World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3116

A RESPONSE SURFACE MODELING METHOD FOR SPAM CLASSIFICATION

ShiChen Chen, YouPeng Fan*, Yue lv
Dalian Polytechnic University, Dalian 116034, Liaoning, China.
**Corresponding Author: YouPeng Fan*

Abstract: Spam identification is an important task in Natural Language Processing. To improve the limited nonlinear representation of traditional linear classifiers, this paper proposes a spam classification model based on Response Surface Methodology. The model constructs a dual response surface on Bag-of-Words features: feature-level quadratic transformations learn nonlinear keyword effects, while a classifier-level quadratic correction adjusts the linear decision output. The proposed structure preserves interpretability through explicit polynomial parameters and improves boundary fitting without relying on large-scale deep models. Experiments on public spam datasets show that the model achieves higher accuracy than Logistic Regression and LSTM, while maintaining much lower computational cost than BERT. The results indicate that response surface modeling provides a lightweight, interpretable, and efficient solution for spam identification. This method is especially suitable for resource-constrained scenarios requiring fast training, low inference cost, and transparent feature-level explanation, and it offers a practical balance between model complexity, classification accuracy, and deployment efficiency.
Keywords: Response surface methodology; Spam classification; Non-linear modeling, Feature transformation; Deep neural networks

1 INTRODUCTION

With the widespread adoption of internet communication, e-mail has become an essential tool for daily interaction and business transactions. However, the proliferation of spam has severely disrupted normal communication order, leading to security risks and resource wastage [1-3]. Spam classification is essentially a binary classification problem. Its core challenge lies in the fact that e-mail text typically possesses characteristics such as high dimensionality, sparsity, and non-linear feature interactions [4-5], which makes it difficult for linear models to fully characterize the complex decision boundaries between spam and legitimate e-mails [6]. Traditional machine learning methods, such as Naive Bayes, Support Vector Machines (SVM), and Logistic Regression, can achieve satisfactory results when feature engineering is sufficient. However, constrained by their inherent linear expressive power, they often rely on complex, manually designed features to improve performance [7-9].

Early spam classification primarily relied on rule-based filtering methods, such as keyword matching and blacklisting mechanisms. For instance, Wang et al. proposed a suite of new intelligent text search technologies encompassing information retrieval, information extraction, and information filtering [10]. Shang et al. developed a method that achieved a classification accuracy of over 99% for text emails [11]. However, these methods are susceptible to adversarial attacks and possess

limited generalization capabilities. Subsequently, statistical machine learning methods became mainstream, among which Naive Bayes was widely adopted due to its simplicity and efficiency. For example, Xu et al. proposed a combined spam filtering method to reduce the misjudgment rate of spam [12], and Chen et al. improved the Naive Bayes classification model to enhance model performance [13]. Nevertheless, the feature independence assumption of Naive Bayes often deviates from the actual text distribution. Support Vector Machines (SVM) can handle non-linear problems through the kernel trick; for instance, G proposed a fusion algorithm of KNN and SVM to improve model performance [14]. Zhang et al. suggested converting email text into vector features and utilized Convolutional Neural Networks (CNN) to identify spam on the internet, verifying the model's effectiveness [15]. However, kernel function selection and parameter tuning for SVM are complex, and its training efficiency on large-scale data is relatively low.

Based on the aforementioned analysis, this paper proposes a dual response surface model for spam classification. By constructing independent quadratic transformation functions ($\varphi_i(x_i) = a_i x_i^2 + b_i x_i + c_i$) for each word frequency feature, the model is able to learn non-linear patterns of features and automatically identify their effective action intervals. A quadratic correction term ($f(s) = s + \gamma s^2$) is superimposed on the output of a traditional linear classifier. Through learnable quadratic coefficients, the decision surface can be "inverted" or "enhanced," thereby improving the model's ability to fit non-linear boundaries.

Experimental results indicate that the model achieves an accuracy of 0.9520 on the test set, significantly outperforming traditional Logistic Regression (0.8960) and LSTM (0.8880), while keeping the performance gap with BERT (0.9680) within 1.6%. More importantly, the training and inference time of the proposed model is reduced by 100 times compared to BERT and 20 times compared to LSTM, achieving an optimal balance between accuracy and efficiency. This work provides an effective and interpretable lightweight solution for high-performance spam identification in resource-constrained scenarios.

