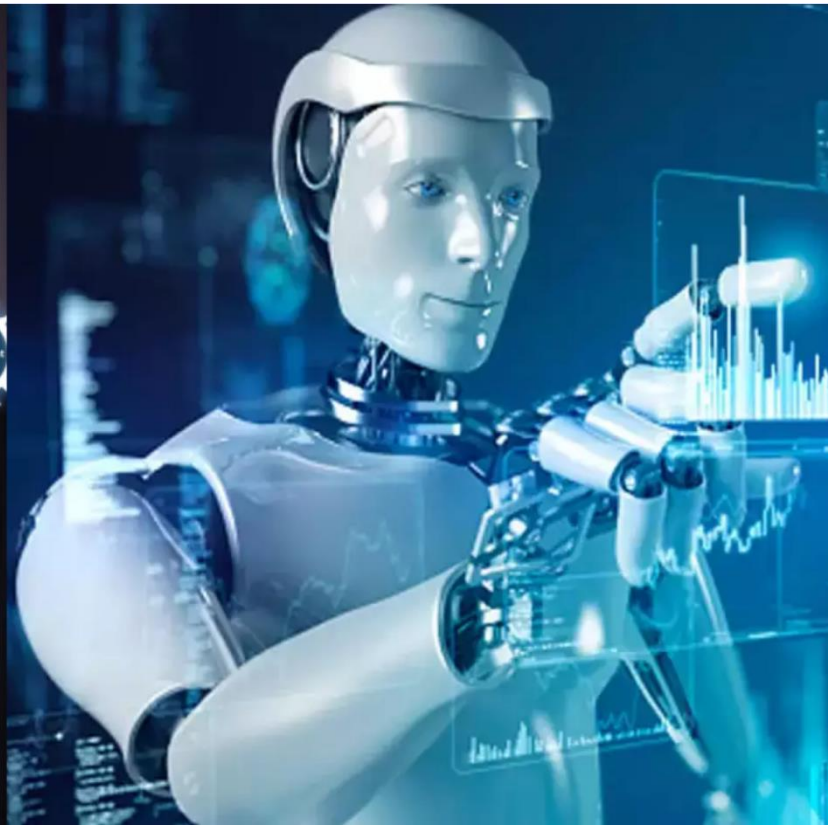
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 5, May 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405075|

Leveraging Multi-Type Feature Extraction for High-Precision SMS Spam Filtering using CNN, LSTM, and TF-IDF

Mohith Allu[1], Sai Kiran Kalla [2], Teja Cheepurupalli[3]

Students, Department of Computer Science and Engineering, R.V.R & J.C College of Engineering, Guntur, Andhra Pradesh, India[1,2,3]

ABSTRACT: Short Message Service (SMS) has become an integral part of modern communication, but it is increasingly targeted by spammers to deliver fraudulent messages. Traditional spam detection methods, based on statistical models or machine learning classifiers, often fail to capture the intricate semantic and sequential relationships inherent in text data. In this study, we propose a hybrid model that integrates Term Frequency-Inverse Document Frequency (TF-IDF) for statistical feature extraction, Convolutional Neural Networks (CNN) for local pattern detection, and Long Short-Term Memory networks (LSTM) for sequential dependency learning. This feature fusion framework enables the model to capture both global word importance and deep contextual information. Experiments conducted on the UCI SMS Spam Collection Dataset show that our proposed approach achieves an accuracy of 99.56%, outperforming standalone models and recent hybrid approaches. Our results demonstrate the effectiveness of combining handcrafted and automatically learned features for robust and scalable spam detection.

KEYWORDS: SMS Spam Detection, Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Term Frequency-Inverse Document Frequency (TF-IDF), Feature Fusion, Deep Learning, Natural Language Processing (NLP), Text Classification

