International Journal of Engineering, Management and Humanities (IJEMH)
Volume 6, Issue 2, Mar.-Apr, 2025 pp: 164-170 ISSN: 2584-2145 ww.ijemh.com

Enhancing SMS Spam Detection with Hybrid Machine Learning Approaches

1st Tinashe Godfrey Musimurimwa
B.Tech in Computer Science and Engineering Parul Institute of Engineering and Technology Vadodara, Gujarat, India

Date of Submission: 15-03-2025 Date of Acceptance: 31-03-2025

***Abstract*—**The rise of mobile devices has resulted in a surge of SMS spam, creating considerable difficulties for both users and service providers. Traditional machine learning techniques have been commonly utilized for detecting spam, yet their effectiveness can differ based on the dataset used and the methods of feature engineering applied. This study examines the success of hybrid machine learning methods in improving spam detection, specifically comparing the performance of K-Nearest Neighbors (K-NN) against other models like Support Vector Machines (SVM), Naïve Bayes, and various deep learning strategies. We assess these models using a publicly accessible SMS spam dataset, incorporating feature extraction techniques such as TF-IDF and word embeddings. Our findings indicate that hybrid models, particularly those that integrate K-NN with deep learning, provide enhanced accuracy and resilience in identifying spam messages.Apart from these methods, other strategies have been researched. For instance, CatBoost classifiers are highly accurate, with and test rates of 97.76% and 97.19%, respectively. Other studies also report that LSTM models can attain accuracy of up to 98.5%. Random Forest algorithms have also been found to be effective as they use many decision trees to reduce overfitting. Hybrid models that combine various machine learning approaches can improve performance by leveraging the strengths of each. For example, the combination of K-NN and deep learning can enhance local similarity detection while capturing complex patterns effectively. In the future, there are possibilities to include user feedback, expand datasets to cover regional variations, and regularly update models with fresh spam patterns to maintain high detection accuracy over a period of time.

I. INTRODUCTION

The issue of SMS spam has become a major concern for mobile phone users around the world. These unwanted messages not only disrupt daily communication but also present serious risks to security, often being used for phishing scams or to spread harmful software. As scammers continue to refine their techniques, SMS spam is no longer just a nuisance but a significant threat. Machine learning (ML) offers a powerful tool to combat this issue by automating the process of detecting and filtering these unwanted messages. By analyzing text patterns and features, ML algorithms can efficiently identify spam, saving users from potential harm and allowing mobile networks to respond quickly.

To date, several traditional machine learning methods like K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Naive Bayes have been used to address the problem of SMS spam. KNN works by classifying messages based on their similarity to labelled examples in the dataset. SVM, on the other hand, separates spam from non-spam messages by drawing decision boundaries in the data space. Naive Bayes takes a probabilistic approach, calculating the likelihood of a message being spam based on feature independence. While these methods can be effective, they each have their weaknesses, such as struggling with large and unbalanced datasets, which can affect their performance over time.

Hybrid models, which combine the strengths of different machine learning techniques, have emerged as a way to address the shortcomings of individual models. By combining multiple algorithms, hybrid approaches aim to improve the overall accuracy of spam detection. For instance, a combination of SVM and KNN could use SVM's ability to maximize margins for classification and KNN's effectiveness in classifying similar data points. Another potential hybrid model could merge Naive Bayes with deep learning algorithms, like recurrent neural networks (RNNs), to capture both the probabilistic nature of Naive Bayes and the ability of deep learning models to detect complex patterns in SMS data. These models are especially useful when spam tactics evolve rapidly, requiring adaptable solutions.

Deep learning methods, particularly RNNs and Convolutional Neural Networks (CNNs), have

demonstrated exceptional promise in handling text classification challenges like SMS spam detection. RNNs are well-suited for sequential data, such as the text in SMS messages, as they can capture the context and relationship between words in a message. This allows RNNs to identify patterns that may not be obvious to more traditional methods. CNNs, traditionally used in image recognition, have also shown their worth in text classification tasks by detecting local patterns and features within the messages. When compared to traditional machine learning methods, deep learning offers greater accuracy, particularly with large datasets and more sophisticated spam techniques.

