iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: III Month of publication: March 2025

DOI: https://doi.org/10.22214/ijraset.2025.67164

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Spam or Ham? A Hybrid Deep Learning Approach for SMS Spam Detection

Srikanth Kandula[1], Charitha Sri Dondapati[2], Yesu Raju Veera[3], Geethika Chowdary Maguluri[4], Eunice Kukati[5]
[1]*Assistant Professor,* [2, 3, 4, 5]*Student, Department of Computer Science and Engineering, Dhanekula Institute of Engineering and Technology, Ganguru, AP, India*

Abstract: With the advent of digital communications, SMS spam has also become a widespread issue, which is inconvenient and even threatening to users. In this project, we advocate a hybrid spam detection model combining Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks and TF-IDF, and efficiently leverages deep learning and text-processing methods to identify spam messages. We train our model using the publicly available UCI SMS spam dataset. The interface gives an easy and convenient means of classifying SMS messages. Upon providing a message for classification, the model processes the message very quickly and provides a real-time classification result as either spam or ham. This project introduces a solution for SMS security through a useful spam detection system with excellent user experience. Following the integration of deep learning and conventional text-processing methods, our model has high accuracy and flexibility in spam detection. With this system implemented, we can mitigate risks from spam and ensure digital communication is safer and more reliable.

Keywords: SMS spam classification, Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), TF-IDF, Hybrid Model

I. INTRODUCTION

As mobile communication developed rapidly, SMS spam has also become a recurring problem, threatening both security and user frustration. Conventional filtering techniques, including keyword filtering, rule-based classifiers, and statistical models like Naive Bayes, usually fail because of their poor capability to comprehend context. Keyword-based filtering, specifically, is susceptible to spam evasion methods, including word obfuscation and adversarial text manipulation, and thus becomes less useful in practical applications. In addition, most traditional methods are based on hand-designed rules or bag-of-words feature representations, which do not appreciate the underlying meaning and natural structure of text.

To overcome these issues, in this paper, a hybrid deep learning model based on Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks is proposed, together with Term Frequency-Inverse Document Frequency (TF-IDF) for feature extraction. CNNs assist in extracting significant local features from text, whereas LSTMs are well-suited for learning sequential patterns, enabling the model to understand contextual subtleties better. As opposed to most classical machine learning methods involving a lot of manual labour for feature engineering, the method here automatically extracts features, thus resulting in improved accuracy and improved versatility to cope with various forms of spam.

By drawing on the respective strengths of LSTM and CNN, the new model overcomes the shortcomings of traditional filtering methods and provides a more accurate, adaptable, and reliable solution to SMS spam filtering. Experimental results show that this method performs better than traditional solutions and is thus a feasible and efficient solution for real-world spam filtering problems.

II. LITERATURE SURVEY

Spam message identification in Short Message Service (SMS) has become important as there is a rise in unwanted and fraudulent SMS. In a quest to improve spam categorization, researchers have examined many machine learning and deep learning techniques. The current review of literature describes an extensive summary of major works that have led the area, which indicates disparate methodologies and their effectiveness.

Menthe et al. [1] compared the use of machine learning algorithms such as decision trees, naive Bayes, and support vector machines (SVM) to identify spam. Their findings show that even though these algorithms are effective in distinguishing spam and non-spam messages, they require large-scale feature engineering in order to operate at their best. In a similar vein, Gawade et al. [3] had compared various classification approaches and highlighted the need for feature extraction methods such as Term Frequency-Inverse Document Frequency (TF-IDF) and word embeddings in order to improve classification accuracy.

Conversely, researchers resorted to deep learning models in order to address the shortcomings of the conventional machine learning. Gadde et al. [2] investigated the application of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks for SMS spam filtering. Their study demonstrated how CNN is specially suited to feature extraction and LSTM in retaining sequential patterns of words in the text, thereby enabling better classification performance. de Luna et al. [5] also explained how hybrid models with a combination of CNN and LSTM were beneficial as they produced superior results when compared to a single-model structure.

Aside from deep learning and machine learning, Gomez Hidalgo et al. [4] presented a content-based filter to detect spam in SMS messages. Their contribution emphasized the importance of text pre-processing and lexicon analysis for spam message detection and suggested that the integration of content-based filtering with machine learning can further enhance classification efficiency. Agarwal et al. [6] also addressed SMS spam detection in the Indian scenario, underlining the difficulty presented by linguistic diversity. Their research indicated the necessity of region-specific datasets to consider regional language differences to enhance model generalization.

Current research reveals that hybrid architectures that integrate CNN, LSTM, and feature extraction techniques such as TF-IDF can lead to a dramatic increase in spam classification accuracy. Although classical machine learning models are a good starting point, deep learning techniques have better performance in sequence learning and feature representation. Moreover, content-based filtering techniques improve spam filtering by improving text pre-processing techniques, making them an enhancement of current classification models.