2 MODEL DESCRIPTION AND MATHEMATICAL FORMULATION

2.1 Rigorous Problem Definition

To accurately describe the proposed model, it is first necessary to provide a
formal definition of the spam classification problem and standardize the relevant notations. A precise definition of the problem facilitates the elucidation of the mathematical structure of inputs and outputs, thereby precluding ambiguity in subsequent derivations. This section begins by introducing the basic settings of the classification task, followed by the notational framework used for input features, output labels, and model parameters, establishing a formal foundation for the model construction and optimization discussed in subsequent chapters.
Assume that the training set consists of n samples {(xi, yi)}i=1^{n}, where xi in Rd denotes the feature vector obtained through the Bag-of-Words (BoW) transformation, and yi in {0,1} represents the class label (0 for non-spam/ham and 1 for spam). Here, d denotes the feature dimensionality, which corresponds to the size of the vocabulary.
Objective: To learn a function f: X -> R such that the predicted value yhati = sigma(f(xi)) is as close as possible to the true label yi, where sigma(z) = 1/(1+exp(-z)) is the sigmoid function.

2.2 Feature-Level Response Surface

RSM approximates the true relationship between inputs and responses by constructing low-order polynomials (typically quadratic), and its parameters possess explicit geometric meanings. Based on this idea, this paper proposes a dual response surface classification model: at the feature level, an independent quadratic transformation is applied to each word frequency feature, enabling the model to learn the non-linear action patterns of features; at the decision level, a quadratic correction is superimposed on the linear classifier's output to achieve non-linear modulation of the classification surface.

$$\varphi_i(x_i)=a_i x_i^2+b_i x_i+c_i, i=1,...,d \tag{1}$$

as a non-linear mapping of the original features, allowing the model to learn feature importance (linear term) as well as saturation/enhancement effects (quadratic term). The entire feature transformation can be written in vector form as:

$$\Phi(x)=a\odot x^2+b\odot x+c \tag{2}$$

2.3 Classifier Response Surface

A quadratic correction is applied to the result of the linear combination. Let the output of the linear layer be defined as:

$$s=w^T\Phi(x)+b_0 \tag{3}$$

$$f(s)=s+\gamma s^2 \tag{4}$$

2.4 Integrated Model

To fully exploit the joint expressive capability of these non-linear transformations, this section integrates both layers into a unified end-to-end model. All parameters are optimized jointly, allowing the model to learn non-linear interactions between features while enabling fine-grained adjustment of the decision boundary. Mathematically, the integrated model constitutes a differentiable composite function. When all non-linear parameters are set to zero or identity, the model strictly degenerates into classical Logistic Regression, representing a natural extension of linear models that balances expressive power with interpretability.
Combining the two aforementioned components yields the complete response surface classifier:

$$\begin{aligned} f_\theta(x)&=w^T(a\odot x^2+b\odot x+c)+b_0 \\ &+\gamma[w^T(a\odot x^2+b\odot x+c)+b_0]^2 \end{aligned} \tag{5}$$

3 PARAMETER LEARNING, OPTIMIZATION, AND LOSS TOPOGRAPHY

3.1 Weighted Binary Cross-Entropy Loss Function

In binary classification problems, cross-entropy loss is the most commonly utilized objective function. Considering that spam datasets typically exhibit class imbalance—where the number of non-spam emails significantly exceeds that of spam—the direct application of standard cross-entropy may cause the model to favor the majority class. To address this, class weights w0 and w1 are introduced to balance the loss contributions from both categories of samples. Specifically, a weighted binary cross-entropy loss is adopted:

$$L(\theta)=-1/N\sum_{i=1}^{N}[\omega_1 y_i\log(yhat_i)+\omega_0(1-y_i) \tag{6}$$

3.2 Adaptive Optimization Algorithm

To efficiently solve for the model parameters θ, this paper adopts the Adam optimizer. Adam combines the advantages of momentum methods and adaptive learning rates, enabling it to effectively handle high-dimensional sparse gradients (such as bag-of-words features) and non-stationary objective functions. The parameter update rules are as follows:

$$g_t=\nabla_\theta L(\theta_t) \tag{7}$$

$$m_t=\beta_1 m_{t-1}+(1-\beta_1)g_t \tag{8}$$
$$v_t=\beta_2 v_{t-1}+(1-\beta_2)g_t^2$$

where $\beta_1=0.9$ and $\beta_2=0.999$ are the exponential decay rates. To correct the bias at the initial time steps, the following are calculated:

$$mhat_t=m_t/(1-\beta_1^t), vhat_t=v_t/(1-\beta_2^t) \tag{9}$$

The final parameter update is defined as:

$$\theta_{t+1}=\theta_t-\eta mhat_t/(sqrt(vhat_t)+\varepsilon) \tag{10}$$

AdamW regularization (weight decay = 1×10^{-4}) is incorporated during the training process, which is equivalent to adding a $\lambda/2||\theta||^2$ term to the loss function to suppress overfitting.

