Machine Learning-Based Detection of SMS Phishing in the Philippine Context

John Exequiel A. Corpuz[1], Sebastian Q. Diaz[1] Luis Benedict M. Palafox[1], Mark Christian T. Tan[1], and Katrina Ysabel C. Solomon*[1]

Advanced Research Institute for Informatics, Computing, and Networking,
De La Salle University,
2401 Taft Avenue, Manila 1004, Philippines
katrina.solomon@dlsu.edu.ph

Abstract. Short Message Service (SMS) phishing, or smishing, is an escalating cybersecurity threat in the Philippines, where widespread mobile usage intersects with informal language and frequent code-switching between Filipino and English. While prior studies have primarily focused on English-language datasets, limited research exists that directly addresses the linguistic complexities and cultural nuances unique to the Filipino context. This study aims to bridge this gap by developing a machine learning-based detection system optimized for smishing in the Philippines. A labeled dataset comprising both phishing and legitimate messages was collected from public sources and surveys, then preprocessed using natural language processing techniques such as tokenization, lemmatization, and TF-IDF vectorization. The study implemented and evaluated five classical machine learning classifiers: Support Vector Machines (SVM), Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), and Multinomial Naive Bayes (MNB). Among these, the SVM model combined with the TF-IDF features achieved the highest performance, recording an F-score of 99.20%, indicating robust precision and recall. These findings affirm the effectiveness of language-aware, content-based smishing detection tailored to the Filipino language, offering a foundational contribution to the development of inclusive, culturally adaptive cybersecurity systems and highlighting the importance of extending protection beyond Anglocentric models.

Keywords: SMS, Smishing, Machine Learning, Information Security, Phishing

1 Introduction

A recent study reveals that most cyberattacks in recent years have emanated from SMS phishing, or smishing, posing a significant threat to individuals and organizations. Attackers employ deceptive text messages impersonating legitimate entities, attempting to extract sensitive information or lure victims into clicking malicious links. Unlike email phishing, smishing takes advantage

J. Caro et al. (eds.), *Proceedings of the Workshop on Computation: Theory and Practice (WCTP 2025)*, Atlantis Highlights in Computer Sciences 24,
https://doi.org/10.2991/978-94-6239-638-8_28

of the immediacy and personal nature of SMS, which often leads to higher user engagement and success rates [1]. Smishing in the Philippine context is further complicated by the nature of local communication. Filipino messages are characterized by non-standard spelling, informal syntax, frequent abbreviation, and code-switching between English and Filipino [8] [10] . Phishing actors exploit these linguistic and cultural features—embedding local idioms, terms, and social cues—to make messages appear credible [13].

As mobile connectivity becomes ubiquitous, so too does the attack surface. In the first quarter of 2024 alone, Smart Communications blocked over 13 million smishing messages [12], while a 2023 Statista report revealed that 43% of Filipino consumers claimed to have received phishing SMS messages [14]. These figures underscore the urgency of building a culturally and linguistically aware smishing detection system.

Natural Language Processing (NLP) has become instrumental in detecting phishing attempts by analyzing text for suspicious patterns. Techniques such as tokenization, stemming, and lemmatization help convert unstructured SMS content into machine readable features [6] [7] . However, standard NLP methods trained on English text fail to capture the nuances of code-switched or informal Filipino SMS [9] [15] . Thus, tailoring language models to local contexts is essential.

Machine learning techniques have shown promising results in phishing detection. Algorithms like Support Vector Machines (SVM), Random Forest, Logistic Regression, K-Nearest Neighbors, and Naive Bayes have been widely used due to their effectiveness in binary classification problems [1] [7]. By training on labeled data and using carefully engineered linguistic features, such models can learn to differentiate between phishing and legitimate SMS messages.

