Advances in Consumer Research
https://acr-journal.com/
Volume-3 | Issue-2 | Feb 2026
Original Researcher Article

A Robust SMS Spam Detection Framework Using Transformer-Based Learning and Real-Time Web Deployment

Mr. P. AnandhaKumar[1], V. Nandhini[2]
[1]Head of the Department, Department of Electronics and Communication Engineering, V.S.B Engineering College, Karur-639111
Email ID : ananthpece@gmail.com
[2]Department of Applied Electronics VSB Engineering College, Karur-639111.
Email ID : nandhujeya35@gmail.com

ABSTRACT
Short Message Service (SMS) spam continues to pose a significant threat to user privacy, security, and trust in mobile communication systems. Traditional rule-based and classical machine learning approaches rely heavily on surface- level lexical features, which makes them vulnerable to obfuscation and evolving spam strategies. In this paper, we propose a robust SMS spam detection framework based on a fine-tuned RoBERTa transformer model that captures deep contextual semantics of short text messages.
To validate the effectiveness of the proposed approach, we con- duct a comparative evaluation against classical baseline models, including Naive Bayes, Support Vector Machines, and Logistic Regression, using the publicly available SMS Spam Collection dataset. Experimental results demonstrate that the transformer- based model achieves near-perfect performance on a balanced evaluation subset, significantly outperforming traditional classi- fiers.
Furthermore, we present a real-time web-based deployment of the proposed system using a Streamlit interface, enabling interactive and user-friendly spam detection. All components of the framework, including the source code, trained model, and live demonstration, are publicly released to support reproducibility. This work bridges the gap between state-of-the-art natural lan- guage processing techniques and practical spam filtering systems.
Keywords: SMS Spam Detection, Natural Language Pro- cessing, Transformer Models, RoBERTa, Text Classification, Deep Learning, Web Deployment, Streamlit.

1. INTRODUCTION:

Short Message Service (SMS) remains one of the most widely used communication channels due to its simplicity, low cost, and accessibility. However, the increasing volume of unsolicited and malicious spam messages has significantly degraded user experience and raised concerns regarding fraud, privacy, and cybersecurity. Spam messages often include deceptive advertisements, phishing attempts, and misleading links, making automatic detection an essential requirement for modern communication platforms.

Early spam filtering techniques relied on manually designed rules and keyword-based heuristics. Although computationally efficient, such systems fail to generalize to new and obfuscated spam patterns. Classical machine learning models, such as Naive Bayes, Support Vector Machines (SVM), and Logistic Regression, introduced statistical learning capabilities and

improved adaptability. Nevertheless, these methods typically depend on handcrafted features such as bag-of-words or TF- IDF representations, which lack deep semantic understanding. Recent advances in transformer-based architectures have revolutionized natural language processing tasks. Models such as BERT and RoBERTa leverage self-attention mechanisms to capture long-range dependencies and contextual semantics. These properties make transformer models particularly suitable for short-text classification tasks such as SMS spam detection, where subtle linguistic cues often differentiate legitimate and

malicious messages.

In this work, we propose a robust SMS spam detection framework based on a fine-tuned RoBERTa model. Unlike traditional approaches, our method learns contextual representations directly from raw text, enabling improved generalization to semantically complex spam messages. To ensure comprehensive evaluation, we compare the proposed approach with classical baseline models.

Beyond model performance, practical usability is a critical factor for real-world adoption. Therefore, we integrate the trained classifier into a real-time, interactive web application built using Streamlit. The complete system, including source code and trained models, is publicly released to support reproducibility and future research.

CONTRIBUTIONS

The main contributions of this paper are summarized as follows:

We propose a transformer-based SMS spam detection framework using a fine-tuned RoBERTa model that captures deep contextual semantics.

We conduct a comprehensive comparative evaluation against classical machine learning baselines, including Naive Bayes, Support Vector Machines, and Logistic Regression.

We present an extensive experimental analysis using standard performance metrics such as accuracy, precision, recall, and F1-score.

