMS messages [2]. Their results demonstrated that traditional machine learning algorithms, when applied with appropriate text representations, can achieve effective spam detection performance.

Further contributions by Almeida et al. focused on extending SMS spam collections and validating classification results

using new datasets [4]. This work highlighted the importance of the quality and representativeness of the dataset when evaluating SMS spam filtering techniques, as well as the need for consistent experimental setups to enable fair comparison between classifiers.

More recent work has expanded the analysis of SMS spam detection by incorporating detailed evaluation strategies and addressing practical issues such as class imbalance. Mohasseb et al. proposed a supervised learning approach for SMS spam identification and emphasised the role of appropriate performance metrics beyond simple accuracy, particularly in imbalanced datasets [3]. Their study reinforces the importance of precision, recall, and related metrics when assessing the effectiveness of spam detection systems.

Building on these studies, the present work adopts established supervised learning techniques and evaluation practices for SMS spam detection. Following a reproducible pipeline aligned with prior literature, this study aims to provide a clear comparison of supervised classifiers applied to a well-known SMS spam dataset, reinforcing existing findings while offering a structured and transparent experimental analysis.

III. Methodology and Machine Learning Models

This section describes the methodology adopted for SMS spam detection, including the dataset used, the data preprocessing steps, the feature representation, the supervised learning models, and the evaluation strategy. The overall workflow follows a standard supervised text classification pipeline, ensuring reproducibility and comparability with prior work.

A. Dataset Description

The research is conducted using a publicly available SMS spam dataset, which consists of a collection of SMS messages labelled as legitimate (ham) messages or spam. Each message is represented as raw text along with its corresponding class label. As commonly observed in SMS spam datasets, the class distribution is imbalanced, with legitimate messages forming the majority of the data, in this dataset, the spam messages represent less than 15%. This characteristic reflects real-world conditions and motivates the use of evaluation metrics beyond simple accuracy.

B. Data Preprocessing

Before model training, SMS messages in the dataset need to undergo a series of preprocessing steps to transform the raw text into a suitable format for machine learning algorithms. These steps include text normalisation, removal of irrelevant characters, and tokenisation of the SMS content. The preprocessing process aims to reduce noise while preserving discriminative textual information relevant to spam detection.

The resulting cleaned messages are then used to construct a new structured dataset, enabling consistent feature extraction and model training across all experiments. The preprocessing pipeline is applied uniformly to both training and testing data in order to avoid data leakage and to ensure fair model evaluation.

C. Feature Representation

In order to apply supervised learning algorithms, the preprocessed SMS messages are transformed into numerical feature vectors using the Term Frequency–Inverse Document Frequency (TF-IDF) representation. TF-IDF assigns a weight to each term based on its frequency within a message and its distribution across the entire dataset, this way emphasising terms that are more discriminative for classification.

This representation is particularly suitable for SMS spam detection, as it captures relevant word usage patterns while reducing the influence of very common terms that carry limited semantic value. TF-IDF has been widely adopted in text categorisation tasks and has demonstrated effective performance in SMS spam filtering when combined with supervised classifiers, especially linear models.

The resulting TF-IDF vectors form a high-dimensional feature space, which is used consistently across all evaluated classifiers to ensure a fair and meaningful comparison of model performance.

D. Supervised Learning Models

Multiple supervised machine learning algorithms are employed to perform binary classification between spam and ham messages. A baseline classifier is first trained to establish a performance baseline. Subsequently, more advanced supervised models are applied in order to assess potential improvements over the baseline.

Among the evaluated models, Support Vector Machines (SVM) are included due to their strong performance in high-dimensional text classification tasks. The hyperparameters of the model are tuned using a systematic search strategy to optimise classification performance. All models are trained using the same splits in the dataset and feature representations to ensure a fair comparison between each model.

E. Evaluation Strategy

The performance of the supervised models is evaluated using a held-out test set that is not seen during training. Given the imbalanced nature of the dataset, multiple evaluation metrics are considered, including precision, recall, and F1-score, in addition to accuracy. These metrics provide a more informative assessment of the model performance, particularly with respect to the detection of spam messages.

The evaluation framework is applied consistently across all models, enabling a direct and meaningful comparison of their strengths and limitations in the context of SMS spam detection.

IV. Results and Discussion

This section presents and discusses the experimental results obtained from the supervised learning models applied to the SMS spam detection task. The models are evaluated using a consistent experimental setup and a common feature representation in order to enable a fair comparison.

A. Classification Performance

The experimental results show that supervised learning models are effective in distinguishing between spam and legitimate SMS messages. The baseline classifier establishes a reference level of performance, confirming that meaningful patterns exist in the dataset that can be exploited by machine learning approaches. More advanced models achieve improved results across most evaluation metrics, indicating their ability to better capture discriminative textual features.

Support Vector Machines demonstrate strong overall performance, particularly in terms of precision and F1-score. This behaviour is consistent with expectations for high-dimensional text data represented using TF-IDF features, where linear decision boundaries are often effective. The observed results align with findings reported in prior studies on SMS spam detection, which highlight the suitability of SVM-based classifiers for this problem domain [1].

The evaluated models include Linear Support Vector Machines (Linear SVM), Logistic Regression, Random Forest, and Multinomial Naive Bayes.

