Machine Learning-Based SMS Spam Detection

K. VASUMATHI[1], S. SELVAKANI[2] , K. SANTHIYA[3]

[1]PG Department of Computer Science,
Government Arts and Science College,
Arakkonam, Tamilnadu,
INDIA

[2]PG Department of Computer Science,
Government Arts and Science College,
Arakkonam, Tamilnadu,
INDIA

[3]PG Scholar, PG Department of Computer Science,
Government Arts and Science College,
Arakkonam, Tamilnadu,
INDIA

Abstract: - The widespread use of mobile devices has led to a significant increase in SMS (Short Message Service) spam, which undermines the integrity of mobile communication. Unlike internet-based platforms like WhatsApp or Facebook, SMS operates without internet connectivity, making it a unique target for spammers. This study presents a machine learning-based approach to detect and filter SMS spam, addressing the shortcomings of traditional email spam filters which struggle due to limited feature sets, informal language, and a lack of robust SMS spam datasets. The proposed methodology involves the integration of multiple publicly available SMS datasets, followed by data preprocessing, exploratory data analysis, and feature engineering. Various classification algorithms—including Naive Bayes, Support Vector Machine (SVM), and Random Forest—are implemented and compared based on their precision, recall, F1-score, and overall accuracy. Experimental results demonstrate that the proposed models can effectively distinguish spam from legitimate messages, with the SVM model achieving the highest classification performance. These findings have important implications for enhancing mobile security and reducing user exposure to fraudulent or intrusive content.

Key-Words: - SMS Spam, Facebook, WhatsApp, Internet Connectivity, Financial Gain, Datasets, Data Preprocessing, Feature Engineering, Naive Bayes, Model Development.

Received: April 16, 2024. Revised: January 5, 2025. Accepted: March 9, 2025. Published: May 8, 2025.

1 Introduction

The widespread accessibility and inherent simplicity of SMS have rendered it an attractive target for malicious actors, leading to unnecessary financial burdens for mobile users and undermining the security of Mobile Message Communication. Numerous individuals and organizations exploit this medium to disseminate unsolicited bulk messages, commonly referred to as Spam SMS. This project seeks to develop a robust SMS spam detection system through the application of Machine Learning algorithms. We will explore various ML techniques, including Naive Bayes, Support Vector Machines (SVM), and Random Forests, to evaluate and classify SMS messages based on their content, linguistic features, and other relevant attributes. Through comprehensive training and evaluation procedures, our objective is to construct a highly accurate and efficient spam detection model capable of identifying subtle patterns and characteristics inherent to spam messages.

Machine Learning presents a promising solution by enabling the automated detection of spam messages through the identification of patterns and features derived from labeled data. A variety of ML models, including Naive Bayes, Support Vector Machines, and neural networks, can be trained using features extracted from the text, such as word frequencies, n-grams, and semantic properties. Additionally, feature engineering methods and preprocessing steps, including tokenization and TF-IDF normalization, play a crucial role in enhancing the performance of spam detection systems. By continually refining and updating these models with new data, SMS spam detection systems can adapt to emerging spamming tactics, providing users with a reliable defense against unwanted messages while maintaining effective communication and preserving the user experience.

In the contemporary era, the relentless evolution of spam masterminds, encompassing the indiscriminate propagation of insidious schemes—predominantly targeting corporate entities while incorporating adversarial components—has escalated into a critical issue for SMS services offered by Internet Service Providers (ISPs), businesses, and individual users alike. Recent analyses indicate that over 60% of all SMS traffic constitutes spam. This deluge overwhelms SMS frameworks, leading to bandwidth congestion and excessive demands on server storage capacity, resulting in annual financial losses for enterprises amounting to several billion dollars.

Moreover, phishing spam messages represent a significant security threat to end users by deceitfully soliciting sensitive personal information, such as passwords and account credentials. These deceptive communications often masquerade as legitimate correspondence from reputable online organizations, particularly financial institutions. While it is widely acknowledged that substantial modifications to Internet regulations could provide a robust solution to the spam crisis, achieving such changes in the short term is deemed impractical. Consequently, a plethora of countermeasures has been proposed, ranging from conservative and aggressive approaches—such as the CAN-SPAM Act in the United States—to more progressive methodologies.