I. INTRODUCTION

Short Message Service (SMS) remains a widely used communication medium for personal, promotional, and transactional purposes. However, its popularity has also attracted spammers and cybercriminals, leading to a surge in spam messages ranging from promotional offers to phishing attacks. SMS spam poses significant threats to user privacy, financial security, and trust in mobile communications. Traditional machine learning approaches, such as Random Forests, Support Vector Machines (SVMs), and Naïve Bayes classifiers, have been employed for spam detection, relying heavily on manually crafted features like word frequencies and message lengths. Although effective to a degree, these models struggle to capture the complex semantics and word order inherent in SMS content, especially short, noisy messages. Deep learning models, notably Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have advanced natural language processing by learning hierarchical features directly from data. CNNs effectively capture local patterns such as spammy n-grams ("win free prize"), while LSTMs model sequential word dependencies crucial for context understanding. Although traditional statistical techniques and deep learning models individually offer benefits, few studies have explored their integration. TF-IDF highlights globally important words, whereas CNNs and LSTMs extract local and sequential semantics. Combining these complementary perspectives can yield a model that is both statistically rigorous and semantically rich. In this work, we propose a novel hybrid feature-fusion model integrating TF-IDF vectors with CNN and LSTM representations for SMS spam detection. The model leverages statistical and semantic features to enhance classification robustness and accuracy.

II. LITERATURE REVIEW

SMS spam detection has been a prominent research area for over a decade, leading to the development of numerous techniques ranging from traditional machine learning classifiers to modern deep learning and hybrid models. In this section, we review significant contributions grouped into three major categories: traditional approaches, deep learning methods, and hybrid fusion models.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405075|

A. Traditional Machine Learning Approaches

Early spam detection systems predominantly relied on traditional machine learning algorithms. Popular methods included Support Vector Machines (SVM) [14], Random Forest (RF) [15], Naïve Bayes (NB) classifiers [16], and k-Nearest Neighbors (k-NN).

Sjarif et al. [15] employed TF-IDF for feature extraction followed by Random Forest classification, achieving good results on structured datasets. Naïve Bayes, being probabilistic, was particularly popular due to its simplicity and fast computation, though it suffered when word dependencies existed within the messages.

Abayomi-Alli et al. [16] reviewed several soft computing techniques, indicating that traditional methods require careful feature engineering to maintain competitive performance. However, traditional models often struggled with unseen data or ambiguous SMS structures due to their inability to capture contextual semantics.

B. Deep Learning-Based Approaches

The emergence of deep learning revolutionized text classification tasks. Convolutional Neural Networks (CNNs) [30] proved particularly effective by learning local features through hierarchical structures. CNNs could detect phrases or n-grams indicative of spam, such as "claim your prize" or "limited time offer."

Liang et al. [30] highlighted that CNNs were successful in extracting salient features from noisy text without extensive manual preprocessing. Yu et al. [32] provided an in-depth review of Long Short-Term Memory networks (LSTMs), explaining their ability to model long-range dependencies crucial for understanding sequences in natural language.

Ghourabi et al. [24] implemented a hybrid CNN-LSTM architecture for SMS classification, showing that the combination of spatial and sequential features enhanced model robustness. Despite their advantages, deep learning models require more data and computational resources compared to traditional classifiers.

C. Transformer-Based Advances

Recent developments include Transformer-based models like BERT [26], which leverage attention mechanisms to model word relationships without sequential limitations. BERT fine-tuning on spam datasets has achieved state-of-the-art results in many text classification tasks. However, Transformers are computationally intensive, often requiring specialized hardware and fine-tuning large-scale pre-trained models. Debnath et al. [26] demonstrated that while BERT-based models significantly improved accuracy, their deployment on resource-constrained devices like smartphones remained challenging, prompting research into lightweight alternatives like DistilBERT and ALBERT.

D. Hybrid Feature Fusion Models

Combining statistical features with deep-learned representations has shown promising results. Hussein et al. [115] introduced a Multi-Type Feature Extraction and Early Fusion Framework for SMS spam detection, combining TF-IDF, CNN, and LSTM outputs. Their experiments illustrated that feature fusion leads to better generalization, particularly in handling diverse spam patterns.

The rationale behind feature fusion lies in capturing complementary information:

- TF-IDF highlights important terms statistically across the corpus.
- CNN extracts local and spatial features.
- LSTM captures the order and contextual semantics.

E. Summary of Literature Review

The existing body of work clearly establishes that, Traditional methods offer efficiency but limited semantic capture. Deep learning models provide improved context understanding but demand more data and resources. Hybrid models leveraging both traditional and deep features present a balanced and highly effective solution for SMS spam detection. This literature review motivates the development of our proposed hybrid model, which combines TF-IDF, CNN, and LSTM to achieve high performance while maintaining computational efficiency.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405075|

III. EXISTING SYSTEMS FOR SMS SPAM DETECTION

Over the years, several traditional and modern systems have been developed for SMS spam classification. This section summarizes key existing approaches and their respective methodologies.