4 MODEL INTERPRETATION AND GEOMETRIC SIGNIFICANCE

4.1 Geometric Interpretation of Feature Response Surfaces

$$x_i^* =-b_i/(2a_i) \tag{11}$$

The coefficient a_i determines the curvature of the feature response surface. If ai > 0, the surface opens upward; if ai < 0, the surface opens downward. By learning these parameters, the model can automatically discover the effective action interval for each feature.

4.2 The Topological Inversion Effect of the Classifier Response Surface

$$f'(s)=1+2\gamma s \tag{12}$$

$$s^* =-1/(2\gamma) \tag{13}$$

4.3 Continuous Relationship with Logistic Regression

$$z=w^T x+b_0 \tag{14}$$

One of the core components of the response surface model is to apply a quadratic transformation to each term frequency feature:

$$xhat_i=a_i x_i^2+b_i x_i+c_i \tag{15}$$

The parameter a_i governs the strength of nonlinearity, and the magnitude of its absolute value reflects the importance of the nonlinear effect of the corresponding feature in the classification task. By analyzing the distribution of a_i and the top 20 features, the feature importance and nonlinear behavior learned by the model can be revealed.

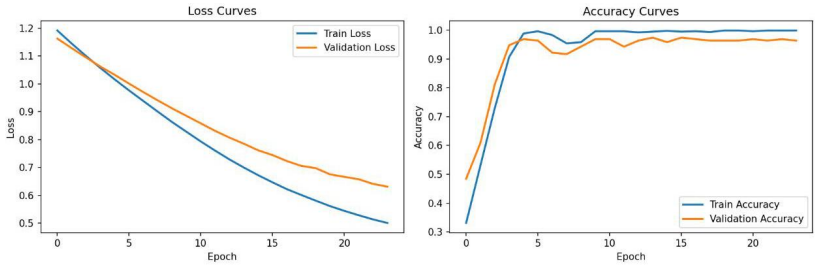

Figure 1 Training Curves

As shown in Figure 1, the training and validation curves remain stable during iteration, indicating that the response surface model converges smoothly and does not show obvious overfitting. This trend supports the effectiveness of the optimization strategy and provides visual evidence for the subsequent performance comparison.

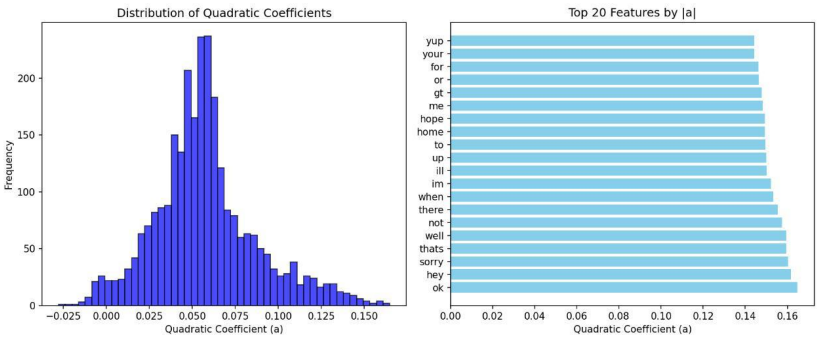

Figure 2 Feature Response Parameters

The left subplot of Figure 2 presents a histogram of all feature quadratic coefficients a_i. The majority of the values are concentrated around zero, with only a few features exhibiting significant positive or negative values. This indicates that the model applies nonlinear transformations only to key features, thereby effectively avoiding overfitting. The right subplot lists the 20 features with the largest absolute quadratic coefficients $|a_i|$, such as "yup," "your," "for," "gt," and "me." These are predominantly high-frequency abbreviations or common words found in spam emails; the model enhances their discriminative power through these quadratic terms. This result validates that the feature response surface can adaptively learn the nonlinear effects of features, further enhancing model interpretability.