This study develops a machine learning-based approach to smishing detection tailored for the Philippine context. A labeled dataset of Filipino and code-switched SMS messages was curated from public sources and user surveys. The messages were preprocessed using NLP pipelines, then transformed via TF-IDF vectorization and RoBERTa-tagalog-base. RoBERTa-tagalog-base is a pre-trained transformer model specifically trained on large-scale Filipino text data [5]. Five machine learning models were trained and evaluated to determine their effectiveness. The results aim to contribute to culturally adaptive phishing detection methods and fill the gap in Philippine-specific cybersecurity solutions.

2 Related Works

Hybrid models that integrate Naive Bayes with other classifiers have shown promise in detecting SMS fraud in linguistically diverse contexts [1]. Simple probabilistic models like Naive Bayes can be adapted effectively to multilingual data when paired with language-specific preprocessing techniques [10]. Deep learning approaches have also proven effective for short-text spam filtering, using neural networks to capture SMS semantics [8]. Precision in SMS phishing detection can be improved through correlation-based feature selection, and

CNN-LSTM hybrid models have demonstrated high performance in multilingual spam detection tasks [12] [13].

Machine learning frameworks using NLP techniques and classifiers have been successfully applied to Filipino spam messages, emphasizing the need to process informal grammar and local vocabulary [14]. Filipino texting behavior—often characterized by abbreviations, hybrid syntax, and informal phrasing—poses additional challenges for traditional NLP pipelines [9]. This is further supported by studies on code-switching, which highlight the need for models that can handle mixed-language tokens common in Philippine SMS communication [15].

Phishing detection strategies in other domains, such as email and URLs, have contributed valuable techniques including hybrid content-header analysis and feature extraction from structured links [4] [11] . Dimensionality reduction techniques have also impacted spam detection model performance, particularly in probabilistic models like Naive Bayes [3]. Meanwhile, phishing has been described as an evolving social engineering threat requiring localized and culturally informed mitigation strategies [2].

Recent smishing statistics in the Philippines emphasize the urgency of the issue, with Smart Communications reporting over 13 million blocked smishing messages in Q1 2024, and Statista data showing that 43% of Filipino users encountered such messages in late 2023 [12] [14]. This underscores the relevance of research focused on localized, language-aware smishing detection systems.

3 Methodology

This study implemented a machine learning-based approach for detecting SMS phishing (smishing) tailored to the Philippine context, focusing on code-switched and informal Filipino messages. The methodology is composed of four key stages: data collection, preprocessing and feature extraction, model training, and evaluation.

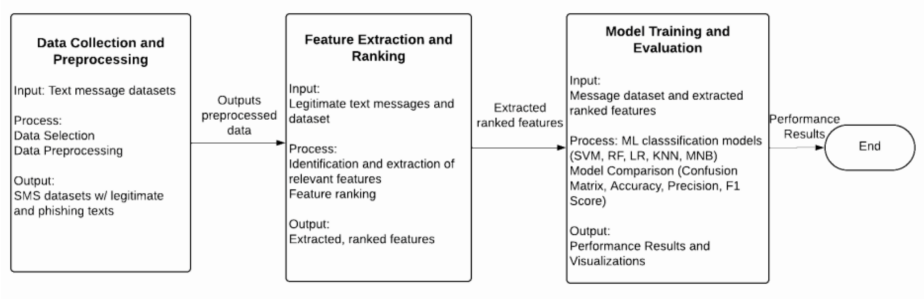

Fig. 1. System Architecture

3.1 Data Collection

To ensure a dataset that captures both general smishing trends and local linguistic patterns, this study combined data from two primary sources. The first is the PH Spam + Marketing SMS dataset from Kaggle, containing approximately 1,475 pre-labeled messages from the Philippine telco environment. The second involved a locally administered campaign that gathered 4,752 SMS messages through voluntary user contributions, for a combined total of 6,227 messages.

