

Arabic SMS Spam filter Using a Random Forest Model Based on Machine Learning

Shahd Khairou Dairi

Table of Content

Abstract	3
Introduction	4
background	4
Problem statement	5
Research objectives	5
Research scope	6
Literature review	7
Research methodology	11
Expected outcomes	12
Gantt chart	13
references	14

Abstract:

This study addresses the problem of Arabic SMS spam, which poses significant risks to user privacy, security, and device efficiency. Due to the increasing reliance on mobile communication, spam messages such as advertisements, phishing attempts, and fraudulent content have become widespread, making effective filtering systems essential. However, existing spam detection models are often designed for English text and perform poorly when applied to Arabic SMS due to its rich morphology and complex linguistic structure.

The main objective of this research is to develop an efficient machine learning-based model, specifically using the Random Forest algorithm, to accurately classify Arabic SMS messages into spam and ham categories based on their textual content and characteristics. The methodology follows a supervised learning approach, involving data collection from a labeled SMS dataset, followed by extensive preprocessing steps such as text cleaning, tokenization, normalization, stop-word removal, and feature extraction. The dataset is then split into training (80%) and testing (20%) sets to evaluate model performance. The Random Forest model is trained on processed data and assessed using evaluation metrics including accuracy, precision, recall, and F1-score.

The expected outcome of the study is that the Random Forest model will achieve high classification performance and demonstrate strong robustness in handling Arabic text complexity and that the proposed approach can effectively reduce unwanted spam messages.

Chapter 1:

Introduction

The rapid advancement of mobile communication technologies has made Short Message Service (SMS) one of the most widely used methods for information exchange, the volume of messages exchanged daily has grown significantly. However, this growth has also led to a rise in unsolicited and harmful messages, commonly known as spam.

Spam messages not only threaten users by attempting to extract sensitive information but also create inconvenience by cluttering inboxes and consuming device resources such as storage and processing power. As a result, users are required to spend additional time and effort identifying and filtering unwanted messages from legitimate ones. This highlights the need for an efficient and automated Arabic SMS spam detection system.

Machine learning techniques have emerged as effective solutions for SMS spam detection such as Random Forest, Naïve Bayes and support vector machines have shown strong performance in identifying spam versus legitimate messages and automatically classify incoming messages.

By analyzing features such as keywords, message structure, and writing patterns, these models can distinguish between spam and non-spam (ham) messages with high accuracy. However, most existing spam detection systems are developed for the English language and are not well-suited for handling Arabic text.

Project Background:

SMS Spam detection has been widely studied using various machine learning approaches. These approaches rely on supervised learning, where models are trained on labeled datasets to recognize patterns associated with spam messages. Common characteristics of spam include the use of promotional keywords, excessive punctuation, unusual capitalization, and requests for personal or financial information.

Despite significant progress in this field, the Arabic language presents unique challenges for spam detection due to its rich morphology, complex grammar, and diverse dialects. These linguistic characteristics make it more difficult to preprocess and extract meaningful features from Arabic text compared to languages like English. Additionally, the limited availability of high-quality Arabic SMS datasets further complicates the development of accurate classification models.

To address these challenges, this research focuses on applying a Random Forest machine learning model for detecting spam in Arabic SMS messages. Random Forest is known for its robustness, ability to handle high-dimensional data, and effectiveness

in classification tasks. The study also incorporates essential preprocessing techniques such as text cleaning, tokenization, and feature extraction to improve model performance.

Furthermore, the research emphasizes the use of real-world datasets to ensure that the model learns realistic spam patterns. The performance of the proposed model will be evaluated using standard metrics such as accuracy, precision, recall, and F1-score, and compared with other machine learning algorithms to determine its effectiveness.

Problem Statement:

The rapid growth of mobile communication has led to an increasing reliance on Short (SMS) as a primary means of information exchange. However, this growth has been accompanied by a significant rise in unsolicited and harmful messages, commonly referred to as spam. Spam poses significant threats to user privacy and information security (Das et al., 2026). These messages often include advertisements, phishing attempts, and fraudulent content, posing risks to users and consuming device resources such as storage and processing capacity.

The Significance of a spam filtering solution using machine learning algorithms incapacity. users time and effort in filtering spam from non-spam messages.

Spam detection in text-based messages essentially involves the use of machine learning for the identification of unwanted or malicious messages in texts such as emails, SMS (Ahmad et al., 2025). should focus on real-life data for training experiments and models, rather than manually generated datasets (Ruchi et al., 2024). as they can be trained on categorized data to predict new message classifications and spam characteristics such as keywords, punctuation, and requests for sensitive data. The Arabic language, with its rich morphology and complex structure, presents unique challenges for text classification tasks such as spam detection (Al Saidat et al., 2024). However, many existing solutions and models are in English and ineffective when dealing with Arabic SMS messages.