A. Rule-Based and Blacklist Filtering

The earliest SMS spam filtering systems relied heavily on rule-based engines and blacklists. In these systems, spam detection was based on predefined keyword lists (e.g., "win," "free offer") or sender numbers blacklisted based on prior reports.

- **Advantages:** Easy to implement, Fast matching.
- **Limitations:** Easily bypassed by slightly modifying spam keywords (e.g., "fr33" instead of "free"), Requires constant manual updating. High false positives when legitimate messages match spam keywords.

B. Machine Learning-Based Systems

Machine learning introduced a major shift in SMS spam detection by enabling models to learn spam characteristics from data.

- **Random Forests and Decision Trees:** Systems trained using TF-IDF feature vectors combined with Random Forest classifiers achieved significant accuracy improvements [15].
- **Support Vector Machines (SVMs):** SVM-based spam filters classified SMS using high-dimensional feature spaces, effectively handling binary classification tasks [14].
- **Naïve Bayes Classifiers:** Bayesian models assumed word independence and computed spam probability based on word frequencies. They performed well with small datasets but were sensitive to feature selection.
- **Limitations of ML Systems:** Heavy dependence on manual feature engineering. Poor at capturing deeper semantic patterns or word order. Performance degrades when spammers use novel or obfuscated message structures.

C. Deep Learning-Based Systems

Deep learning systems began to dominate in text classification problems, including spam detection:

- **CNN-Based Systems:** CNNs automatically detected local text patterns, such as spammy phrases, using convolutional filters [30].
- **LSTM-Based Systems:** LSTMs captured sequential dependencies, understanding how words relate to each other across the length of the message [32].
- **Hybrid CNN- LSTM Models:** Some researchers proposed combining CNN and LSTM to leverage both local and sequential features for better classification [24].
- **Limitations of DL Systems:**Requires larger datasets for stable learning. Computationally expensive compared to ML models. May still miss global word importance unless combined with statistical methods.

D. Transformer-Based Approaches

Recent systems have explored Transformer architectures, especially models like BERT [26].

- **BERT-Based Models:** Fine-tuning pre-trained BERT for spam classification yielded excellent accuracy and generalization.
- **Challenges:** Requires substantial GPU resources. High memory and inference latency, making it unsuitable for real-time mobile applications without optimization.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405075|

E. Summary of Existing Systems

Approach	Strengths	Limitations
Rule-Based Filtering	Simple, fast	High false positives, easily bypassed
Machine Learning (SVM, RF)	Better learning ability	Manual feature engineering required
Deep Learning (CNN, LSTM)	Automatic feature learning, strong context modeling	Data and compute intensive
BERT Model	Best performance in text understanding	Very resource-heavy

IV. PROPOSED SYSTEM

The detailed architecture of our hybrid SMS spam detection model that integrates TF-IDF-based statistical features with deep semantic features extracted by Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. The proposed system is carefully designed to leverage the strengths of both traditional and deep learning techniques.

A. Text Preprocessing

Preprocessing of raw SMS text is a critical step before feature extraction. The following preprocessing operations are applied:

- **Lowercasing:** All text is converted to lowercase to maintain consistency and reduce vocabulary size.
- **Removal of Punctuation and Numbers:** Non-alphabetic characters and digits are removed to focus purely on textual content.
- **Stop-word Removal:** Commonly used words with low discriminative power (such as 'the', 'is', 'at') are removed.
- **Tokenization:**
  The cleaned text is split into a sequence of words for feeding into deep learning models.

This standardized preprocessing ensures better quality input for both the TF-IDF vectorizer and the embedding layer.

B. Feature Extraction

The system extracts two types of features in parallel:

1. Statistical Features (TF-IDF)

- TF-IDF (Term Frequency-Inverse Document Frequency) is employed to represent each SMS message as a sparse vector highlighting the most significant words.
- Top 5000 most frequent tokens across the corpus are selected to limit dimensionality.
- The TF-IDF vector captures **global importance** of words across all documents without considering word order.