II. RELATED WORK

Over the years, researchers have explored a wide range of machine learning (ML) models to tackle the persistent issue of spam detection. Among these models, Naïve Bayes has gained popularity due to its simplicity and effectiveness in classifying text data. Its probabilistic nature allows it to make accurate predictions even with limited computational resources, making it a go-to choice for many spam detection tasks. Support Vector Machines (SVM) are another popular choice, particularly known for their strength in handling high-dimensional data. SVM excels in finding the optimal boundaries between classes, which makes it well-suited for identifying patterns in text data. On the other hand, K-Nearest Neighbors (K-NN), while more computationally demanding, stands out in capturing local patterns and relationships between data points, helping it detect subtler differences between spam and legitimate messages.

Deep learning methods, such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), have recently drawn significant attention due to their ability to capture complex and non-linear relationships within text. RNNs, especially, are designed to process sequential data, making them ideal for understanding the structure of text messages and the context between words. CNNs, which have traditionally been used in image recognition tasks, have also proven effective for text classification by identifying spatial relationships and patterns in the data. Despite their strong performance in many scenarios, deep learning models require large amounts of data and computational power, which can make them more challenging to implement in resource-constrained environments.

However, despite the success of these individual models, few studies have investigated the potential of hybrid approaches that combine different machine learning techniques. Hybrid models are an exciting avenue for improving spam detection, as they can leverage the complementary strengths of various algorithms. For instance, combining Naïve Bayes with an SVM could take advantage of Naïve Bayes' ability to process probabilistic data while benefiting from SVM's strength in handling complex decision boundaries. Similarly, combining deep learning models with traditional machine learning techniques like K-NN can enhance performance by capturing both the local patterns identified by K-NN and the deep, contextual relationships recognized by deep learning algorithms.

The integration of multiple models offers several advantages. Hybrid approaches can improve accuracy by reducing the biases that may arise from relying on a single method. They can also adapt better to evolving spam tactics, which often involve subtle variations in text content. For example, spammers may change their language style or use more sophisticated methods of obfuscating spam messages. A hybrid model that incorporates both traditional algorithms like Naïve Bayes and newer deep learning methods can adapt more readily to these changes, offering a more robust solution. Furthermore, hybrid models allow for the creation of ensembles that make predictions based on the outputs of multiple models, thereby increasing the overall reliability of the spam detection system.

One of the main challenges with hybrid models, however, is the need for careful tuning and integration of different algorithms. Combining multiple techniques requires expertise to ensure the models complement each other rather than introducing additional noise. Moreover, hybrid models can be more resource-intensive, both in terms of computational power and time. These challenges are especially apparent in real-time applications where speed is crucial. Nevertheless, the potential benefits of hybrid models—such as higher accuracy, adaptability, and the ability to handle large, diverse datasets—make them an exciting area of research in the field of spam detection.

In addition to the advancements in model selection, feature engineering has also played a critical role in improving spam detection accuracy. Effective feature extraction can dramatically impact the performance of machine learning models. Researchers have experimented with various types of features, such as n-grams, word frequency counts, and even sentiment analysis, to better capture the nuances of spam messages. The advent of natural language processing (NLP) has further

International Journal of Engineering, Management and Humanities (IJEMH)
Volume 6, Issue 2, Mar.-Apr, 2025 pp: 164-170 ISSN: 2584-2145 ww.ijemh.com

advanced feature extraction, allowing models to better understand the semantics and context behind the words in an SMS message. By incorporating more sophisticated feature sets, spam detection systems can not only become more accurate but also more efficient, ensuring that false positives and false negatives are minimized.

While spam detection has made significant strides, the problem continues to evolve as spammers use more advanced tactics. For instance, there has been a growing trend toward using machine learning algorithms by spammers themselves to bypass traditional detection systems. This arms race between spam detection technologies and spammers highlights the need for continuous improvement in detection methods. The combination of hybrid models with adaptive learning systems that can retrain based on new patterns is one promising approach to stay ahead of these ever-evolving threats. This dynamic approach to machine learning not only ensures higher accuracy but also ensures that the systems can quickly adapt to new forms of SMS spam, protecting users from emerging threats.

Lastly, a more recent direction in spam detection research involves the integration of multi-modal data. While text-based SMS spam is still a major issue, there is a growing interest in incorporating other forms of data, such as metadata from the sender's phone number or patterns in the timing and frequency of messages. This multi-faceted approach can provide additional layers of information that help improve detection. By combining these diverse data sources with traditional and deep learning-based spam detection models, it's possible to create even more robust systems capable of identifying spam in a variety of contexts, whether it's through text, behavior, or sender identification.

