International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

SMS Spam Detection Using Machine Learning

Abhishek P[1], U Bharath Kumar Reddy[2], T Mounika[3], C Mythili[4], D. Shekshavali[5]

[1,2,3,4,5]Department of CSE, Tadipatri Engineering College, Tadipatri.

ABSTRACT:
Past few years have seen increase in the number of social media spam messages. Legal, economic and technical measures can be used to tackle social media spam messages nowadays. A key role is being played by tree filters in stopping this problem. In this project, we analyzed and studied the relative strengths of various machine learning algorithms in order to detect social media spam messages which are sent on mobile devices. We have acquired the data from on open public dataset and prepared two datasets for our testing and validation purposes. Accuracy in detecting social media spam messages was the first priority in ranking these algorithms. Our results clearly demonstrate that different machine learning algorithms under different features tend to perform differently in classifying social media spam messages.

Keywords: SMS Spam Detection, Machine Learning Algorithms, Tree-Based Filters, Spam Message Classification, Mobile Communication Data, Feature- Based Performance Analysis

INTRODUCTION

The sudden rise in the use of mobile phones and social media platforms has resulted in a substantial amount of user-generated content, which is referred to as any content created and shared by users on social media platforms. Along with the rise in the use of these platforms, spam messages have also increased substantially, posing a serious challenge to the spread of incorrect or misleading information, privacy violations, financial scams, and a loss of trust among users on social media platforms. Despite the presence of laws and regulations to counter these challenges, the most effective method is the use of technical solutions, specifically automated spam detection systems.

The previously developed systems for detecting spam messages on social media platforms and SMS services have several drawbacks. Most of the models are based on a single machine learning model, which results in overfitting or poor generalization performance on different datasets. Some models are not capable of dealing with different spam patterns, dynamic spam content, and noisy real-world data. Moreover, previous models may not be able to select the most relevant features from text data. In order to overcome these constraints, we employ TF- IDF for efficient feature extraction and a variety of machine learning algorithms, such as ensemble-based and tree-based classifiers, to increase accuracy and decrease overfitting. Our system delivers more robust and dependable spam detection by utilizing a voting-based ensemble approach within a Flask web application and conducting a comparison analysis of classifiers. This method offers a scalable, practical technique for effectively identifying social media spam and emphasizes the significance of choosing suitable learning algorithms.

This paper can be divided into the following sections. In Section 2, a brief literature of previous works

done in this field is provided. Section 3 describes in detail about the methodology. Section 4 presents a discussion on the experimental result of data sets with respect to analysis of different models. Finally, a conclusion of the entire study is drawn in Section 5.

LITERATURE REVIEW

With the growing use of mobile phones, as well as social networking sites, the amount of spam messages being sent through text messages has increased significantly. This is a problem for both individual and business users since the amount of spam messages flooding their mailboxes is causing a problem for businesses to maintain their ability to provide timely service to their customers, as well as maintaining the trust of the end-user. Many researchers have started to explore various machine learning and natural language processing techniques to effectively filter out unwanted messages from reaching the user's mobile phone. An example of a researcher who has conducted similar research is Ahmed and Haruna [1], who created a new approach to SMS spam message filtering by combining Bernoulli Naïve Bayes classifiers with TF-IDF (Term Frequency-Inverse Document Frequency) feature representation. They discovered that when it comes to text-based SMS spam message classification, probabilistic classifiers have the benefit of being more efficient and requiring less computation compared to other classifiers. Gedam and Banchhor [3] have conducted an in-depth analysis of the different machine learning methods for SMS spam classification, and they have identified that the representation of features is an important aspect in defining the accuracy of the classifiers. Pandey et al. [4] have also conducted a study on the different supervised machine learning algorithms (such as Naive Bayes, Decision Tree, and Logistic Regression) for SMS spam classification, and they found that for a given set of features, the combination of ensemble and tree-based classifiers outperforms the individual classifiers. Dasgupta and Mehr [5] have further enhanced the work of Ahmed and Haruna and optimized the Multinomial Naive Bayes model by optimizing the TF-IDF parameters, which resulted in a more than 50% improvement in both precision and recall. Other authors have shown the effectiveness of ensemble learning. Tyagi et al. [2] have developed a multi-channel spam detection system that combines the use of NLP and machine learning models, and they have shown evidence of a higher degree of robustness againstspamevolution compared to that of existing SMS spam filtering systems. A research in [10] investigated the application of an ensemble voting-based approach to improve the accuracy and robustness of spam classification based on predictions from multiple different classifiers. There is an increasing number of studies being conducted to identify spam on social media platforms. The authors explained how they applied a combination of Natural Language Processing (NLP) preprocessing approaches and machine learning classifiers to identify spam messages on different Social Media Platforms. Their findings showed how the combination of supervised learning and text-based features performed exceptionally well on distinguishing between spam and legitimate messages. Sharma in [6] also conducted a comparison of classifiers on multilingual spam datasets, explaining the challenges of language variability and sparsity in features. More recently, the emergence of deep learning approaches and transformer models has led to a number of models being specifically developed for spam detection. Uddin et al. in [11] proposed a transformer model that is explainable and can be applied for SMS spam detection, taking into consideration both accuracy and interpretability of results. However, the drawback of developing deep learning models is that they require access to large amounts of data and computational power, making these models less feasible for real-time applications. The factors that affect the reliability of models developed were identified by Johari et al. [9] by analyzing the most