TABLE I
PERFORMANCE COMPARISON OF SUPERVISED MODELS FOR SMS SPAM DETECTION

Model	Acc.	Prec.	Rec.	F1	ROC-AUC
Linear SVM	0.982	0.952	0.901	0.925	–
Logistic Reg.	0.977	0.908	0.908	0.908	0.991
Random Forest	0.975	0.991	0.809	0.891	0.994
Multinomia NB	0.970	1.000	0.763	0.866	0.979

Table I summarises the classification performance of the evaluated supervised learning models. All models achieve high accuracy, indicating that SMS spam detection is a well-defined classification problem when appropriate text representations are used.

The Linear SVM classifier achieves the highest overall accuracy and F1-score, demonstrating a strong balance between precision and recall. Logistic Regression shows comparable performance, with slightly lower accuracy but a more balanced precision–recall trade-off. Random Forest achieves the highest precision but exhibits a lower recall, indicating a tendency to miss a larger proportion of spam messages. Multinomial Naive Bayes also demonstrates high precision; however, its lower recall suggests reduced effectiveness in detecting all spam instances.

These results highlight the importance of considering multiple evaluation metrics, particularly in imbalanced datasets, as accuracy alone does not fully reflect classifier performance.

B. Impact of Class Imbalance

As the SMS spam dataset exhibits a clear class imbalance, evaluation metrics beyond accuracy are essential for a meaningful assessment of model performance [1]. While accuracy values remain high across different models, precision and recall provide a more informative perspective on the ability to correctly identify spam messages without misclassifying legitimate ones.

The results indicate that some classifiers prioritise recall, successfully detecting a larger portion of spam messages at the cost of increased false positives. Other models achieve higher precision, reducing false alarms but potentially missing a subset of spam messages. This trade-off is particularly relevant in practical deployment scenarios, where the cost of misclassifying legitimate messages may outweigh the benefit of detecting all spam instances.

C. Comparison with Related Work

The results obtained are consistent with those reported in previous SMS spam filtering studies. Almeida et al [2]. demonstrated that traditional supervised classifiers can achieve reliable performance when applied to curated SMS datasets, while Mohasseb et al [3]. emphasised the importance of evaluation metrics that account for class imbalance. The behaviour observed in this study reinforces these conclusions, confirming that model performance should be assessed using a combination of precision, recall, and F1-score rather than accuracy alone.

By adopting a reproducible pipeline and a consistent evaluation strategy, this work validates existing findings while providing a transparent comparison of supervised learning approaches applied to SMS spam detection.

D. Discussion and Limitations

Although the evaluated models achieve strong classification performance, several limitations must be considered. The experiments are conducted using a single publicly available dataset, which may not fully capture the diversity of SMS spam messages encountered in real-world environments. In addition, the dataset contains a relatively limited number of spam messages, with fewer than eight hundred spam samples available for training and evaluation. This restricted sample size, combined with class imbalance, may limit the generalisation capability of the trained models when exposed to previously unseen spam patterns.

Furthermore, the use of traditional text representation techniques, such as TF-IDF, does not explicitly capture contextual or semantic relationships between words beyond their statistical co-occurrence. While this approach is effective and computationally efficient, it may overlook more complex linguistic patterns present in SMS messages.

Despite these limitations, the results demonstrate that supervised learning techniques remain a robust and practical solution for SMS spam detection. The simplicity, interpretability, and effectiveness of the evaluated models make them suitable as baseline approaches and as components of more advanced spam filtering systems.

V. CONCLUSION

This paper investigated the application of supervised machine learning techniques for SMS spam detection using a publicly available dataset. A reproducible experimental

pipeline was adopted, that include text preprocessing, TF-IDF feature representation, and the evaluation of multiple supervised classifiers under consistent conditions. The results demonstrate that supervised learning approaches are effective in distinguishing between legitimate and spam SMS messages when appropriate text representations are employed.

Among the evaluated models, linear classifiers, particularly Support Vector Machines, achieved the strongest overall performance, providing a favourable balance between precision and recall. The comparative analysis also highlighted important trade-offs between different classifiers, especially in the presence of class imbalance, reinforcing the need to consider multiple evaluation metrics rather than accuracy alone.

Despite the strong performance observed, the study is subject to limitations related to dataset size and diversity, as well as the use of traditional text representation techniques that do not capture deeper semantic relationships. Future work may explore the use of larger and more diverse datasets, as well as more advanced representation methods, to further improve robustness and generalisation.

In the end, the findings confirm that supervised machine learning models constitute a robust and practical baseline for SMS spam detection and provide a solid foundation for the development of more advanced spam filtering systems.

REFERENCES

[1] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002. [Online]. Available: https://arxiv.org/abs/cs/0110053

[2] T. A. Almeida, J. M. G. Hidalgo, and T. P. Silva, "Towards sms spam filtering: Results under a new dataset," *International Journal of Information Security Science*, vol. 2, no. 1, pp. 1–18, 2011. [Online]. Available: https://www.researchgate.net/publication/258514273_Towards_SMS_Spam_Filtering_Results_under_a_New_Dataset

[3] A. Mohasseb, B. Aziz, and A. Kanavos, "Sms spam identification and risk assessment evaluations," in *Proceedings of the 16th International Conference on Web Information Systems and Technologies (WEBIST 2020)*. SCITEPRESS, 2020, pp. 417–424. [Online]. Available: https://www.scitepress.org/Link.aspx?doi=10.5220/0010022404170424

[4] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, "Contributions to the study of sms spam filtering: New collection and results," in *Proceedings of the 11th ACM Symposium on Document Engineering (DOCENG 2011)*. ACM, 2011, pp. 259–262.