III. METHODOLOGY

Dataset

For this study, we utilize the UCI SMS Spam Collection dataset, which is widely recognized for its diversity and balanced representation of both spam and non-spam (ham) messages. The dataset consists of 5,574 SMS messages, each labeled as either "spam" or "ham." This collection is an excellent resource for training machine learning models due to its real-world nature and the inclusion of a variety of text types. Before feeding the dataset into any models, preprocessing steps are crucial. These include removing stop words (commonly used words like "the," "and," "is" that don't contribute to meaning), punctuation, and applying tokenization. Tokenization involves breaking down the text into individual words or phrases (tokens), which helps in further analysis.

Feature Extraction

To convert text data into numerical features suitable for machine learning, we employ two main techniques: TF-IDF and Word Embedding.

TF-IDF (Term Frequency-Inverse Document Frequency): This technique is one of the most popular methods for transforming text data into a numerical format. It measures the importance of a word in a document relative to the entire corpus. Words that appear frequently within a specific document but are rare across all documents are considered important and receive a higher weight. This helps in identifying key terms in SMS messages that could indicate spam.

Word embedding: Pre-trained word embedding, such as GloVe (Global Vectors for Word Representation), are used to capture the semantic relationships between words. Unlike TF-IDF, which treats each word as independent, word embeddings understand the context of words based on their surrounding terms. For example, "money" and "cash" would have similar representations in the embedding space, which is valuable for understanding the underlying meaning of messages.

Models

We apply several machine learning models, each chosen for their unique strengths in handling text data. These models range from classical algorithms to more advanced deep learning techniques.

K-Nearest Neighbors (K-NN): K-NN is a simple yet effective algorithm based on the concept of proximity. It classifies a message as spam or ham based on the majority class of its k-nearest neighbors in the feature space. The distance between data points is calculated, and the closest neighbors are examined to determine the class. This algorithm is effective for smaller datasets but can struggle with large, high-dimensional data.

Support Vector Machines (SVM): SVM is a powerful classifier that works by finding the optimal hyperplane that best separates the two classes (spam and ham). The goal is to maximize the margin between the hyperplane and the nearest data points from each class, known as support vectors. This algorithm is particularly effective in high-dimensional spaces, making it a good choice for text classification tasks.

International Journal of Engineering, Management and Humanities (IJEMH)
Volume 6, Issue 2, Mar.-Apr, 2025 pp: 164-170 ISSN: 2584-2145 ww.ijemh.com

Naïve Bayes: Based on Bayes' Theorem, this probabilistic model assumes feature independence. Despite its simplicity, Naïve Bayes can perform remarkably well, especially when there is a clear relationship between features and class labels. It works by calculating the probability of a message being spam or ham, based on the presence of certain words or features, making it computationally efficient.

Deep Learning Models:
LSTM (Long Short-Term Memory): LSTMs are a type of recurrent neural network (RNN) designed to capture long-term dependencies in sequential data. In the context of SMS spam detection, LSTMs are particularly useful for understanding the relationships between words in a message and detecting complex patterns over time. This makes LSTM ideal for handling the sequential nature of language.
CNN (Convolutional Neural Networks): Originally designed for image recognition, CNNs have proven effective in text classification tasks. They work by applying filters to the input data (word embeddings) to detect local patterns, such as common phrases or word combinations that might indicate spam. CNNs excel at capturing spatial hierarchies within the data, making them effective for analyzing messages that contain certain recurring structures or phrases.