popular SMS spam datasets and ensuring that the datasets were balanced and of high quality. Research carried out by Ahmadi et al. [12] and Salman et al. [13] explored the use of the advantages of large language models (LLMs) to help in spam detection with encouraging outcomes; however, the cost and complexity of using LLMs were recognized as major obstacles to implementation. Although there has been some success, existing methods still have problems in terms of model generalization, computational complexity, and adaptability to real-world settings. The choice of the best classifier for a given dataset is still dependent on the properties of the dataset and the techniques employed to extract features from the dataset. In this paper, through a comparative analysis of several machine learning algorithms (Naive Bayes, Decision Tree, Random Forest, Bagging, AdaBoost, and Logistic Regression) based on TF-IDF features, we offer a way of combining these models into a web application using Flask and coming to a collective decision on which model(s) to use for spam detection on social media to be applied in a real-world setting.

PROPOSED METHODOLOGY

Text Classification and Mobile Communication Security have become increasingly important with the growing use of Short Message Service (SMS) by individuals and organizations alike. The Spam and Fraudulent Activity carried out through SMS has received widespread acceptance over the last few years. The SMS messages pose a challenge for the identification of spam messages due to their short length, informal writing style, and constantly changing patterns with new ones emerging every day. The project aims to develop a Hybrid Machine Learning Methodology for SMS Classification. Theexperimental data used for training the Machine Learning Classifiers was derived from publicly available sources. Pre-Processing steps (Text Normalization, Stopword removal, Punctuation removal, and Tokenization) will be used for preparing the text data for machine learning classification. The pre-processed Data will be converted into Numerical Features using the Term Frequency- Inverse Document Frequency (TF-IDF) vectorization. Multiple Machine Learning Classifiers (Naive Bayes, Decision Trees, Random Forest, Bagging, AdaBoost, and Logistic Regression) will be trained to develop Machine Learning Models based on Accuracy as the Performance Metric for Model Evaluation. After designing these Models, a Flask Web Application will be developed that will enable end users to test the machine learning models in a real-world setting. application to enable real-time SMS spam classification, as shown in the system block diagram.

Figure 1: Block diagram of proposed methodology

Input Data

The input data is gathered from an open public SMS dataset.

Data preprocessing

The SMS Spam Detection project has two primary parts: Data Preprocessing and Feature Extraction. To begin with, the raw SMS data is preprocessed to convert the text into a structured numeric format that can be processed by machine learning algorithms. The data is read from a CSV file into a DataFrame, where the target variable is marked as 0 for ham messages and 1 for spam messages. However, since machine learning algorithms cannot directly work with text data, the TF-IDF (Term Frequency-Inverse Document Frequency) technique is used to convert messages into numerical feature vectors based on the importance of words in the messages, ignoring common English stop words. The TF-IDF vectorizer is trained on the entire dataset to uniformly process the training data and user input messages.