III. PROPOSED METHODOLOGY

A. Working

As explained in the paper, SMS spam detection relies on a combination of CNN and LSTM networks and TF-IDF feature extraction. The suggested methodology consists of a sequence of pivotal steps, as shown below:

1) Dataset Selection and Pre-processing

The UCI SMS Spam Collection Dataset was utilized for training and selection. The dataset was pre-processed for text quality enhancement and noise elimination.

Data Cleaning: This consisted of the dropping of unnecessary columns, missing values handling, and the dropping of duplicate records.

Text Normalization: This involved making the text lowercase.

Tokenization: This is white-space tokenization of the text into various given words using NLTK.

Stop word Removal: This is the removal of all general English stop words having no contribution towards the classification.

Punctuation Removal: This process is the removal of all punctuation and special characters.

Stemming: The stemming technique reduces words to their root word: this paper used the Porter Stemmer algorithm.

2) TF-IDF for Feature Extraction

Feature extraction was performed with the Term Frequency-Inverse Document Frequency (TF-IDF) Vectorizer, which converts any text data into numerical form. TF-IDF, from its name, assigns weights to the words in a dataset based on its significance, thus enhancing classification accuracy.

3) Hybrid Model: CNN + LSTM for Classification

A hybrid deep learning model was constructed to classify messages as spam or ham through:

Text tokenization & Padding: The text was tokenized and the sequence was padded to a fixed length.

Embedding Layer: This layer converts the tokenized input into dense vector representation.

CNN Layer: A Conv1D layer extracts local features from text sequences. A MaxPooling1D layer is applied to reduce dimensionality and extract significant patterns.

LSTM Layer: Long dependencies in the sequence of text are learned through this layer that contributes to spam classification accuracy.

Fully Connected Layers: In this, a dense layer with ReLU activation is used for extracted features and then an output layer that does binary classification with sigmoid activation.

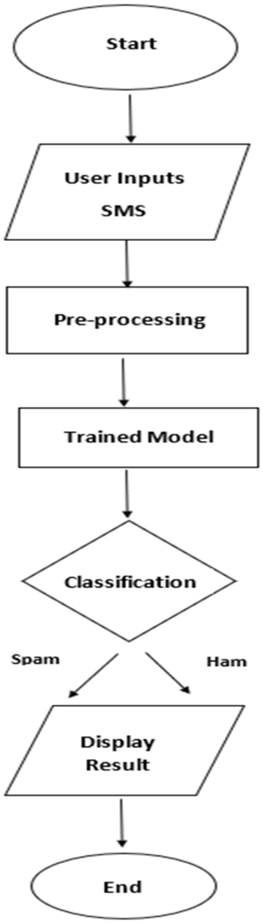
4) *Model Training and Evaluation*

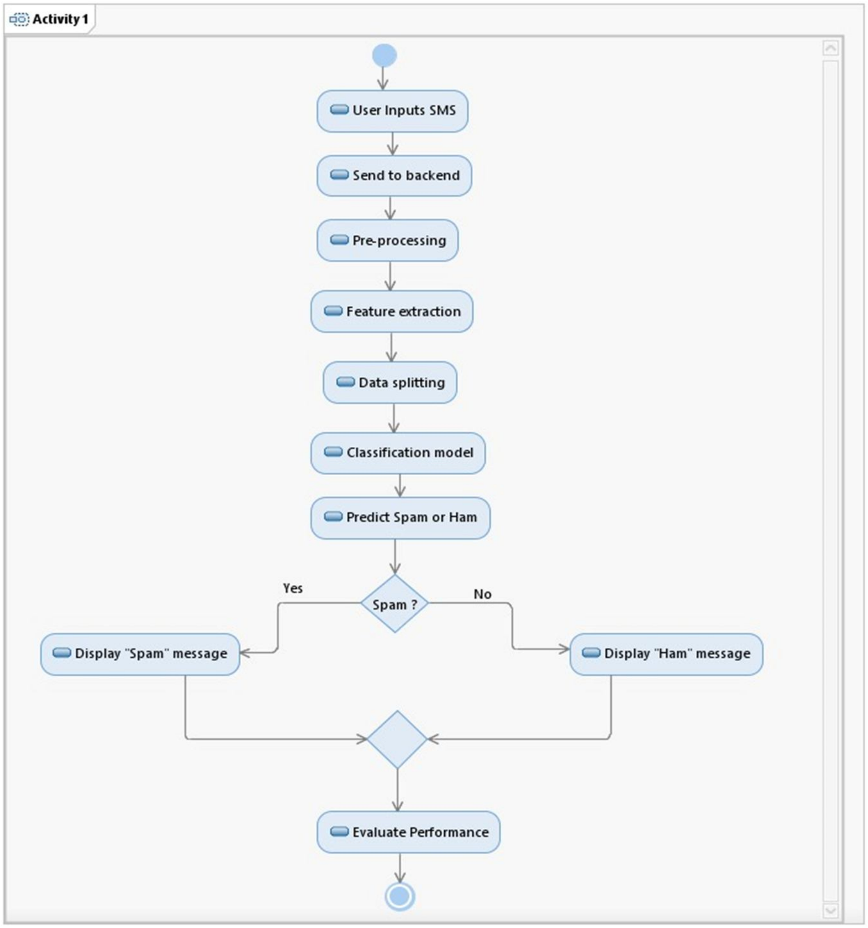
The data set was divided 80%-20% for training/testing. Training was conducted with Adam optimizer and binary cross-entropy loss function for 5 epochs and a batch size of 64. Accuracy and performance metrics were determined to test model efficiency. Model was serialized and saved for making it possible to deploy it in the future and along with fitted tokenizer.