One such progressive approach involves the deployment of software filters, either at ISP email servers or at the client level, designed to identify and automatically eliminate or appropriately manage spam messages. Although server-side spam mitigation is regarded as pivotal in alleviating the problem (Geer, 2004; Holmes, 2005), it is not without its drawbacks. For instance, these solutions may inadvertently delete legitimate communications falsely classified as spam and fail to address bandwidth overload, as spam messages still traverse the network before being filtered.

Initially, anti-spam systems relied predominantly on keyword detection within email subjects and bodies. Short Message Service (SMS) stands as the most prevalent and widely utilized messaging medium. The term “SMS” refers both to user engagement and all forms of concise text-based communication across numerous regions globally. However, spammers have adeptly

introduced variations to evade detection, thereby driving the advancement of more sophisticated spam filtering mechanisms.

It has evolved into a powerful conduit for disseminating announcements, promoting products, delivering banking notifications, sharing agricultural insights, providing flight updates, and extending internet-related offers. Moreover, SMS plays a pivotal role in direct marketing, commonly termed SMS marketing.

However, SMS marketing occasionally becomes a source of concern for recipients. Such unsolicited messages are classified as spam SMS. Spam encompasses one or multiple unwelcome communications, which are undesirable to recipients and disseminated as part of a larger batch of messages with substantially identical content. The primary objectives of SMS spam include promotional advertising, political propaganda, circulation of inappropriate adult content, and distribution of internet-based offers. Consequently, the inundation of spam SMS has emerged as a pressing global issue.

SMS spamming has garnered widespread prominence over other spamming methodologies, such as email and Twitter-based spam, owing to the ever-increasing reliance on SMS communication.

2 Related Works

Gangare et al. (2022) [1] employed the Count Vectorizer for feature selection and implemented the Naïve Bayes Multinomial classifier to distinguish SMS messages. Their model demonstrated remarkable proficiency, attaining an efficiency rate of 94%, underscoring its effectiveness in accurately identifying and classifying spam messages with high precision.

Boudaa, El Mohajir, and Boutkhoum (2021) [2]. Their study focused on utilizing the Naïve Bayes classifier, a probabilistic model known for its simplicity and efficiency, to detect spam messages in SMS datasets. The authors demonstrated the algorithm's effectiveness in distinguishing between spam and legitimate messages by leveraging relevant features such as message content and word occurrence patterns. Their research highlighted the suitability of Naïve Bayes for real-time, large-scale SMS spam filtering, offering a practical solution for enhancing mobile communication security.

Julis and Alagesan (2020) [3].The authors developed a model that effectively differentiated between spam and legitimate SMS messages by leveraging advanced feature extraction methods and classification algorithms. Their research evaluated multiple machine learning approaches, including Naïve Bayes, Support Vector Machines (SVM), and Decision Trees, to identify the most effective techniques for spam detection. By analyzing the linguistic and structural features of SMS data, their model achieved notable accuracy. This study contributed valuable insights into the efficacy of machine learning in combating SMS spam in many situation with a vast content about different algorithms.

Lota and Hossain (2017) [4]. Their work comprehensively analyzed existing methodologies for identifying and mitigating SMS spam, focusing on various machine learning, statistical, and heuristic approaches. By evaluating a wide range of techniques, including Naïve Bayes, Support Vector Machines (SVM), and hybrid models, the study highlighted the strengths and limitations of each method. The review also emphasized the importance of feature selection and preprocessing techniques in improving detection accuracy. This research provided a valuable foundation for future advancements in SMS spam filtering systems. This paper insist about the filtering of the SMS using machine learning concepts.

Liu (2021) [5] employed five distinct classification algorithms in conjunction with Word2Vectorizer and TF-IDF Vectorizer. Additionally, they utilized Syntactic Parsing to scrutinize the syntactic relationships among words within SMS messages. Among the evaluated models, Logistic Regression, when integrated with Word2Vec embedding technique, exhibited superior performance in classifying spear phishing messages.

Krishnaveni and Radha (2021) [6], a comparative analysis was performed between the Naïve Bayes (NB) and Support Vector Machine (SVM) algorithms for spam SMS detection utilizing Natural Language Processing (NLP). The Count Vectorizer was employed to ascertain the frequency of unique words within the given dataset. Upon evaluation of both classifiers, SVM demonstrated superior efficacy over Naïve Bayes across all assessed metrics. The SVM model achieved an accuracy of 94.32%, a precision of 92.84%, a recall of 93.07%, and an F-measure of 94%.