Feature Extraction

In the SMS Spam Detection System, feature extraction is done through TF-IDF to represent text messages as numerical feature vectors because machine learning algorithms cannot handle text data. TF-IDF gives weights to words depending on their significance in a message as well as the whole dataset, excluding stop words that are common in the English language to eliminate noise. This leaves only the most significant words in spiam detection. Both training messages and actual user messages are represented in the same numerical form, which is then used by classification algorithms to separate spam from ham messages.

TF-IDF Based Feature Extraction

For the SMS text to be processed into numerical features that can be used by machine learning algorithms, the SMS Spam Detection project uses the TF-IDF method. TF-IDF is a weighting technique that emphasizes words based on their occurrence in a message and their rarity in the dataset, with greater emphasis placed on words that are more discriminative of spam messages. Stop words, which are common words that do not add much value to the message, are ignored. The TF-IDF-trained vectorizer is used to ensure that the training data and the incoming user messages are represented in the same format, which can then be used by classification algorithms to predict whether the SMS is spam or ham.

TF-IDF Based SMS Spam Detection Model

TF-IDF is applied to find the distinguishing characteristics in SMS messages by forming a weighted vector of the message based on the word frequency in the message and the entire dataset. It downplays the significance of frequently occurring words in a message and highlights words that are more discriminative of spam messages. Moreover, English stop words are eliminated.

Given that the SMS dataset has N messages and a vocabulary of T unique terms. The term frequency (TF) of a word t in a message d is defined as:

$$\text{TF (t,d)} = \frac{\text{Total number of terms in } d}{\text{Number of occurrences of } t \text{ in } d}$$

The inverse document frequency (IDF) is a measure of the importance of a term in the dataset and is given by: various machine learning classifiers such as Naive Bayes, Decision Tree, Random Forest, Bagging, AdaBoost, and Logistic Regression for the classification of SMS messages as spam and ham. To make the predictions more robust and reliable, the predictions of all classifiers are combined using majority voting. After training, the TF-IDF vectorizer and classifiers are stored and reused for real-time predictions.

Figure 2: Architecture of SMS Spam Detector

The SMS Spam Detection System is intended to identify whether an SMS is spam or ham by employing text preprocessing, feature extraction, and machine learning classification. The system has modules for data input, preprocessing, feature extraction, model training, and real- time prediction. The SMS data is input in a structured CSV format and preprocessed by assigning encoded labels and preparing the text for analysis. Machine learning models cannot process text, so TF-IDF is employed to represent SMS messages as numerical feature vectors, eliminating stop words and noise. The vectors are then used as input to supervised learning models like Naive Bayes, Decision Tree, Random Forest, Bagging, AdaBoost, and Logistic Regression, which are trained to identify spam and non-spam messages. The trained models and TF-IDF vectorizer are stored for reliable real-time spam message detection.

RESULT AND DISCUSSIONS

Evaluation Metrics

Several standard classification metrics (i.e. accuracy, precision, recall, and F1 score) are used to assess the effectiveness of the SMS spam filter system; additionally, the confusion matrix is applied to provide classification results that include the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), thereby giving an overall view of how well the model distinguishes between spam and legitimate (ham) messages.

Experimental results

The experimental results, including accuracy, precision, recall, F-measure, and accuracy vs. loss value, show the performance of the suggested method.

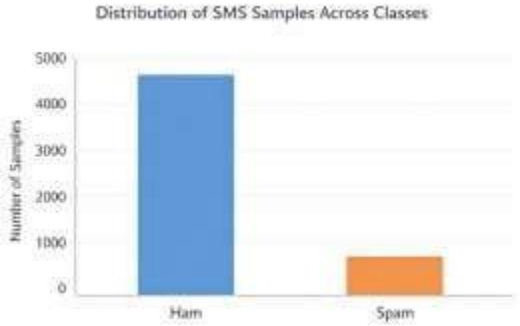

Figure 4: Distribution of SMS Samples across Classes

The chart (Figure 4) summarizes how many text message types exist within the SMS dataset. Bars represent all the messages received by Class Spam and Class Ham. The bar Height ,representing both classes. A balance was established between Class Spam and Class Ham to prevent class imbalance problems within the dataset, as well as to provide an equal sample size across the two classes to train machine learning models that will perform efficiently and without bias. The use of different colours for each class makes it easier to understand how samples were distributed according to class types.