a) Algorithms Used:

- TF-IDF for Feature Extraction: TF-IDF converts text to numerical features by approximating word importance in a document. TF-IDF assigns greater weights to rare but meaningful words and reduces the impact of frequent words. This improves the classification performance by separating spam-specific words. These are then transferred to the deep learning model.
- CNN: Local dependencies are simulated by CNN in text by the application of the filter to sequence words. Max Pooling is utilized to lessen dimensionality with no loss of important patterns. These are applied to the LSTM layer.
- LSTM for Sequence Learning: LSTM learns word long-term dependency and processes sequence text data. LSTM holds helpful past information by eliminating noise, and this improves spam classification precision. CNN-learned features are fed into LSTM in an attempt to capture context. This is a two-edged approach that boosts the robustness of spam detection.
- Fully Connected (Dense) Layer for Classification: The dense layer learns feature representations to distinguish between spam and ham messages and classifies them using ReLU and sigmoid activations. The last layer outputs a probability from 0 (ham) to 1 (spam). This provides precise message classification.
- Adam Optimizer & Binary Cross-Entropy Loss: Adam optimizer dynamically adjusts the learning rates for effective convergence during training. Binary cross-entropy loss approximates model predictions and reduces classification errors. The function is achieved with greater accuracy through weight updates optimization. They provide effective and stable model training.

The following figures represent the Flow chart and Activity diagram of the proposed system:





IV. PERFORMANCE EVALUATION

A. *Evaluation Metrics*

To evaluate the performance of our hybrid model (CNN + LSTM + TF-IDF) for spam message detection, we employed some important metrics that assist in measuring its accuracy and efficiency:

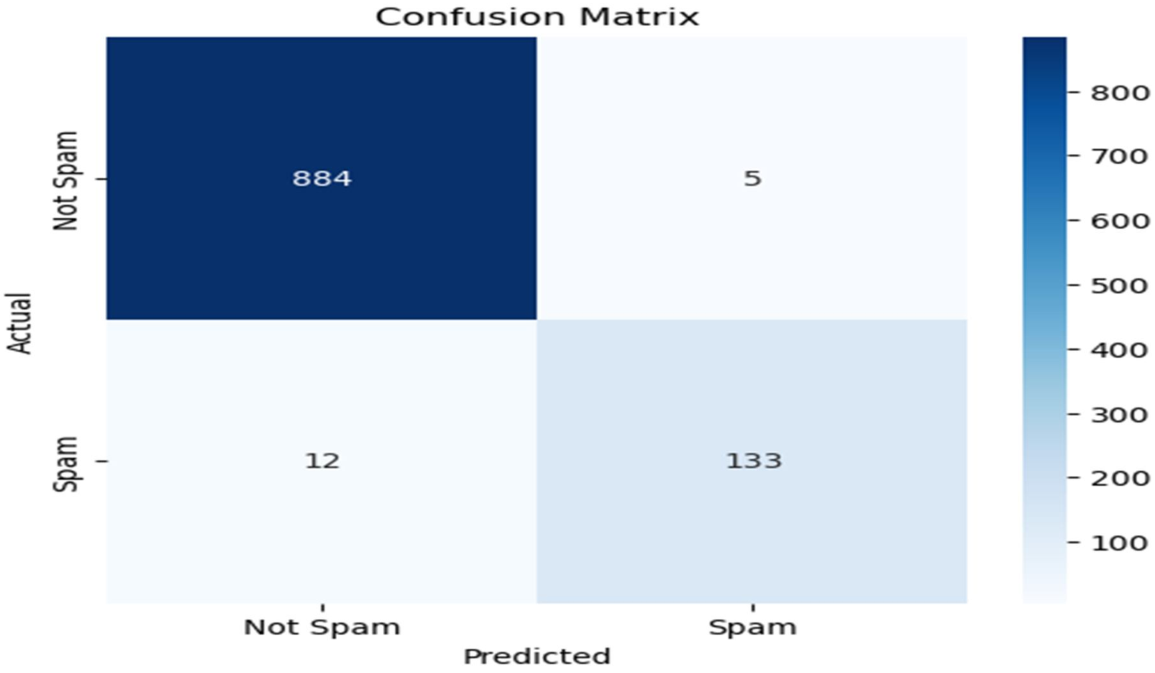
1) Accuracy: This informs us about how many SMS messages were labelled correctly in total, either as spam or not.
2) Precision: This shows the number of messages that the model had anticipated as spam that actually were spam. The higher the precision, the fewer false alarms.
3) Recall (Sensitivity): Depicts the potential of the model to identify real spam messages. High recall implies fewer spam messages are missed.
4) F1-Score: A harmonic metric that considers both precision and recall. It informs us regarding the overall consistency of the model in distinguishing spam from normal messages.
5) Confusion Matrix: A detailed breakdown of the performance of the model, showing how many correctly and mistakenly classified messages exist in each class (spam or not spam).