Paras Sethi et al. [7] elaborated on the global severity of SMS spam and its far-reaching implications. They analyzed various algorithms available for model evaluation, identifying the most effective among them. Additionally, their study encompassed an in-depth examination of diverse filtering methodologies to enhance spam detection efficiency.

Gupta S. et al. (2021) [8] utilized a Python-based Flask framework as their development platform, implementing TF-IDF vectorization to construct a word cloud vector. Through this approach, they attained a model with an impressive accuracy of 95.90%.

Sahadevan and Subramanian (2019) [9]. The authors focused on leveraging the probabilistic nature of the Naïve Bayes classifier to effectively classify SMS messages as either spam or legitimate. Their approach involved extracting key features from the message content, such as word frequencies and message length, to train the model. The study demonstrated the Naïve Bayes algorithm's efficiency and accuracy in SMS spam detection, emphasizing its potential for real-time applications in mobile environments where speed and resource efficiency are critical.

Warade, Tijare, and Sawalkar [10]. Proposed a comprehensive approach for SMS spam detection, focusing on the application of various machine learning algorithms and feature extraction techniques to improve the accuracy and efficiency of spam classification. Their method combined the analysis of textual content with advanced data preprocessing steps to extract meaningful features from SMS messages. By leveraging algorithms such as Naïve Bayes, Decision Trees, and Support Vector Machines, they were able to design a robust system capable of distinguishing between spam and legitimate messages. Their work contributed to the ongoing efforts to enhance SMS security, providing a practical solution for real-time spam filtering in mobile communication systems.

3 Methodology

3.1 Data Collection Module

This module is responsible for aggregating and preprocessing a labeled dataset of SMS messages, which serves as the foundation for training the Naïve Bayes algorithm. Additionally, it has the capability to acquire incoming SMS messages for real-time classification, ensuring dynamic adaptability Feature Extraction Module: This module systematically derives the most salient attributes from SMS messages, including word occurrences and their respective probabilistic distributions. Serving as a cornerstone for the Naïve Bayes algorithm, it ensures the provision of pertinent and refined data, thereby enhancing the algorithm's efficacy in accurately categorizing incoming messages.

3.2 Evaluation Module

This module is designed to assess the efficacy of the Naïve Bayes algorithm by juxtaposing its predicted classifications of incoming messages against their actual classifications. Furthermore, it facilitates continuous monitoring of the classifier's accuracy over time, enabling the implementation of necessary refinements to enhance overall performance.

3.3 Naïve Bayes Classifier Module

This module encompasses the Naïve Bayes algorithm, which has been meticulously trained on a labeled dataset of SMS messages. The classifier module computes the probabilistic likelihood of an incoming message belonging to each predefined category and subsequently

assigns it to the class with the highest probability—either spam or ham.

3.4 Prediction

The prediction phase of SMS spam detection leveraging machine learning, the trained model analyzes new, previously unseen SMS messages to ascertain their spam classification. This determination is based on salient features extracted from the text and intricate patterns discerned from the training dataset. The model assigns a probabilistic score or categorical label to each message, signifying whether it falls under the spam or non-spam category. This predictive mechanism is paramount for real-time spam detection, where the prompt and accurate classification of incoming messages is crucial for ensuring user protection.