The research problem focuses on SMS spam messages in Arabic. Given some limitations and gaps on SMS filtering models.

The aim of this research is to apply an efficient and reliable machine learning-based model, specifically using the Random Forest algorithm, for accurately detecting and classifying Arabic SMS Spam messages into spam categories, based on their content and characteristics to reduce unwanted messages.

Research objectives:

The goal is to Apply Random Forest model an effective machine learning-based filtering model capable of accurately distinguishing Arabic non- Spam and Spam incoming messages based on their content and characteristics enhancing user protection from spam and potentially harmful content

1-To Apply a machine learning-based model capable of classifying SMS messages into spam and ham categories.

2-To train the proposed model using a suitable SMS dataset to ensure effective learning of message patterns.

3-To apply appropriate data preprocessing techniques, including text cleaning, tokenization, and feature extraction, to improve model performance.

4-To evaluate the performance of the model using measurable metrics such as accuracy, precision, recall, and F1-score.

5-To compare the performance of different machine learning algorithms to identify the most effective approach for SMS spam detection.

Research Scope:

The research primarily works to apply the Random Forest algorithm for accurately detecting and classifying Arabic SMS Spam messages into spam categories, to reduce unwanted messages.

A key aspect of the study is the use of real-world SMS datasets to ensure that the model captures realistic patterns and characteristics of spam messages. The research includes preprocessing such as text cleaning, normalization, tokenization, and feature extraction to address the linguistic complexity of the Arabic language, including its morphology and variability in writing styles.

The scope further includes training and testing the proposed model using labeled datasets, enabling the system to learn distinguishing

Model performance will be evaluated using standard metrics, including accuracy, precision, recall, and F1-score, to measure its effectiveness and reliability.

To create an effective spam filtering solution, it must incorporate key functionalities as outlined below:

1- Real-time filtering capability: The system should be able to swiftly make decisions to filter out SMS

spam for the benefit of mobile device users.

2- Adaptive learning: Utilizing machine learning algorithms is essential for the SMS spam filter to

continuously evolve and counter new spamming tactics employed by spammers.

3- Precision: The system must be engineered to deliver the utmost accuracy in identifying and blocking

spam messages.

However, this research is limited to SMS text data and does not extend to other communication platforms such as email, social media, or multimedia messaging services. It also focuses solely on content-based filtering and does not incorporate behavioral analysis, network-level detection, or real-time deployment considerations.

Chapter 2:

Literature reviews:

(Maqsood, U. 2023) performed In this paper the application of multiple machine learning classifiers and a deep learning classifier to detect spam in SMS and email datasets with improved accuracy. Experiments were conducted on real datasets to evaluate the performance of the proposed system. The researcher found that the support vector machine (SVM) classifier outperformed other models, indicating its effectiveness for classification tasks. The study's findings suggest that the proposed system offers enhanced accuracy compared to existing models for detecting spam in SMS and email messages. The SVM classifier emerged as the top-performing model, showcasing its suitability for classification purposes in this context.

In their study (Al Saidat et al., 2024) proposed a hybrid deep learning model that combines Convolutional Neural Networks (CNN) with Bidirectional Long Short-Term Memory (Bi-LSTM) networks to improve spam detection performance. This study aimed to detect potentially harmful and spam messages in Arabic SMS communications using deep learning techniques. The proposed model was applied to a dataset consisting of Arabic SMS messages that included both spam and legitimate (ham) messages. Prior to training, the dataset underwent several preprocessing steps, such as text cleaning, tokenization, and padding, in order to prepare the textual data for the deep learning model. The hybrid model was trained using the Adam optimizer and evaluated using several performance metrics, including accuracy, precision, recall, and F1-score. Additionally, an early stopping technique was implemented during training to prevent overfitting. The experimental results showed that the proposed model achieved high classification performance, with an accuracy of 0.9699, precision of 0.9739, recall of 0.9675, and an F1-score of 0.9707. These results demonstrate the effectiveness of the hybrid CNN–BiLSTM model in accurately detecting spam messages in Arabic SMS datasets. The study concluded that the proposed approach provides a robust and reliable solution for Arabic SMS spam classification.