We deploy the proposed system through a real-time web interface using Streamlit, enhancing accessibility and usability.

We release all resources publicly, including source code, trained models, and a live demo, to support reproducibility.

DATASET DESCRIPTION

We evaluate the proposed framework using the publicly available SMS Spam Collection dataset [1], which is widely used for benchmarking spam detection systems. The dataset contains 5,572 real-world SMS messages labeled as either legitimate (ham) or spam.

The original dataset exhibits a significant class imbalance, with approximately 86.5% ham messages and 13.5% spam messages. Such imbalance can bias learning algorithms toward the majority class. To mitigate this issue, we apply controlled oversampling of the minority class during training.

After preprocessing and cleaning, a total of 5,149 messages remain. The dataset is then balanced to obtain 5,000 samples per class, resulting in 10,000 messages. A stratified split is used, with 9,200 samples for training and 800 samples for testing.

A. Exploratory Data Analysis

Prior to model training, we perform exploratory data analysis to understand the characteristics of the dataset. The label distribution reveals a strong imbalance between legitimate and spam messages. Word frequency analysis shows that spam messages commonly contain promotional and action-oriented terms, while ham messages are dominated by conversational vocabulary.

We further analyze the distribution of message lengths. Spam messages tend to be longer on average and exhibit higher variance, motivating the use of models capable of capturing long-range dependencies.

2. METHODOLOGY

This section describes the proposed SMS spam detection framework, including preprocessing, baseline models, and the transformer-based classifier.

Preprocessing

Each SMS message undergoes a series of preprocessing steps, including lowercasing, removal of special characters, normalization of whitespace, and masking of URLs and phone numbers.

For classical machine learning models, the preprocessed text is converted into numerical feature vectors using the TF-IDF representation. For the transformer-based model, raw text is tokenized using the RoBERTa tokenizer, which applies byte- pair encoding and inserts special tokens.

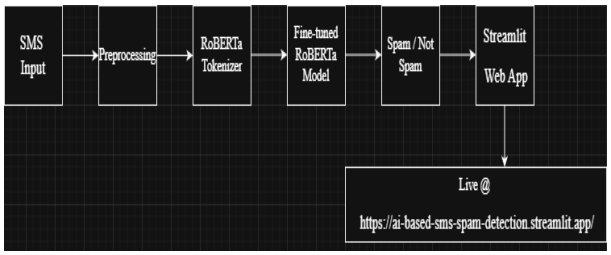
Baseline Models

We evaluate three classical machine learning models as baselines:

Naive Bayes (NB): A probabilistic classifier based on Bayes' theorem with independence assumptions.

Support Vector Machine (SVM): A discriminative classifier that constructs a hyperplane to maximize class separation.

Fig. 1. Proposed system architecture.



Logistic Regression (LR): A linear classifier that models the probability of class membership using a sigmoid function.

These baselines serve as reference points to highlight the advantages of transformer-based learning.

Proposed Transformer-Based Model

The core of the proposed framework is a fine-tuned RoBERTa model [3]. RoBERTa is an optimized variant of BERT that improves training efficiency by using dynamic masking and larger batch sizes.

The input SMS message is tokenized and passed through multiple transformer encoder layers. The final hidden state corresponding to the classification token is fed into a fully connected layer with a softmax activation to produce class probabilities:

$$\hat{y} = \text{softmax}(W \cdot h_{cls} + b) \quad (1)$$

The model is fine-tuned using cross-entropy loss and optimized with AdamW.

SYSTEM ARCHITECTURE

The overall architecture of the proposed SMS spam detection framework is illustrated in Fig. 1. The system is designed as a modular pipeline consisting of preprocessing, tokenization, classification, and deployment components.

Initially, raw SMS messages are passed to the preprocessing module, where normalization and noise removal are per- formed. The cleaned text is then tokenized using the RoBERTa tokenizer, which converts the input into subword tokens.