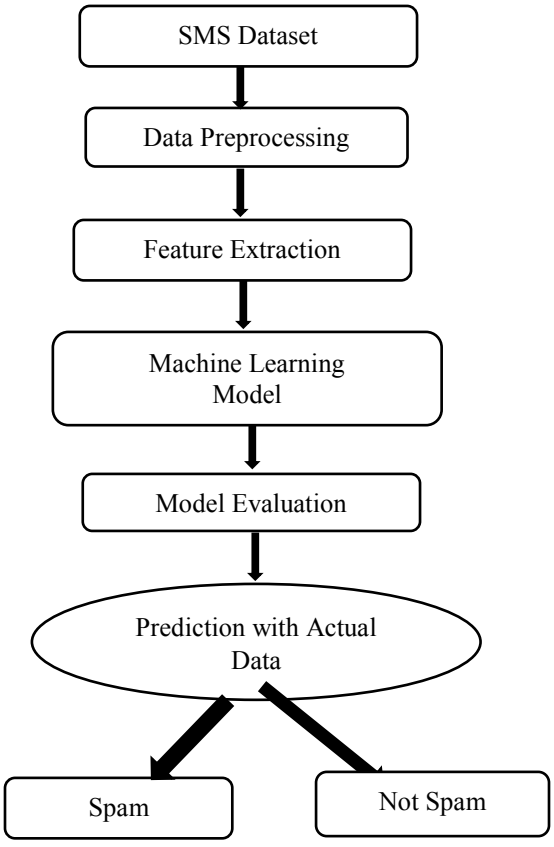


Fig 1: System Model

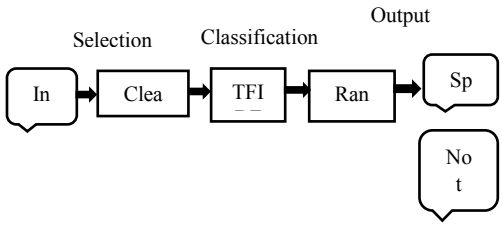


Fig 2: Propose model

5 Experiment and Results

Upon selecting the optimal features, machine learning models such as Naïve Bayes, Random Forest were employed. The evaluation of these models was conducted utilizing a Stratified 10-Fold Cross-Validation technique, based on key performance metrics including Accuracy, Precision, and Execution Time. Among these, accuracy and execution time were deemed the most pivotal, serving as fundamental benchmarks to address the research inquiry.

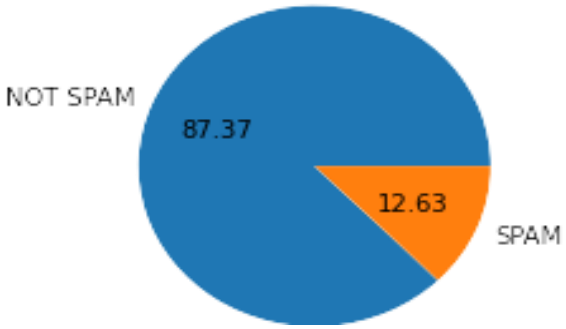


Fig 3: To display data form most common spam with vertical x-axis labels

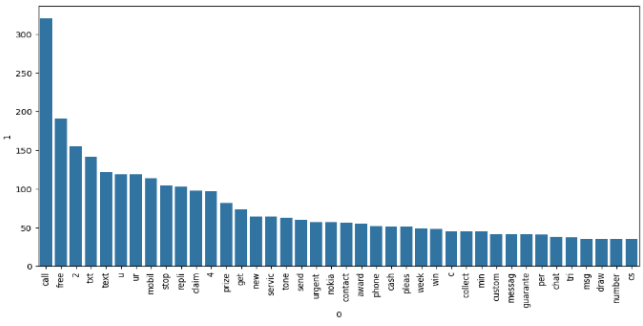
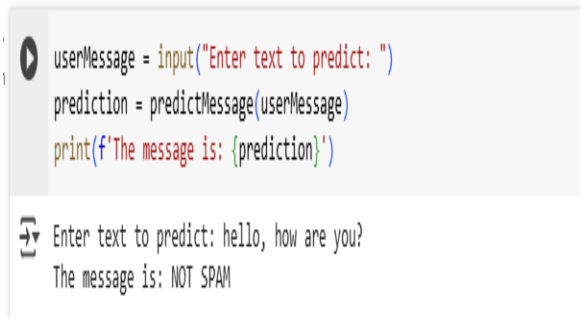


Fig.4 Common spam

6 Final Result

Code trains a **stacking classifier** using multiple models and evaluates its performance.

Accuracy = 0.9854932301740812
Precision = 0.9838709677419355



It is clear from the system identifies the message ” Hello, how are you?” as not spam which is true , with accuracy of 99.46%, and from the system Future work identified the message “congratulati on!! You have won 5000$”as spam.