B. *Experimental Results*

After training and testing the model, we obtained the following results:

Metric	**Score**
Accuracy	98.36%
Precision	96.37%
Recall	91.72%
F1-Score	93.99%

1) *Confusion Matrix*



C. *Advantages of the Proposed System*

The hybrid system (CNN + LSTM + TF-IDF) has numerous advantages:

1) Improved Feature Representation – Both semantic and sequential patterns are captured to detect spam more effectively.
2) High Precision and Accuracy – Achieves 98.36% accuracy and 96.37% precision, lowering the number of false positives.
3) Effective Learning of Sequences – LSTM enhances the detection of context-based spam sequences.
4) Low False Alarms – Low misclassification of genuine messages.

V. FUTURE WORK

Although the suggested CNN + LSTM + TF-IDF model provides high accuracy in spam filtering, there are some points for improvement:

1) Multi-Language Support – Scaling the model to identify spam in various languages.
2) Real-time Deployment – Leverage the cloud or mobile devices for real-time filtering by deploying the model.
3) Hybrid Solutions – Research transformer-based models, for example, BERT or GPT, to provide favourable performance.

VI. CONCLUSION

This work proposes a hybrid deep learning model, which contains Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and Term Frequency-Inverse Document Frequency (TF-IDF) for SMS spam detection. The CNN is responsible for the extraction of spatial features while the LSTM is in charge of the sequential relation-ships extraction. In doing so it manages to learn about the patterns in the spam messages.

Experimental results demonstrate the model's high accuracy of 98.36%, with strong precision (96.37%) and recall (91.72%), confirming its reliability in detecting spam messages. The confusion matrix analysis shows minimal misclassifications, indicating the system's robustness. Given its efficiency and adaptability, this approach can be deployed in real-time SMS filtering applications to enhance mobile security.

REFERENCES

[1] S. Menthe, K. Rawal, M. Hirave, and A. J. Patil, "SMS Spam Detection Using Machine Learning," International Journal of Advanced Research in Computer and Communication Engineering, vol. 13, no. 3, Mar. 2024. https://www.researchgate.net/publication/379058545_SMS_SPAM_DETECTION_USING_MACHINE_LEARNING

[2] S. Gadde, A. Lakshmanarao, and S. Satyanarayana, "SMS Spam Detection using Machine Learning and Deep Learning Techniques," 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2021. https://ieeexplore.ieee.org/document/9441783

[3] A. L. Gawade, S. S. Shinde, S. G. Sawant, R. S. Chougule, and A. A. Mahaldar, "A Research Paper of SMS Spam Detection,"International Journal of Novel Research and Development, vol. 9, no. 3, Mar. 2024. https://www.ijnrd.org/papers/IJNRD2403165.pdf

[4] J. M. Gomez Hidalgo, G. C. Bringas, E. P. Sanz, and F. C. García, "Content-Based SMS Spam Filtering," in Proceedings of the 2006 ACM Symposium on Document Engineering, Amsterdam, The Netherlands, Oct. 2006. https://www.researchgate.net/publication/221353070_Content_based_SMS_spam_filtering

[5] R. G. de Luna et al., "A Machine Learning Approach for Efficient Spam Detection in Short Messaging System (SMS)," TENCON 2023 - 2023 IEEE Region 10 Conference (TENCON), Chiang Mai, Thailand, Oct. 2023. https://ieeexplore.ieee.org/document/10322491

[6] Sakshi Agarwal, Sanmeet Kaur, Sunita Garhwal, "SMS spam detection for Indian messages" ,2015 1st International Conference on Next Generation Computing Technologies (NGCT), Dehradun, India, Sept. 2015. https://ieeexplore.ieee.org/document/7375198