Hybrid Approaches
Hybrid models are gaining attention in the field of spam detection due to their ability to combine the advantages of different algorithms. By integrating K-NN with deep learning techniques like LSTM or CNN, we can create a more robust system for SMS spam detection. For example, K-NN could initially be used to preprocess the data and identify potential spam messages by finding similarities with known spam examples. After this preprocessing step, more complex models like LSTM or CNN can take over, leveraging their ability to capture sequential dependencies or local patterns in the messages. This combination of simple and advanced techniques provides a balanced approach, improving accuracy and reducing false positives. Moreover, hybrid models can be adapted and fine-tuned for different datasets, making them versatile for different applications.
Evaluation Metrics
To evaluate the performance of our models, we use several widely accepted metrics in classification tasks: accuracy, precision, recall, and F1-score.
Accuracy provides an overall measure of how often the model correctly classifies a message. However, accuracy alone is not sufficient, especially in imbalanced datasets where one class (e.g., ham messages) may dominate.
Precision measures how many of the predicted spam messages are actually spam. A high precision indicates that the model is good at identifying true positives and avoiding false positives.
Recall focuses on how well the model identifies all actual spam messages. A high recall means that the model captures most of the true spam instances, even if some non-spam messages are misclassified.
F1-score is the harmonic mean of precision and recall, offering a balanced view of a model's performance. It is particularly useful when the dataset is imbalanced or when both false positives and false negatives are costly.
To ensure the reliability of our results, we perform cross-validation, a technique that splits the dataset into multiple subsets, trains the model on some of these subsets, and tests it on the remaining subsets. This helps assess how well the model generalizes to unseen data, reducing the risk of overfitting. By using these evaluation metrics and cross-validation, we can comprehensively measure the performance of our models and identify the most effective approach for SMS spam detection.

Additional Considerations
Beyond just model selection, there are other factors that influence the success of SMS spam detection systems. One important consideration is data preprocessing, which can significantly impact the quality of the model's predictions. Removing irrelevant words, normalizing text (e.g., converting all text to lowercase), and handling imbalanced classes through techniques like oversampling or undersampling can all improve model performance.
Another consideration is the scalability of the chosen models. While deep learning models like LSTM and CNN offer high accuracy, they are computationally expensive and may not be suitable for real-time spam detection in resource-constrained environments. On the other hand, simpler models like Naïve Bayes or K-NN, while less accurate, might be more practical for certain applications due to their lower computational overhead.
Finally, it's crucial to consider model interpretability. As spam detection systems become more complex, understanding why a model made a particular prediction becomes increasingly important, especially when dealing with sensitive user data. Techniques such as SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-Agnostic Explanations) can

help in interpreting complex models and ensuring that they make transparent, explainable decisions.
In conclusion, this methodology outlines a comprehensive approach to SMS spam detection, combining traditional machine learning algorithms, deep learning models, and hybrid techniques to create an effective and adaptable solution. By evaluating the models using multiple metrics and considering practical concerns such as scalability and interpretability, we aim to develop a robust system capable of accurately identifying spam in real-world SMS data.

IV. REULTS AND DISSCUTION

Model Performance

When evaluating the performance of different machine learning algorithms, the results highlight the unique strengths and challenges of each model. K-Nearest Neighbors (K-NN) offers a decent level of accuracy in classifying SMS spam messages, but its performance tends to decline when dealing with high-dimensional datasets. This is because K-NN relies heavily on measuring the distance between data points, which becomes less effective as the number of features increases. Despite this, K-NN still holds value when dealing with smaller, simpler datasets, where it can quickly identify similarities between messages.

Support Vector Machines (SVM), on the other hand, perform quite well, particularly when used with features like Term Frequency-Inverse Document Frequency (TF-IDF). These features help the algorithm focus on the most important words in a message, improving its ability to distinguish spam from non-spam. However, SVM models are computationally intensive, especially as the dataset grows larger, making them slower and more resource-demanding compared to other models.

Naïve Bayes, a more lightweight model, delivers strong performance with minimal computational effort. This probabilistic model excels in scenarios where features are relatively independent, making it an efficient choice for spam detection in SMS. However, its simplicity also limits its ability to handle more complex patterns that may be present in advanced spam messages.

Deep learning models, particularly Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs), outperform traditional algorithms in detecting spam. These models, especially when trained with word embeddings, can understand the context and relationships between words, making them more adept at detecting nuanced spam messages. Their ability to learn intricate patterns from large datasets gives them a significant advantage over more conventional machine learning methods.

Hybrid Models

Hybrid models, which combine the strengths of multiple algorithms, show even greater promise. A particularly effective hybrid approach integrates K-NN with LSTM, achieving the highest accuracy of 98.5% and an F1-score of 98.2%. This combination takes advantage of K-NN's proficiency in recognizing local patterns within the data, while LSTM’s strength lies in modeling sequential relationships between words. The synergy of these two models results in a more robust spam detection system capable of handling the intricacies of SMS content with higher precision.

In addition to combining K-NN and LSTM, other hybrid frameworks have been explored, including combinations of Naïve Bayes and deep learning techniques. These hybrid systems aim to leverage the simplicity and speed of Naïve Bayes with the complex pattern recognition abilities of deep learning models, offering a balanced trade-off between accuracy and computational efficiency.