7 Future work

The future trajectory of this project will encompass the incorporation of additional feature parameters. A greater number of considered parameters will invariably enhance the accuracy of the model. Furthermore, these algorithms can be leveraged for the analysis of public comment content, thereby facilitating the identification of intricate patterns and relationships between customers and corporations. The application of conventional algorithms alongside data mining methodologies can also contribute to forecasting the structural performance of enterprises holistically.

Looking ahead, we intend to integrate neural networks with complementary techniques such as genetic algorithms and fuzzy logic. The synergistic implementation of these methodologies in conjunction with neural networks holds the potential to yield significant advancements in SMS spam prediction.

8 Conclusion

In summation, the Naïve Bayes algorithm stands as a widely utilized and highly efficacious machine learning technique for SMS spam detection. It excels in text classification tasks due to its capability to manage high-dimensional feature spaces and noisy data with remarkable efficiency. The algorithm is inherently simple and computationally expedient, rendering it well-suited for real-time SMS spam detection applications.

To attain optimal performance, meticulous preprocessing and feature extraction from textual data are imperative, alongside fine-tuning the algorithm’s hyper parameters and rigorously assessing its efficacy through appropriate evaluation metrics. Furthermore, the utilization of a representative and diverse dataset that accurately mirrors the target population of SMS messages is crucial.

In essence, the Naïve Bayes algorithm offers a pragmatic and dependable solution for SMS spam detection, with broad applicability across various real-world implementations.

Spam detection is paramount in fortifying message and email communications against unsolicited and potentially malicious content. Attaining precise spam detection remains a formidable challenge, compelling researchers to devise and propose a multitude of innovative detection methodologies.

Leveraging machine learning for SMS spam detection presents a dynamic and robust approach, adaptable to specific requirements. Through continuous advancements and iterative model refinement, we can adeptly navigate the ever-evolving landscape of SMS spam, ensuring users experience heightened security and a seamless messaging environment.

These models can achieve exceptional levels of accuracy and precision in discerning spam messages, thereby serving as indispensable tools for SMS filtration and protection.

References

[1] A. Gangare, J. Rathore, A. Tadge, A. Shrivastav, R. Yadav, and P. Sisodiya, *Implementation of Spam Classifier using Naïve Bayes Algorithm*, International Research Journal of Engineering and Technology (IRJET), Vol. 09, 02,2022, https://www.irjet.net/archives/V9/i2/IRJET-V9I272.pdf.

[2] Boudaa, N., El Mohajir, B., & Boutkhoum, O. (2021). SMS Spam Detection Based on Naive Bayes Algorithm. In Proceedings of the 2nd International Conference on Computer Science, Information Technology and Engineering (pp. 31-35). Springer.

[3] Julis, M. R., Alagesan, S., "Mobile SMS Spam Detection using Machine Learning Techniques," INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH, Volume 9, Issue 02, February 2020.

[4] Lota, L. N., Hossain, B. M. M., "A Systematic Literature Review on SMS Spam Detection Techniques," I.J. Information Technology and Computer Science, 2017, 7, 42-50, Published Online July 2017 in MECS (http://www.mecs-press.org/), DOI: 10.5815/ijitcs.2017.07.05.

[5] M. Liu, Y. Zhang, B. Liu, Z. Li, H. Duan and D. Sun, "Detecting and Characterizing SMS Spear phishing Attacks", 930-943, 2021, doi: 10.1145/3485832.348

[6] N. Krishnaveni and V. Radha, "Comparison of Naïve Bayes and SVM Classifier for Detection of Spam SMS using Natural Language Processing", International Journal of Semantic Computing, vol. 11, no. 2 pp. 2260-2265, 2021, Doi: 10.21917/ijsc.2021.0323.

[7] P. Sethi et al. "SMS spam detection and comparison of various machine learning algorithms," International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN), 2017, pp. 28-31

[8] S. D. Gupta, S. Saha, and S.K. Das "SMS Spam Detection Using Machine Learning", Journal of Physics: Conference Series (JPCS), 2021, doi: 10.1088/1742-6596/1797/1/012017.

[9] Sahadevan, M. S., & Subramanian, K. (2019). SMS Spam Detection Using Naive Bayes Algorithm. In Proceedings of the International Conference on Computing and Communications Technologies (pp. 43 47). Springer.

[10] Warade, S. J., Tijare, P. A., Sawalkar, S. N., "An approach for SMS spam detection.