Each message was independently annotated by ten annotators, with final labels (SPAM or HAM) determined via majority consensus. No ties were recorded. This process helped mitigate individual bias and supported consistent labeling across code-switched and informal texts, which are common in Filipino SMS. In the final dataset, 2,318 messages were labeled SPAM and 3,910 as HAM, naturally reflecting real-world SMS class imbalance.

3.2 Data Preprocessing and Feature Extraction

Messages underwent multiple transformation steps including lowercasing, removal of special characters, links, emojis, and redundant punctuation. Financial and promotional keywords were normalized into tokens like "money" to reduce lexical variability. Tokenization was performed using NLTK, while bilingual stop word removal (English and Filipino) and stemming with the Porter Stemmer helped reduce linguistic noise.

For feature extraction, TF-IDF vectorization captured unigrams and bigrams using a filtered vocabulary (min_df=2, max_df=0.95), yielding 5,983 features. Sender information, categorized by top 10 frequent senders and grouped otherwise under "OTHER," was one-hot encoded and appended as metadata. These were horizontally stacked with the TF-IDF vectors to form the final feature matrix.

RoBERTa-tagalog-base embeddings were also explored as they are effective particularly in code-switching and informal messaging contexts in Filipino-language [5]. Cleaned messages were tokenized and passed through the model, with the [CLS] token used as the semantic representation. Sender encodings were concatenated with these dense embeddings to form a second feature matrix used for alternative model evaluation. Due to resource constraints, final training and evaluation focused on the TF-IDF-based pipeline.

3.3 Model Selection and Training

Five machine learning models were implemented and evaluated for smishing detection: Support Vector Machines (SVM), Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbors (KNN), and Multinomial Naive Bayes (MNB). All models were trained using the same feature set—TF-IDF vectors combined with encoded sender information—and implemented in Python using NumPy, scikit-learn, and supporting libraries.

SVM was configured with both linear and RBF kernels, with hyperparameter tuning on C and gamma using grid search. Logistic Regression used the liblinear solver and tested both l1 and l2 regularization to reduce overfitting. Random Forest models were tuned for n_estimators, max_depth, and min_samples_split, and also provided interpretable feature importance scores. KNN tested different values of k and distance metrics (e.g., cosine), with both uniform and distance-weighted voting. MNB, selected for its efficiency on sparse data, optimized alpha (Laplace smoothing) and fit_prior.

The dataset was split using stratified train-test splitting (70-30) to preserve class balance. Performance metrics—accuracy, precision, recall, F-score, and AUC—were computed on the test set and validated using 5-fold cross-validation.

While contextual embeddings from RoBERTa-tagalog-base were initially tested (with [CLS] token vectors), final evaluation and tuning focused on TF-IDF pipelines due to their computational efficiency and strong empirical performance.

3.4 Evaluation Metrics and Validation

Model performance was measured using accuracy, precision, recall, F-score, and area under the curve (AUC). Five-fold stratified cross-validation was employed to ensure fair assessment across classes. Confusion matrices were generated to visualize correct vs incorrect classifications.

4 Results and Analysis

This section presents the outcomes of the experimental evaluation of various machine learning models applied to the SMS phishing detection dataset. The analysis is organized into two primary segments: Exploratory Data Analysis (EDA) and Model Performance Analysis. EDA provides insight into the linguistic structure, sender behavior, and distributional characteristics of the dataset, while the performance analysis evaluates how well different models generalize in detecting phishing messages, using standard classification metrics. These two perspectives offer a comprehensive understanding of both the data properties and the system's effectiveness.

The dataset comprises a total of 6,227 SMS messages, with 2,318 labeled as SPAM and 3,909 as HAM. Each message is paired with metadata including the sender's identity and timestamp of receipt. The data was compiled from a Kaggle-hosted public dataset and a locally conducted collection effort, ensuring both generalizability and contextual relevance. All messages underwent an annotation process involving ten independent annotators, with labels finalized through majority consensus. No missing values were identified in the final dataset, allowing for streamlined preprocessing during the data cleaning phase. The structured nature of the dataset, including sender fields and timestamp information, facilitated not only text classification but also exploratory analysis involving behavioral patterns, such as frequency of spam-associated sender IDs and message content variations over time.