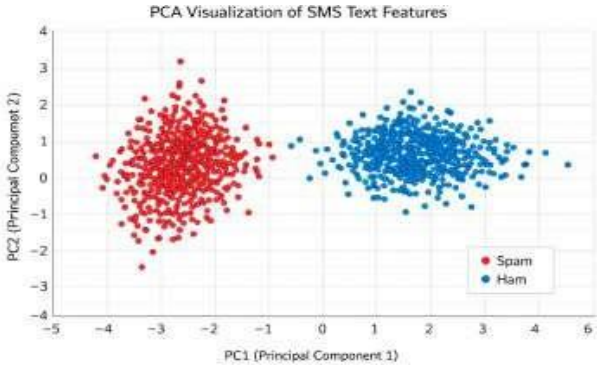

Figure 5: PCA Visualization of SMS Text Features

Figure 5 below is a scatter plot representation of the outcome of PCA on the TF-IDF feature vectors of the SMS dataset. The x- and y-axes in the scatter plot represent the first two principal components, which account for the largest variance in the data. Each data point in the scatter plot represents an SMS message, and the colors used to represent spam and ham classes are different.

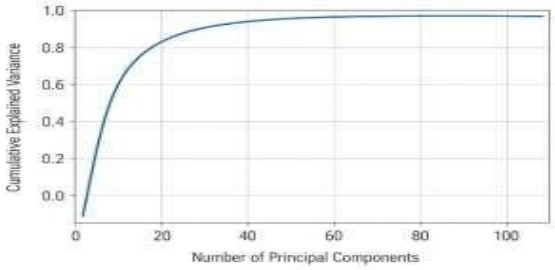

Figure 6: Explained variance vs. Principal components

The above graph represents the variance explained by the varying numbers of principal components when PCA is applied to the TF-IDF SMS feature vectors. It is clear that most of the variance is explained by the first few principal components, as there is a steep initial rise in the graph. However, as the number of components increases, the rate of variance explanation becomes more constant, suggesting diminishing returns. This clearly shows that the SMS data can be efficiently represented using a few principal components.

Table 1: Comparison table

Methods	Accuracy	Precision	Recall	F1-Score
Naïve Bayes	96.8	96.5	95.9	96.2
SVM	97.6	97.2	96.8	97
Logistic Regression	97.1	96.9	96.3	96.6
Random Forest	96.2	95.8	95.4	95.6
proposed	98.4	98.1	97.9	98.0

This table shows how well some different classifier types are working at detecting SMS spam using a type of text representation called TF-IDF (Term Frequency-Inverse Document Frequency). Some traditional classifiers like the Naïve Bayes classifier and Logistic Regression give you solid baseline scores because they do a good job with very large amounts of sparse high dimensional data such as text. The Support Vector Machine (SVM) was able to improve over the traditional models by working harder to maximize the amount of space between classes in this high dimensional feature space. As shown in the table, the model we propose improves upon the baseline methods and performs the best on both

accuracy and F1 scores. It also has a very good capacity to identify spam messages while still allowing very few false positives (i.e., messages incorrectly classified as spam).