This research paper which (Shinde, 2024) explored a method for developing a model designed to detect smishing messages via SMS by understanding the relationships between contextual factors, textual content, and images, which are often overlooked in previous studies. The research utilized Optical Character Recognition (OCR) to analyze a dataset containing both legitimate and spam SMS messages. The study evaluated two main approaches. The first approach employed deep learning models,

including Recurrent Neural Networks (RNN-Flatten), Long Short-Term Memory (LSTM), and Bidirectional LSTM (Bi-LSTM). The second approach involved traditional machine learning techniques such as K-Nearest Neighbors (KNN) and K-means, along with feature extraction methods including Term Frequency–Inverse Document Frequency (TF-IDF) and Principal Component Analysis (PCA). The experimental results showed that the KNN-Flatten model achieved the best performance, reaching an accuracy of 94.13%, indicating its potential effectiveness for real-time smishing detection systems. However, the study also noted that the model may face scalability challenges due to its computational complexity, which could limit its performance when applied to large-scale datasets.

Nawaz et al (2025) focuses on their study developing an accurate system for detecting and classifying spam messages using machine learning techniques. The researchers evaluated several classification algorithms and incorporated multiple features to enhance the model's performance. The methodology involved applying five machine learning classifiers—Logistic Regression (LR), Naïve Bayes Classifier (NBC), Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Machine (GBM). The models were trained and tested using the SMS Spam Collection dataset obtained from Kaggle. In addition, the study utilized eleven different features, including links, character count, date information, email addresses, emojis, and phone numbers, to improve the effectiveness of the classification process. The performance of the models was evaluated using several metrics such as accuracy, sensitivity, precision, recall, F1-score, and specificity. The results indicated that the Light Gradient Boosting Machine (LGBM) model achieved the best performance among the tested algorithms. It obtained a sensitivity, precision, recall, and F1-score of 100%, along with a specificity score of 100%, resulting in an overall accuracy of 100%. The study concluded that the proposed approach significantly outperformed many recent methods in the field of SMS spam detection.

In this study Machine Learning-Based Optimization Algorithms for Spam SMS Classification proposed (Das et al.,2026) a novel framework for SMS spam detection that integrates machine learning techniques with optimization algorithms. The researchers combined the Golden Search Optimizer (GSO) and the Whale Optimization Algorithm (WOA) to develop a hybrid optimization approach known as GSOWOA. This optimization method was integrated with the Random Forest (RF) classifier in order to enhance classification performance and reduce convergence time.

The proposed model focused on improving feature selection and optimizing model parameters to achieve more accurate spam message classification. Experimental results the model achieved an overall accuracy of 98.9% and a precision of 97.85%, while also improving convergence speed by approximately 35% compared with conventional methods.

Furthermore, the GSOWOA approach showed strong resistance to overfitting and maintained high computational efficiency. These results indicate that the proposed framework is effective for real-time SMS spam detection, particularly in environments with limited computational resources such as edge devices

proposed (Yadav et al., 2025). a scalable SMS spam detection framework that integrates Natural Language Processing (NLP) techniques with an ensemble approach combining both machine learning and deep learning models. The system applied preprocessing steps such as tokenization, stemming, and lemmatization, while feature extraction was performed using TF-IDF and CountVectorizer. Several classifiers, including Random Forest, Support Vector Machine, Extra Trees, Bi-LSTM, GRU, and a transformer-based model, were combined using a soft-voting ensemble. The results showed that the proposed framework achieved a high accuracy of 98.92% and an F1-score of 0.987. In addition, Explainable AI techniques such as SHAP and LIME were integrated to enhance model interpretability. The system was deployed using a Flask API to enable near real-time SMS classification, making it suitable for practical applications in sectors such as banking and e-commerce.

Researchers (Ahmad et al., 2025). provided develop an effective machine learning–based framework for accurately detecting and classifying spam messages. To achieve this, the study employs a supervised learning approach using labeled SMS datasets, along with preprocessing techniques such as text cleaning and feature extraction (e.g., TF-IDF) to improve classification performance. Multiple machine learning models, including ensemble or tree-based algorithms, are trained and compared using evaluation metrics such as accuracy, precision, recall, and F1-score. The dataset used consists of real-world SMS messages categorized into spam and ham, ensuring realistic model training. The key findings show that machine learning models especially ensemble approaches, can achieve high accuracy in spam detection, and that proper preprocessing and feature representation significantly enhance model performance, making these approaches effective for real-world spam filtering systems.

The researchers (Sumathi V P. et al. 2023) explored various machine learning techniques for effective spam detection. The objective of this study is to compare the performance of different machine learning algorithms in classifying SMS messages as spam or ham and to determine the most accurate method. The methodology involves applying supervised learning algorithms such as Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes, Decision Tree, and Random Forest to labeled SMS datasets, followed by evaluating their performance using accuracy metrics. The findings reveal that although all models perform reasonably well, Support Vector Machine (SVM) achieves the highest accuracy and proves to be the most effective technique for SMS spam classification.