Table 1. Dataset Columns

SMS Message	Raw textual content of the SMS message provided.
Sender Identifier	Classification of the message as either "SPAM" (phishing) or "HAM" (legitimate).
Class Label	Information pertaining to the origin of the SMS message, which could be a numerical phone number or registered name in NTC.

Early analysis revealed distinct lexical and structural differences between message types. SPAM messages frequently featured shortened URLs, promotional phrases, and direct call-to-action instructions, often using persuasive or urgent tones. In contrast, HAM messages displayed more varied sentence structures and often originated from verified service providers, personal contacts, or transaction-based systems. Linguistically, many messages featured informal syntax, abbreviations, and extensive code-switching between Filipino and English. These hybrid linguistic forms presented challenges for traditional natural language processing and influenced preprocessing choices.

Following preprocessing and feature extraction using both TF-IDF and RoBERTa-tagalog-base embeddings, five classical classifiers: Support Vector Machine (SVM), Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), and Multinomial Naive Bayes (MNB)—were trained and evaluated. Each model underwent stratified five-fold cross-validation, and performance was assessed using four standard metrics: accuracy, precision, recall, and F-score. Based on the results, TF-IDF features combined with sender encoding consistently delivered the most robust performance across classifiers.

Table 2. Performance Using TF-IDF Vectorizer

Classifier	Accuracy	Precision	Recall	F-Score
SVM	**99.25%**	**99.23%**	**99.17%**	**99.20%**
Logistic Regression	99.14%	99.11%	99.06%	99.08%
Random Forest	98.50%	98.48%	98.31%	98.39%
KNN	97.32%	97.59%	96.70%	97.11%
MNB	98.98%	98.84%	98.99%	98.91%

Presented in Table 2, the SVM classifier emerged as the top-performing model under the TF-IDF pipeline, achieving 99.25% accuracy, 99.23% precision, 99.17% recall, and an F-score of 99.20%. Logistic Regression followed closely, with respective scores of 99.14%, 99.11%, 99.06%, and 99.08%. MNB, despite its simplicity, maintained competitive performance, reaching 98.98% accuracy and a balanced F-score of 98.91%. Random Forest achieved good results (98.50% accuracy, 98.48% precision, 98.31% recall, and 98.39% F-score), though it showed

slightly lower recall, suggesting reduced sensitivity in detecting SPAM messages. KNN, on the other hand, posted 97.32% accuracy, with strong precision at 97.59% but comparatively lower recall at 96.70%, indicating occasional misclassifications of SPAM messages.

Table 3. Performance Using Tagalog-based RoBERTa Transformer

Classifier	Accuracy	Precision	Recall	F-Score
SVM	**99.14%**	**99.11%**	**99.06%**	**99.08%**
Logistic Regression	98.66%	98.56%	98.58%	98.57%
Random Forest	97.59%	97.44%	97.41%	97.42%
KNN	98.56%	98.47%	98.44%	98.45%
MNB	85.66%	84.51%	85.25%	84.83%

RoBERTa-tagalog-base embeddings were also evaluated using the same classifiers as seen in Table 3. While the SVM still performed well (99.14% accuracy and 99.08% F-score), results across other classifiers were less consistent compared to TF-IDF. Notably, MNB's performance declined significantly under RoBERTa, achieving only 85.66% accuracy and an F-score of 84.83%. This decline is attributed to the incompatibility between MNB and the dense, contextual embeddings generated by RoBERTa, which do not resemble the word frequency distributions MNB is optimized for. Additionally, the study notes that models using RoBERTa were subject to greater processing overhead and resource constraints, which limited their use in final evaluations.