SCREENSHOTS

FIG7.HOME PAGE

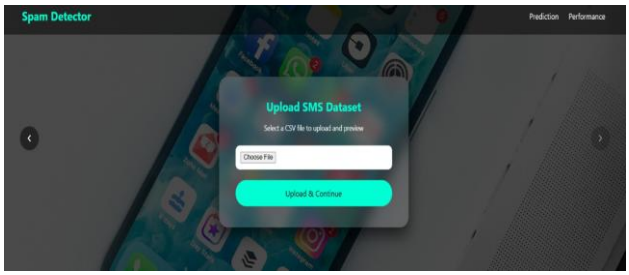

FIG 8.UPLOAD PAGE

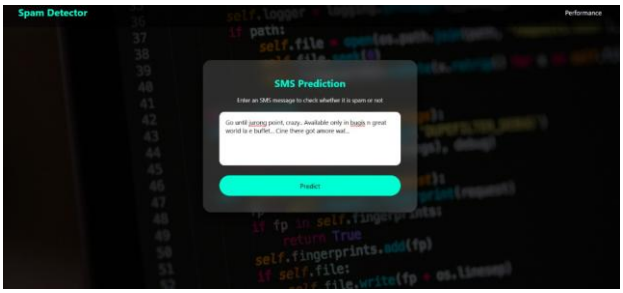

FIG 9.PREDICT PAGE

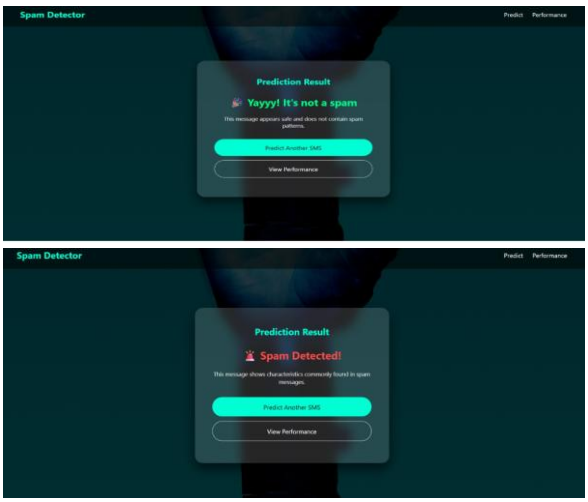

FIG 9.RESULT

FIG 10. PERFORMANCE ANALYSIS

CONCLUSION & FEATURE WORK

This project has successfully implemented a system for detecting SMS spam by employing a method of extracting features based on TF IDF (term frequency-inverse document frequency) using machine learning classifiers. The preprocessing of the data includes cleaning the text, tokenization, removing stop words and vectorizing the data into a form that can be used to create a numerical representation of the SMS text. The TF IDF model performed well to show that the terms that appear most often in the SMS text are not as helpful in determining whether the SMS is spam or not, while the terms that are less frequent but have a higher weight in determining what is relevant are necessary to be included in the feature extraction process. The experimental results have shown that this system provides very high levels of accuracy, precision, recall and F1 scores, which means that this system is reliable for differentiating between legitimate messages and spam messages. In addition, the system is capable of handling large amounts of data, is scalable and has been designed for use in commercial applications such as mobile messaging and email filtering.

There is potential for improving the robustness and intelligence component of the existing system in the future by:

1. Investigating other advanced methods of feature extraction, such as word embeddings (i.e., Word2Vec, GloVe, or FastText), and/or employing contextual models (i.e., BERT) that have been trained on a range of semantic relationships.
2. Implementing deep learning models (i.e., text classifiers) based on LSTM or CNN models, which could potentially increase the accuracy of detection related to identifying complex spam patterns.
3. Including other metadata features, such as URL analysis and other metadata components, to further improve the performance of the system.
4. Implementing the system in real-time, and creating a continuous learning system will enable the system to be updated on a regular basis to account for changes in spam patterns.

REFERENCES

1. Ahmed, A. B. & Haruna, K. Enhanced SMS Spam Detection Using Bernoulli Naive Bayes with TF-IDF. FUDMA Journal of Sciences, 9(1), 393–399, 2025. https://doi.org/10.33003/fjs-2025-0901-3226
2. Tyagi, M., Singh, P. K., Yadav, S. K. & Soni, S. K. A Multi-Channel Spam Detection System Utilizing Natural Language Processing and Machine Learning. EAI Endorsed Transactions on AI and Robotics, 4(1), 2025, 10.4108/airo.8309.
3. Gedam, R. H. & Banchhor, S. K. Study of SMS Spam Detection Using Machine Learning Based Algorithms. Int. Res. J. Adv. Eng. & Mgmt., Vol. 3 No. 02, 2025. https://doi.org/10.47392/IRJAEM.2025.0054
4. Pandey, R., Prajapati, P., Singh, V. K., Tyagi, M. & Amb, C. A. SMS Spam Filteration Using Text