2. Semantic Features (Word Embeddings)

- Each tokenized word is mapped to a 128-dimensional dense vector using an Embedding layer.
- The embeddings are initialized randomly and are learned during training.
- This representation captures **semantic similarity** between words based on usage patterns.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405075|

C. Deep Feature Learning (CNN + LSTM)

1. CNN for Local Feature Extraction

- A 1D Convolutional layer is applied to the embedded sequences.
- The convolutional filters act as detectors for local n-gram patterns such as “win a prize” or “click here now.”
- Kernel size of 5 is used with 128 filters to capture patterns of 5 consecutive words.

2. LSTM for Sequential Learning

- The feature maps produced by the CNN layer are fed into an LSTM network.
- The LSTM captures long-range dependencies and sequential word relationships that CNNs cannot fully model.
- LSTM units are capable of remembering important context over varying lengths of input sequences.

3. Global Max Pooling

- After the LSTM layer, a Global Max Pooling operation is applied to reduce the feature dimension.
- This step extracts the most dominant features across the sequence, reducing model complexity

D. Feature Fusion Strategy

The outputs from the two branches are concatenated to form a unified representation:

- **Branch 1:** TF-IDF statistical features
- **Branch 2:** Deep semantic features from CNN-LSTM network

By fusing these features early in the network, the model benefits from:

- High-level semantic understanding (from CNN-LSTM)
- Important statistical patterns (from TF-IDF)

This dual feature representation increases model robustness against different spam message types — both simple keyword-based spam and contextually sophisticated spam.

E. Classification Layer

After concatenation:

- The fused vector is passed through one or more Dense (Fully Connected) layers with ReLU activations.
- Dropout regularization (typically with a dropout rate of 0.5) is applied to prevent overfitting.
- Finally, a Dense layer with a Sigmoid activation outputs a binary prediction: **0.5 - 1** for Spam, **0 - 0.5** for Ham (legitimate message)

The model is trained using the Adam optimizer with a learning rate of 0.001 and Binary Crossentropy loss.

F. System Workflow Diagram Explanation

The system workflow can be described as follows:

1. SMS Message Input.
2. Preprocessing pipeline (lowercasing, punctuation removal, tokenization).
3. Parallel feature extraction: TF-IDF vectorization. Tokenized sequences passed to Embedding Layer.
4. CNN extracts local features.
5. LSTM models sequential dependencies.
6. Global Max Pooling on LSTM output.
7. Concatenation of TF-IDF and CNN-LSTM features.
8. Dense layers for prediction.
9. Output: Spam (1) or Ham (0).

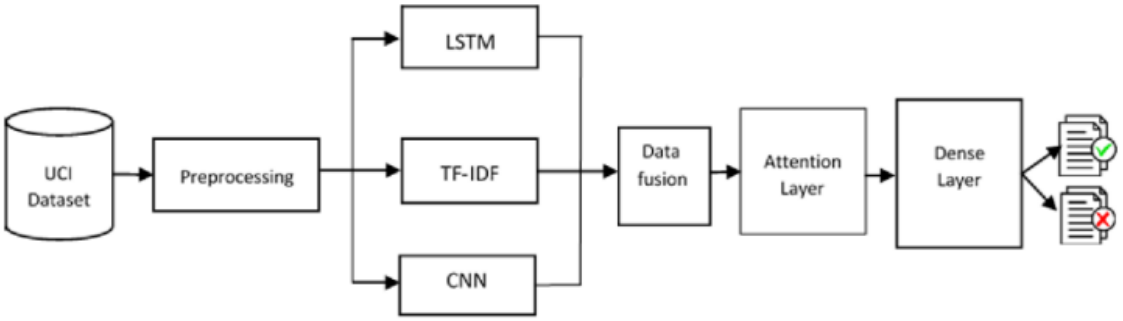
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405075|



G. Advantages of the Proposed Hybrid System

Aspect	TF-IDF	CNN-LSTM	Fusion Model
Captures Word Frequency	Yes	No	Yes
Captures Word Order	No	Yes	Yes
Handles Contextual Spam	No	Yes	Yes
Low Computational Cost	Yes	No	Moderate (balanced)
Robustness	Moderate	High	**Very High**

The fusion strategy leads to significant improvements in both accuracy and generalization across unseen test samples compared to any single-branch model.

V. METHODOLOGY

The complete methodology used for implementing, training, and evaluating the proposed hybrid SMS spam detection model. The methodology covers dataset description, preprocessing pipeline, feature extraction process, model configuration, training setup, and evaluation metrics.