TF-IDF paired with linear classifiers such as SVM and Logistic Regression proved more effective and reliable, with clearer interpretability and more consistent results across evaluation folds. These findings validate the suitability of keyword-based lexical models for Filipino SMS phishing detection, particularly when sender metadata is incorporated.

5 Conclusion and Recommendations

This study presented a machine learning-based approach to SMS phishing detection tailored to the Filipino context. The dataset used in this study combined publicly available messages and locally sourced contributions, yielding 6,227 labeled messages consisting of both HAM and SPAM written in English, Filipino, and code-switched formats. After thorough preprocessing and feature extraction using TF-IDF and sender information, five classical classifiers were evaluated. Among them, the Support Vector Machine (SVM) classifier achieved the best overall performance, followed by Logistic Regression and Naive Bayes.

The results demonstrate the viability of using traditional machine learning models in detecting smishing messages within the Philippine setting. The use of bilingual preprocessing steps and inclusion of sender metadata contributed to improved detection accuracy. These findings support the development

of localized anti-smishing tools that are both effective and efficient in resource-constrained environments.

While the study addressed Taglish and incorporated Filipino preprocessing, deep or regional dialects (e.g., Cebuano, Ilocano) and sociolects remain under-represented. Although the NLP techniques employed were designed to address code-switching and general linguistic nuances of Filipino, the vast linguistic diversity of the Philippines, encompassing numerous regional dialects beyond standard Filipino (Tagalog-based), presents a complex challenge. The models' ability to accurately interpret and detect phishing in less common or highly localized dialects might be constrained by the training data's representation of these linguistic variations. This limitation suggests that while the system performs well for widely understood Filipino, its efficacy might decrease when confronted with messages heavily reliant on specific regional linguistic features or deeply embedded cultural references not present in the training corpus.

The restriction to classical ML methods, specifically Support Vector Machines, Random Forest, and Logistic Regression, represents a methodological limitation. Transformer-based models (e.g., RoBERTa) were evaluated but proved less effective on sparse SMS data and MNB's performance notably degraded due to its count-based assumptions. While these algorithms proved effective and provided a solid comparative analysis, the study did not explore more advanced machine learning paradigms, such as deep learning models (e.g., Recurrent Neural Networks or Transformers), which are often more adept at capturing complex sequential patterns and long-range dependencies in text data. The decision to focus on classical ML was partly due to computational resource constraints and the scope of a comparative analysis. However, this means that the full potential for higher accuracy and more nuanced linguistic understanding, which deep learning models might offer, remains unexplored within this research. Future work could leverage these advanced techniques to potentially overcome some of the current linguistic challenges.

Future work may include conducting region-based data collection campaigns to include under-sampled dialects and demographics (e.g., senior citizens, rural communities). This would also involve active collaboration with telecommunication companies, cybersecurity organizations, and local communities to gather real-world, anonymized data that encompasses a broader spectrum of Filipino dialects, slang, and code-switched messages. A more comprehensive dataset will significantly enhance the generalization and robustness of machine learning models, allowing them to better adapt to the dynamic nature of phishing tactics and the rich linguistic diversity of the Philippines.

Another suggested future work is to develop and pilot a lightweight, on-device plugin or telco-level filter that leverages the trained TF-IDF/SVM pipeline. Real-world testing in controlled environments such as university labs or volunteer cohorts that will surface deployment challenges (latency, battery usage) and user-experience feedback. The current study provides a

foundational understanding of effective detection algorithms; however, for practical application, these models need to be seamlessly integrated into platforms that can offer immediate protection to users. This could involve creating a standalone mobile application that scans incoming SMS messages or, more effectively, collaborating with telecommunication providers to implement detection systems directly within their network infrastructure. Real-time scanning capabilities would allow for instant identification and blocking of smishing attempts, significantly reducing the window of vulnerability for users.