These tokens are fed into the fine-tuned RoBERTa classifier, which computes contextual embeddings through multiple transformer encoder layers. The output

representation is passed to a classification head that predicts whether the message is spam or legitimate.

Finally, the predicted label is displayed through a real-time web interface built using Streamlit.

EXPERIMENTAL SETUP

This section describes the training configuration, evaluation metrics, and experimental protocol used to assess the proposed framework.

A. Data Splitting

After preprocessing and balancing, the dataset is divided into training and test sets using a stratified split. A total of 9,200 messages are used for training, while 800 messages are reserved for evaluation.

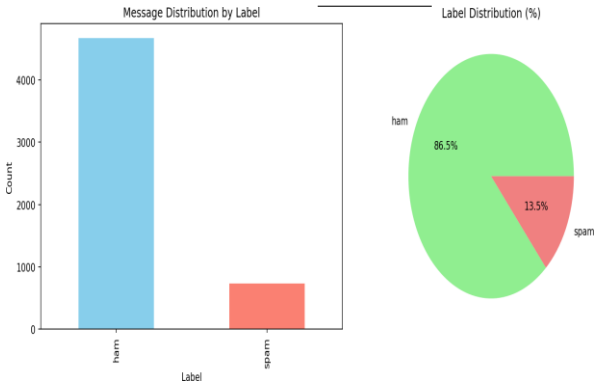


Fig. 2. Original label distribution of the dataset.

B. Training Configuration

The RoBERTa model is fine-tuned using the AdamW optimizer with a learning rate of 2×10^{-5}. Training is conducted for four epochs with a batch size of 32. A linear warm-up schedule is applied to stabilize early training.

Cross-entropy loss is employed as the objective function. All experiments are conducted on a CUDA-enabled GPU.

C. Evaluation Metrics

We evaluate model performance using accuracy, precision, recall, and F1-score. These metrics are defined as follows:

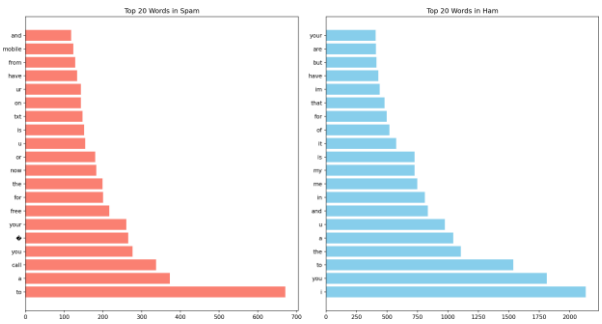


Fig. 3. Top frequent words in spam and ham messages.

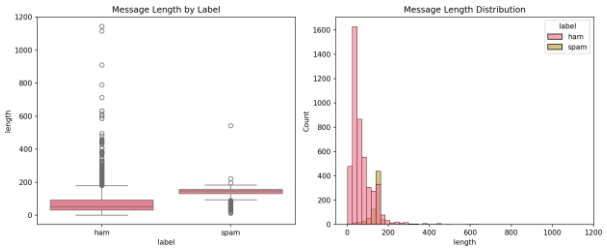


Fig. 4. Message length distribution by class.

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} \times \text{Recall} \quad (2)$$

(3)

E. Proposed Model Results

The RoBERTa-based classifier demonstrates strong generalization capability on the balanced evaluation subset. The model effectively captures semantic and contextual features, reducing both false positives and false negatives compared to

$$\text{F1-score} = 2 \times \frac{\text{Precision} + \text{Recall}}{} \quad (4)$$

RESULTS AND DISCUSSION

This section presents the quantitative evaluation of the proposed framework and compares it with classical baseline models.

Dataset Distribution

Fig. 2 shows the original class distribution of the dataset, highlighting a strong imbalance between legitimate and spam messages.

Word Distribution Analysis

Fig. 3 illustrates the most frequent words in spam and ham messages. Spam messages are dominated by promotional terms, while ham messages contain more conversational words.