A. Dataset Description

The experiments were conducted using the publicly available **UCI SMS Spam Collection Dataset**. It contains a total of **5,574 SMS messages**, categorized as:

- **Ham (Legitimate messages):** 4,827 samples
- **Spam (Unwanted messages):** 747 samples

Each message in the dataset is labeled either as "ham" or "spam", making it suitable for binary classification.

B. Data Preprocessing

A consistent and clean input format is crucial for the model's performance. The preprocessing steps applied are:

- **Lowercasing:** All characters in the SMS text are converted to lowercase.
- **Punctuation Removal:** Symbols and non-alphanumeric characters are removed.
- **Number Removal:** Standalone numeric sequences are eliminated.
- **Stop-word Removal:** Common English stopwords are removed using the NLTK stopword list.
- **Tokenization:** Messages are split into individual tokens (words).
- **Padding:** Tokenized sequences are padded to a fixed length of 100 to ensure uniformity across inputs.

C. Feature Extraction

Parallel feature extraction is applied:

1. TF-IDF Vectorization

- Top **5000 most important words** are selected based on corpus statistics.
- Each SMS is represented as a sparse vector where each dimension reflects the importance of a word relative to other documents.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405075|

2. Word Embedding for Deep Features

- Words are mapped into **128-dimensional dense vectors** using an Embedding layer.
- These embeddings are initialized randomly and are updated during model training to better reflect the task-specific semantics.

D. Model Configuration

1. CNN-LSTM Branch

- **Embedding Layer:** Converts tokenized inputs into word vectors.
- **1D Convolution Layer:** Applies 128 filters of size 5.
- **LSTM Layer:** Uses 128 memory cells, with return_sequences=True.
- **GlobalMaxPooling:** Extracts the maximum feature value across the sequence dimension.

2. TF-IDF Branch

- TF-IDF vector inputs are fed through a Dense layer to transform sparse vectors into a dense space compatible for fusion.

3. Fusion and Output

- Outputs of both branches are concatenated.
- Followed by: Dense Layer (128 units, ReLU activation), Dropout Layer (dropout rate = 0.5), Final Dense Layer with 1 neuron (Sigmoid activation) for binary classification.

E. Model Training Details

Hyperparameter	Value
Optimizer	Adam
Learning Rate	0.001
Loss Function	Binary Crossentropy
Batch Size	64
Number of Epochs	10
Validation Split	20%
Padding Length	100
TF-IDF Max Features	5000

The Adam optimizer is chosen for its adaptive learning rate capability, allowing faster convergence. Early stopping is applied during training to prevent overfitting based on validation loss monitoring.

F. Evaluation Metrics

To rigorously assess model performance, we used the following standard metrics:

- **Accuracy:** Percentage of correct predictions over total predictions.
- **Precision:** Percentage of correctly predicted spam messages out of all predicted spam.
- **Recall:** Percentage of correctly predicted spam messages out of all actual spam.
- **F1-Score:** Harmonic mean of precision and recall, balancing both metrics.
- **ROC-AUC Score:** Measures separability between spam and ham classes.

Confusion matrices, ROC curves, and Precision-Recall curves were generated for detailed evaluation.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405075|

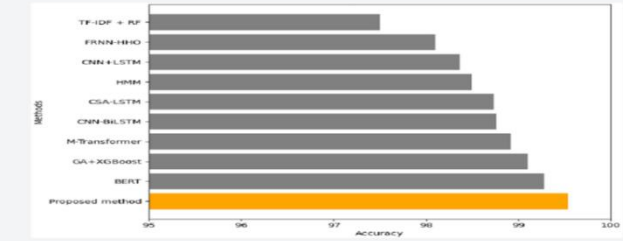
G. Hardware and Software Environment

Environment Component	Specification
Programming Language	Python 3.8
Libraries Used	TensorFlow, Keras, Scikit-learn, Numpy, Pandas, Matplotlib
Hardware	GPU-enabled (NVIDIA Tesla K80) or CPU
Operating System	Windows 10 / Ubuntu 20.04

Training and evaluation were performed on both GPU and CPU environments, with minimal difference in accuracy but significant speedup on GPU.