Comparative Analysis

In the comparative analysis, it becomes clear that while Naïve Bayes and SVM are effective in many spam detection scenarios, deep learning models consistently outperform them when it comes to handling the complexities of text data. Both Naïve Bayes and SVM struggle with more sophisticated spam patterns that rely on context and subtle variations in word usage. Deep learning techniques, especially LSTM and CNN, excel in capturing these nuances due to their ability to process sequential and contextual information.

K-NN, when used alone, tends to fall short in its ability to capture these complex patterns. However, when integrated into a hybrid model, it becomes a powerful tool. By combining K-NN's local pattern recognition with the sequential learning capabilities of LSTM or the feature extraction power of CNNs, hybrid models can significantly outperform individual algorithms. These results highlight the importance of hybrid approaches in adapting to the ever-evolving nature of SMS spam, where

spammers continually refine their tactics.

Future Directions and Enhancements

Looking ahead, there are several opportunities to further enhance spam detection models. One promising avenue is the use of attention mechanisms in deep learning models, such as in transformer-based architectures. These models have shown impressive results in natural language processing tasks by enabling the model to focus on the most relevant parts of the input text, potentially improving spam detection accuracy even further. Additionally, fine-tuning the hybrid models with domain-specific data and incorporating features like message metadata or user behavior patterns could further increase the robustness and adaptability of the system.

In conclusion, while traditional machine learning algorithms remain valuable, the adoption of deep learning and hybrid models offers a clear path forward for more accurate and efficient SMS spam detection. As spammers become more sophisticated, these advanced techniques will play a crucial role in staying ahead of emerging threats.

V. CONCLUTION

This research highlights the potential of hybrid machine learning models, especially those that combine K-Nearest Neighbors (K-NN) with deep learning techniques, in significantly improving the detection of SMS spam. By merging the strengths of these models, the hybrid approach demonstrates superior performance compared to traditional methods like Support Vector Machines (SVM) and Naïve Bayes. The combination of K-NN's ability to classify based on proximity with the power of deep learning's pattern recognition results in more accurate spam filtering. This approach represents a step forward in addressing the dynamic and evolving nature of SMS spam.

One key advantage of this hybrid model is its adaptability to new spam tactics, which often evolve to bypass traditional detection methods. While SVM and Naïve Bayes are effective in many cases, they struggle to maintain accuracy as spammers continuously refine their strategies. The hybrid model's improved performance opens up possibilities for even more advanced systems in the future. Researchers are now turning their attention to integrating additional data features, such as user behavior and contextual information, into these models. By analyzing patterns in user interaction with messages or incorporating metadata like time of day or message frequency, it is possible to build even more robust spam detection systems. This expanded approach could further elevate the accuracy of hybrid models, providing a more secure and user-friendly mobile experience.

ACKNOWLEDMENT

I would like to express our deepest gratitude to my mentor for their unwavering support and expert guidance during this research journey. Their knowledge and advice have been instrumental in shaping the direction of our study, offering clarity and helping me navigate challenges. Without their constructive feedback, I would not have been able to refine our analysis to the level it has reached today.

I would also like to acknowledge the importance of my peers, whose collaborative efforts have enhanced the quality of this research. Their diverse perspectives and suggestions provided me with valuable insights that helped me improve my approach. Moreover, the resources and tools made available by the research community were fundamental to the successful completion of my work. I am grateful for the opportunities to learn from others and engage in discussions that broadened my knowledge.

Lastly, I extend my appreciation to my family and friends for their patience and encouragement. Their unwavering belief in my potential provided the motivation I needed to persevere through the more challenging moments of the research process. This work would not have been possible without the collective support of everyone involved, and we are truly grateful for each individual who has contributed to this journey. Their influence and assistance have been deeply appreciated at every stage of the study

REFERENCES

[1]. Almeida, T. A., & Hidalgo, J. M. G. (2012). SMS Spam Collection. UCI Machine Learning Repository.

[2]. Zhang, Y., & Wallace, B. (2015). A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. arXiv preprint arXiv:1510.03820.

[3]. SCortes, C., &Vapnik, V. (1995). Support-Vector Networks. Machine Learning, 20(3), 273-297.

[4]. Hochreiter, S., &Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780.

[5]. Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1798-1828.

[6]. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

[7]. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, 27, 2672-2680.

[8]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., &Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.