Message Length Analysis

Fig. 4 shows the distribution of message lengths. Spam messages tend to be longer on average compared to legitimate messages.

Baseline Model Performance

We evaluate Naive Bayes, Support Vector Machine, and Logistic Regression models using TF-IDF features. These models serve as baseline references.

classical baselines.

DEPLOYMENT

To demonstrate practical usability, the trained model is deployed through an interactive web application using Streamlit. The interface allows users to input SMS messages and receive real-time predictions.

The application performs preprocessing, tokenization, and inference in real time, making the system accessible to non- technical users.

LIMITATIONS

The proposed system is evaluated on English-language SMS messages, which limits its multilingual applicability. Addition- ally, the balanced evaluation setup may not fully reflect real- world class distributions.

Transformer-based models also require higher computational resources than classical approaches.

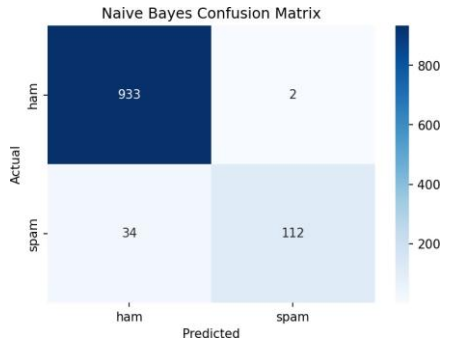


Fig. 5. Confusion matrix for Naive Bayes classifier

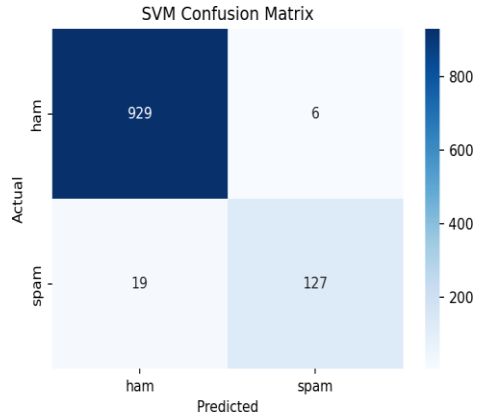


Fig. 6. Confusion matrix for Support Vector Machine classifier

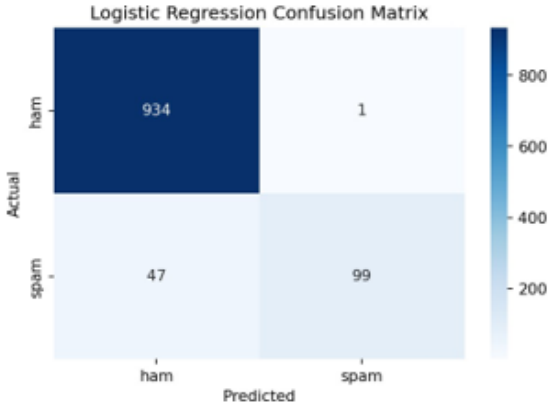


Fig. 7. Confusion matrix for Logistic Regression classifier.

3. CONCLUSION

This paper presented a robust SMS spam detection frame- work based on a fine-tuned RoBERTa transformer model and a real-time web deployment interface. Experimental results demonstrate that transformer-based contextual learning significantly outperforms classical machine learning approaches.

The proposed system bridges the gap between research and real-world deployment, offering a scalable and user-friendly solution for spam detection.

4. FUTURE WORK

Future research directions include extending the system to multilingual datasets, exploring lightweight transformer variants for edge deployment, and incorporating continuous learning mechanisms..

REFERENCES

[1] J. M. G. Hidalgo, "SMS Spam Collection Dataset," UCI Machine Learning Repository, 2012.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre- training of Deep Bidirectional Transformers for Language Understand- ing," NAACL, 2019.

[3] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv:1907.11692, 2019.

[4] Streamlit Inc., "Streamlit: The fastest way to build and share data apps," 2024

.