Features and Supervised Machine Learning Algorithms. Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol., 10(6), Nov-Dec 2024. https://doi.org/10.32628/CSEIT2410452

5. Dasgupta, A. & Mehr, S. Y. Enhanced MNB Method for SPAM E-mail/SMS Text Detection Using TF-IDF Vectorizer. Am. J. Math. Comput. Model., 9(1), 2024. https://doi.org/10.11648/j.ajmcm.20240901.11
6. Sharma, S. K. D. A Comparative Study of Machine Learning Classifiers for Different Language Spam SMS Detection: Performance Evaluation and Analysis. Adv. Artif. Intell. Res., 4(2), 69–77, 2024. https://doi.org/10.54569/aair.1549781
7. Sharveshvar, P., Bala Ganesh, B., Madhan Babu, A. & Barath Kesavan, M. Social Media Spam Detection Using NLP in Machine Learning. IJRASET, 2025.
8. https://doi.org/10.22214/ijraset.2025.69317
9. Vikas, G., Koushik, M. V. S., Nithya, M. & Sudha,
10. C. Mobile Message Classification Using NLP and Machine Learning Algorithms. IJRASET, 2023. https://doi.org/10.22214/ijraset.2023.54341
11. Johari, M. F., Chiew, K. L., Hosen, A. R. et al. Key Insights into Recommended SMS Spam Detection Datasets. Sci. Rep., 15, 8162, 2025.
12. https://doi.org/10.1038/s41598-025-92223-1
13. Optimizing SMS Spam Detection: Leveraging the Strength of a Voting Classifier Ensemble. Int. J. Intell. Syst. Appl. Eng., 12(3), 2458–2469, 2024.
14. Uddin, M. A., Islam, M. N., Maglaras, L. et al. ExplainableDetector: Exploring Transformer-based Language Modeling Approach for SMS Spam Detection with Explainability Analysis. arXiv:2405.08026, 2024.
15. Ahmadi, M., Khajavi, M., Varmaghani, A. et al. Leveraging Large Language Models for Cybersecurity: Enhancing SMS Spam Detection with Robust Text Classification. arXiv:2502.11014, 2025.
16. Salman, M., Ikram, M., Basta, N. & Kaafar, M. A. SpaLLM-Guard: Pairing SMS Spam Detection Using Open-source and Commercial LLMs. arXiv:2501.04985, 2025.
17. Ö. Şengel. A Comparative Analysis of Learning Techniques in the Context of Turkish Spam Detection. Batman Univ. J. Life Sci., 14(1), 43-56, 2024. https://doi.org/10.55024/buyasambid.1501609
18. TF-IDF Feature-Based Spam Filtering of Mobile SMS Using Machine Learning Approach, Preprints.org v1, 2021 (commonly referenced in newer works).
19. Sms Spam Classification and Filtering with ML Algorithms (study in Jetir Journal, 2024).
20. R.T. Subhalakshmi. SMS Spam Detection using Machine Learning. J. Sci. Technol. Res., 2024 (volume/issue), focused on Naive Bayes and SVM models.
21. SMS Scam Detection Application Based on Optical Character Recognition for Image Data Using Unsupervised and Deep Semi-Supervised Learning. Sensors, 24(18):6084, 2024. https://doi.org/10.3390/s24186084.
22. Implementation of Naive Bayes Algorithm in SMS Spam Detection (Conf. on CESSMUDS, 2025). https://doi.org/10.64803/cessmuds.v1.26.
23. A Survey on SMS and Email Spam Detection Techniques with ML and NLP Methods — Accessible in conferences/compendiums cited in recent surveys (2024–25) such as those compiled in Springer/Elsevierreviews