The researcher (Asgher, Tayba.2023). In this research paper, the researchers applied classification methods in addition to machine learning algorithms. The aim of the study was to determine the number of unwanted short text messages. The

methodology included different classification methods on a dataset collected by a Weka tool; the results showed 100% accuracy using a random forest.

A study titled “Measure Effectiveness of SMS Spam Detection Model Based on Machine Learning Techniques” by A. Osman, M. Al-Khalifa. (2023) addresses the increasing issue of unwanted SMS spam and the need for efficient filtering systems. The objective of the study is to evaluate the performance of various machine learning algorithms, with a particular focus on the Random Forest algorithm for classifying messages as spam or ham. The methodology involves using a labeled SMS dataset consisting of thousands of text messages, where Natural Language Processing (NLP) techniques such as text preprocessing and feature extraction are applied before training multiple models, including Naïve Bayes, Decision Tree, and Random Forest. The models are then evaluated using performance metrics such as accuracy and execution time. The findings reveal that the Random Forest algorithm outperforms the other models, achieving an accuracy of approximately 99% along with efficient processing speed, making it a highly effective approach for SMS spam detection.

Chapter 3:

Research Methodology:

This study adopts a supervised machine learning approach to apply an effective model for detecting Arabic SMS spam. The methodology consists of several key stages, including data collection, preprocessing, model development, training, and evaluation.

Phase 1: Data collection: The SMS Spam Collection is a dataset of SMS messages that have been tagged for SMS spam research.

Phase 2: Preprocessing data set: Due to the complexity of the Arabic language, several preprocessing techniques are applied to prepare the data for analysis. These include: (Data cleaning, Tokenization, normalization, Removing stop words, vectorization)

Phase 3: Data splitting: Data splitting is a crucial process in machine learning, will divide the dataset into 80% that will be used to train the machine learning model, and 20% will be used to test the machine learning model on unseen data.

Phase 4: Machine learning model used:

The primary model used in this study is the Random Forest algorithm, selected for its robustness, ability to handle high-dimensional data, and strong performance in classification tasks. The model is trained to classify SMS messages into spam or ham based on features such as:

1. Robustness: Combines multiple decision trees to reduce overfitting and improve generalization.
2. Feature Importance: Provides insights into feature importance, helping in feature selection and understanding the data.
3. Versatility: Can be used for both classification and regression tasks.

Phase 5: Evaluation: The effectiveness of the model is assessed using standard evaluation metrics

1. Accuracy: The ratio between the correct positive predictive value and the total predictive value. The higher it is, the better the model. Accuracy is a

good measure only when the negative and false positive predictive values are similar.

2. Precision: The ratio between the correct positive predictive value and the all positive predictive value; A high value is associated with a low false positive predictive value.

3. Sensitivity (Recall Sensitivity): The ratio between the correct positive predictive value and the entirely correct predictive value: positive and negative.

4. (F1-score): The weighted arithmetic average of the precision and sensitivity value of the results.

Expected Outcomes:

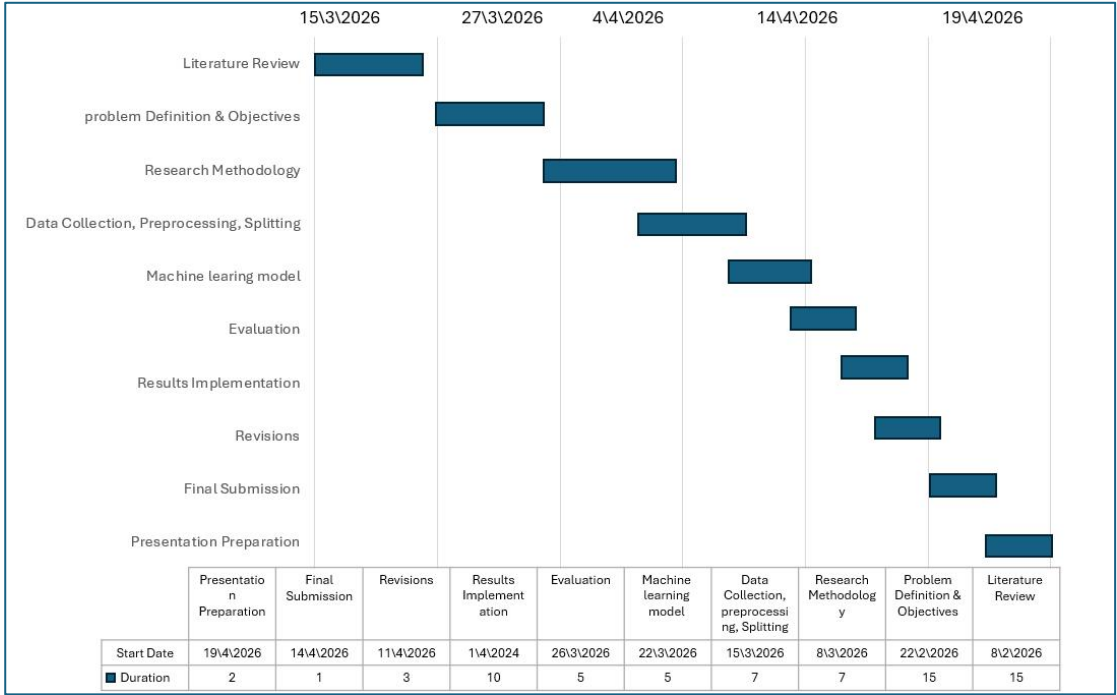
The study is expected to demonstrate that the proposed model can accurately classify Arabic SMS messages into spam and ham categories despite the linguistic complexity of the Arabic language. By applying appropriate preprocessing techniques such as text cleaning, tokenization, normalization, and feature extraction, the model is likely to achieve improved data representation and enhanced classification performance. It is anticipated that the Random Forest model will produce high evaluation results in terms of accuracy, precision, recall, and F1-score, indicating its effectiveness in minimizing false positives and false negatives.

Furthermore, the study is expected that using real-world SMS datasets improves the model's ability to generalize and detect unseen spam messages more effectively. In comparison with other machine learning algorithms, Random Forest is likely to demonstrate superior or competitive performance due to its robustness and ability to handle high-dimensional data.

Overall, the research is expected to Random Forest model, for accurately detecting and classifying Arabic SMS Spam messages into spam categories, based on their content and characteristics to reduce unwanted messages.

while providing a scalable and practical solution for Arabic spam detection systems.

Gantt chart



References

1. Ahmad, M, Rozlan, M., Yusri, M. (2025). Spam Detection Using Machine Learning Techniques. IEEE.
<https://www.techrxiv.org/doi/full/10.36227/techrxiv.173933255.51566942/v1>
2. Ruch, A, Anshita, D. Surya, K. (2024). A novel approach for spam detection using natural language processing with AMALS models. IEEE Access.
3. Maqsood, U. (2023). An intelligent framework based on deep learning for SMS and e-mail spam detection. Wireless Communications and Mobile Computing. *Wiley online library*. <https://doi.org/10.1155/2023/6648970>
4. Al Saidat, M. R., Yerima, S. Y., & Shaalan, K. (2024). A novel approach for Arabic SMS spam detection using hybrid deep learning techniques. Procedia Computer Science.
<https://www.sciencedirect.com/science/article/pii/S187705092403000X>
5. Shinde, A. (2024). SMS scam detection application based on optical character recognition for image data using unsupervised and deep semi-supervised learning. Sensors. <https://www.mdpi.com/1424-8220/24/18/6084>
6. Nawaz, I., Khosa, S., Fatima, R., Saeed, M., & Hashmi, M. S. A. (2025). Smart filters for SMS spam: A machine learning approach to SMS classification. Spectrum of engineering sciences.
<https://thesesjournal.com/index.php/1/article/view/333>
7. Yadav, A., Yadav, R., & Kiruthika, M. (2025). SMS spam detection with NLP and ensemble learning for enhanced classification. IEEE.
<https://ieeexplore.ieee.org/document/11159039>
8. Das, L., Ahuja, L., & Pandey, A. (2026). Machine learning-based optimization algorithms for spam SMS classification. Engineering, Technology & Applied Science Research.
<https://www.etasr.com/index.php/ETASR/article/view/14670>
9. Sumathi V P, Vanitha V, Tania Toyo T, (2023). Performance Comparison of Machine Learning Algorithms in Short Message Service Spam Classification. IEEE Xplore. <https://ieeexplore.ieee.org/document/10199265>
10. Asher, Tayba (2023) A Study of SMS Spam using Machine Learning. ResearchGate. https://www.researchgate.net/publication/345311834_A_Study_of_SMS_Spam_using_Machine_Learning
11. A. H. Osman and M. B. Al-Khalifa (2023). journal of Engineering Sciences and Information Technology.
<https://journals.ajsrp.com/index.php/jesit/en/article/view/6303/5998>