Given the linguistic diversity of the Philippines, a system that can detect phishing across various indigenous languages and dialects would provide more comprehensive protection. This would involve dedicated research into the unique linguistic features of each major dialect, developing tailored NLP models, and potentially leveraging transfer learning techniques from the current Filipino model. Such an expansion would ensure that all linguistic communities within the Philippines are adequately protected from smishing threats. This may be achieved by adapting the framework to incorporate multilingual embeddings or dialect-aware tokenizers and explore hybrid architectures that blend lexical and contextual features. This could improve detection of phishers' use of localized idioms and code-mixed messages beyond Taglish.

Additional exploration may also be performed on transformer-light models (e.g., MobileBERT, DistilBERT) fine-tuned on Filipino SMS data, as well as convolutional or recurrent neural networks optimized for brevity. Comparative trials against the TF-IDF baseline will clarify trade-offs between performance, interpretability, and resource requirements.

By pursuing these directions, future work can strengthen the system's generalization, resilience, and real-world effectiveness—ultimately enhancing the cybersecurity posture of Filipino mobile users against rapidly evolving smishing threats.

References

1. Agrawal, N., Bajpai, A., Dubey, K., and Patro, B. An effective approach to classify fraud sms using hybrid machine learning models. In *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)* (2023), pp. 1–6.
2. Alkhalil, Z., Hewage, C., Nawaf, L., and Khan, I. Phishing attacks: A recent comprehensive study and a new anatomy. *Front. Comput. Sci. 3* (Mar. 2021).
3. Almeida, T. A., Almeida, J., and Yamakami, A. Spam filtering: how the dimensionality reduction affects the accuracy of naive bayes classifiers. *J. Internet Serv. Appl. 1*, 3 (Feb. 2011), 183–200.
4. Asif, A. U. Z., Shirazi, H., and Ray, I. Machine learning-based phishing detection using url features: A comprehensive review. In *Stabilization, Safety, and Security of Distributed Systems* (Cham, 2023), S. Dolev and B. Schieber, Eds., Springer Nature Switzerland, pp. 481–497.
5. Cruz, J. C. B., and Cheng, C. Improving large-scale language models and resources for filipino. *arXiv preprint arXiv:2111.06053* (2021).
6. de Luna, R. G., Magnaye, V. C., L. Reaño, R. A., Enriquez, K. L., Astorga, D., Celestial, T., Española, A. M., Lanting, B. A., Mugar, D.,

Ramos, M., and Redondo, J. A machine learning approach for efficient spam detection in short messaging system (sms). In *TENCON 2023 - 2023 IEEE Region 10 Conference (TENCON)* (2023), pp. 53–58.
7. Ghourabi, A., Mahmood, M. A., and Alzubi, Q. M. A hybrid cnn-lstm model for sms spam detection in arabic and english messages. *Future Internet 12*, 9 (2020), 156.
8. Gomaa, W. The impact of deep learning techniques on sms spam filtering. *International Journal of Advanced Computer Science and Applications 11* (01 2020).
9. Ignacio, A., and De Jesus, F. Texting and chatting styles of grade 11 students : A case in the philippines. *International Journal of English Literature and Social Sciences 6* (01 2021), 096–109.
10. K, A., and Halder, S. Detection of multilingual spam sms using naïvebayes classifier. In *2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)* (2023), pp. 89–94.
11. Maturure, P., Ali, A., and Gegov, A. "hybrid machine learning model for phishing detection". In *2024 IEEE 12th International Conference on Intelligent Systems (IS)* (2024), pp. 1–7.
12. Rosales, E. F. Smart blocks 13 million smishing messages, Apr 2024.
13. Sonowal, G. Detecting phishing sms based on multiple correlation algorithms. *SN Comput. Sci. 1*, 6 (Nov. 2020).
14. Statista. Share of consumers in the philippines targeted by smishing in q4 2023, Jul 2024.
15. Villanueva, L., and Bert, G. Analysis on code switching manifested by filipino high school teachers. *Diversitas Journal 8* (07 2023).