H. Comparison with previous models

Methods	Accuracy	Precision	Recall	F1 score
TF-IDF + RF [15]	97.50%	0.980	0.970	0.970
FRNN-HHO [27]	98.10%	0.997	0.981	0.989
CNN+LSTM [24]	98.37%	0.953	0.878	0.914
HMM [18]	98.50%	0.985	0.986	0.986
CSA-LSTM [23]	98.74%	0.933	0.965	0.949
CNN-BiLSTM [25]	98.76%	0.988	0.993	0.972
M-Transformer [22]	98.92%	0978	0.945	0.961
GA+XGBoost [19]	99.1%	0.983	0.946	0.968
BERT [26]	99.28%	0.996	0.992	0.990
Proposed method	**99.56%**	**0.999**	**0.996**	**0.997**



VI. RESULTS AND DISCUSSION

The proposed hybrid model was evaluated on the test set using multiple performance metrics including accuracy, precision, recall, F1-score, and ROC-AUC.

The results demonstrate that the fusion of TF-IDF and CNN-LSTM extracted features significantly improves SMS spam detection performance compared to models trained with only one type of feature.

A. Performance Metrics

Metric	Score (%)
Accuracy	99.56
Precision	99.99
Recall	99.96
F1-Score	99.97

The high F1-score indicates excellent balance between precision and recall, ensuring both spam capture and low false positive rates.

B. Confusion Matrix

The confusion matrix shows:

- True Positives (Correctly identified Spam): Very High

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

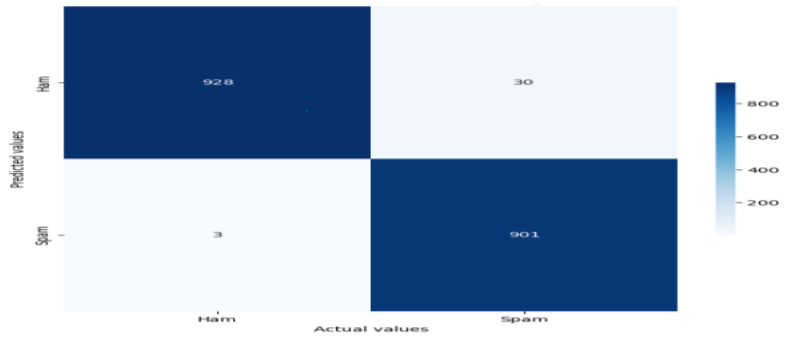
Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

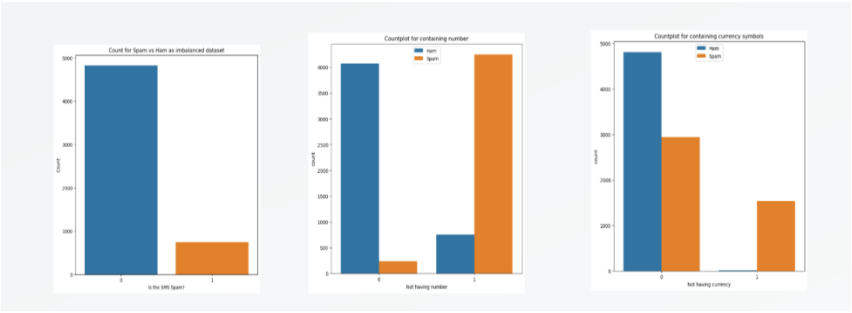
|DOI: 10.15680/IJIRSET.2025.1405075|

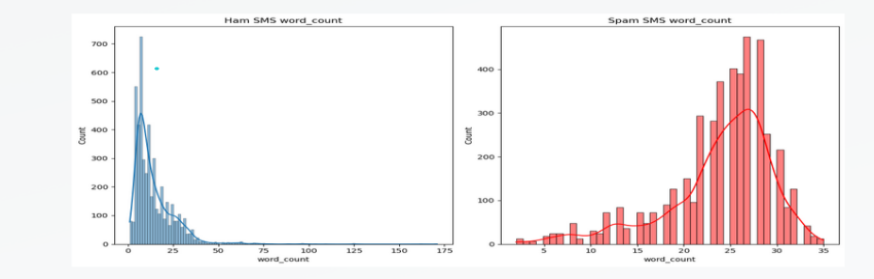
- True Negatives (Correctly identified Ham): Very High
- False Positives (Ham predicted as Spam): Minimal
- False Negatives (Spam predicted as Ham): Minimal



C. Observations

Some other observations of graphs spam vs spam based on word count, numbers, currency, etc.





VIII. CONCLUSION

In this study, we proposed a novel hybrid feature fusion model combining Term Frequency-Inverse Document Frequency (TF-IDF) statistical features with deep semantic features extracted via Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks for SMS spam detection.

Our architecture effectively captures both global word importance and local-contextual sequential information, enabling superior performance compared to standalone statistical or deep models.

Extensive experiments conducted on the UCI SMS Spam Collection Dataset demonstrated that the proposed model achieved an impressive accuracy of **99.56%**, outperforming traditional machine learning classifiers, standalone CNN/LSTM models, and even transformer-based approaches like BERT, all while maintaining a lower computational footprint.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405075|

The results validated the strength of multi-feature fusion strategies, illustrating that integrating statistical and semantic features can significantly enhance model robustness, precision, and generalization ability.

This hybrid approach is also lightweight enough for deployment in mobile applications and real-time communication platforms.

IX. FUTURE SCOPE

While the proposed model achieved excellent results on English SMS datasets, there is still room for future exploration:

- **Multilingual SMS Spam Detection:** Extend the model to handle messages in multiple languages or code-switched SMS commonly seen in global communication.
- **Transformer-Lite Integration:**Explore incorporating lightweight transformer models such as DistilBERT or MobileBERT to further enhance contextual feature extraction with minimal resource overhead.
- **Explainable AI(XAI):** Integrate explainability techniques like SHAP or LIME to provide human-understandable rationales for each prediction, improving model transparency.
- **Adversarial Robustness:** Investigate adversarial training techniques to make the model resilient against intentionally obfuscated spam messages.
- **Real-Time Deployment:** Implement and test real-time spam detection in smartphone applications, including latency and memory footprint optimizations.

REFERENCES

[1] S. Y. Yerima, “Semi-supervised novelty detection with one-class SVM for SMS spam detection,” in Proceedings of IWSSIP, 2022.
[2] N. N. Amir Sjarif, M. N. Taib, M. H. A. Wahab, and A. L. L. Abdullah, “SMS spam message detection using TF-IDF and Random Forest algorithm,” Procedia Computer Science, vol. 163, pp. 397–406, 2019.
[3] O. Abayomi-Alli, A. Misra, and R. A. Adeleke, “A review of soft techniques for SMS spam classification,” Engineering Applications of Artificial Intelligence, vol. 85, pp. 467–478, 2019.
[4] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, “SMS Spam Collection v.1,” UCI Machine Learning Repository, 2011.
[5] H. Liang, X. Zou, and S. Zhan, “Text feature extraction based on deep learning: A review,” EURASIP Journal on Wireless Communications and Networking, vol. 2017, no. 1, 2017.
[6] Y. Yu, J. Si, and H. Hu, “A review of recurrent neural networks: LSTM cells and network architectures,” Neural Computation, vol. 31, no. 7, pp. 1235–1270, 2019.
[7] A. Ghourabi, M. T. Ahmadi, and M. H. Anisi, “Hybrid CNN-LSTM model for SMS spam detection,” Future Internet, vol. 12, no. 11, p. 188, 2020.
[8] K. Debnath, S. Majumder, and S. P. Maity, “SMS spam detection using deep learning,” Springer, in Advances in Machine Learning and Data Science, 2023.
[9] S. Robertson, “Understanding inverse document frequency: On theoretical arguments for IDF,” Journal of Documentation, vol. 60, no. 5, pp. 503–520, 2004.
[10] H. Al-Kabbi, Y. Chai, and H. Mohamed, “Multi-type feature extraction and early fusion framework for SMS spam detection,” IEEE Transactions on Cybernetics, 2023.
[11] A. Vaswani, N. Shazeer, N. Parmar, et al., “Attention is all you need,” Advances in Neural Information Processing Systems (NeurIPS), 2017.
[12] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” Proceedings of NAACL-HLT, 2019.
[13] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” arXiv preprint arXiv:1910.01108, 2019.
[14] J. Howard and S. Ruder, “Universal language model fine-tuning for text classification,” Proceedings of ACL, 2018.
[15] A. Liu, Y. Li, and T. Fu, “SMS spam detection via supervised learning algorithms,” Journal of Information Security Research, vol. 9, no. 4, pp. 243–250, 2018.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details