The classifier response surface maps the linear score s to the final logits: $z = f(s)$. The parameter γ determines the degree of nonlinearity. If $\gamma > 1$ (or $\gamma > 0$ depending on the specific convexity), the function is convex, providing a greater response to high-confidence samples, which helps improve the clarity of the classification boundaries.

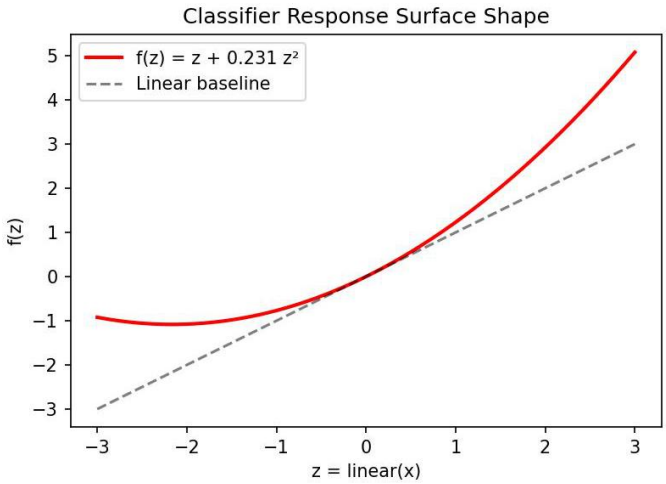

Figure 3 Classifier Response Surface Shape

Figure 3 illustrates the classifier response surface curve, where $\gamma > 1$ (positive). Compared to the linear baseline (dashed line), the actual curve exhibits an accelerated upward trend in the region where $s > 0$, implying that the model assigns stronger positive outputs to samples with high predictive confidence (e.g., $s >> 0$) and provides more pronounced suppression for samples with negative confidence ($s < 0$). Consequently, this widens the decision margin between the positive and negative classes. This nonlinear transformation enhances the discrimination capability of the classifier, validating the effectiveness of the classifier response surface. Figure 3 further illustrates how the classifier response surface reshapes the linear score. The nonlinear curve enhances high-confidence positive outputs and suppresses negative-confidence samples, so the decision margin becomes clearer than that of the linear baseline.

Table 1 Result Comparison

Model Architecture Strategy	Training Set Accuracy	Testing Set Accuracy
Response Surface Modeling	0.9987	0.9520
BERT	0.9895	0.9680
LSTM	0.9583	0.8880
TF-IDF + Standard Logistic Regression	1.0000	0.8960
Ablation Variant: Only Classifier Response Surface	0.9974	0.9360
Ablation Variant: Only Feature Response Surface	0.9974	0.8880

The Response Surface Model (RSM) achieves an optimal balance between accuracy and computational efficiency, see Table 1. Its test accuracy reaches 0.9520, which is second only to BERT (0.9680), yet significantly higher than LSTM (0.8880), TF-IDF + Logistic Regression (0.8960), and variants utilizing only a single component of the response surface. Most importantly, the training and inference times of this model are reduced by more than 100 x compared to BERT and by 20 x compared to LSTM.

Through the joint optimization of the feature response surface and the classifier response surface, the model successfully captures the nonlinear characteristics inherent in spam classification. The feature response surface applies adaptive quadratic transformations to keyword frequencies, while the classifier response surface widens the decision margin for high-confidence samples via γ , effectively enhancing discriminative power. Although the training accuracy approaches 1.0, the test accuracy remains high, indicating that the model possesses strong generalization alongside its robust fitting capabilities. In terms of both precision and efficiency, the Response Surface Model outperforms traditional linear models and common deep learning architectures like LSTM. With computational costs far lower than BERT, it represents a lightweight, high-performance spam classification method suitable for practical deployment.

5 CONCLUSIONS

This paper proposes a spam classification model based on a Response Surface Mechanism. The model combines feature-level quadratic transformation with classifier-level quadratic correction, enabling it to capture nonlinear relationships in Bag-of-Words features while retaining clear parameter interpretability. Experimental results show that the proposed model achieves better accuracy than Logistic Regression and LSTM, and obtains a favorable balance between classification performance and computational efficiency compared with BERT. The main limitation is that the current model still relies on BoW features and therefore cannot fully represent word order, contextual semantics, or subtle adversarial expressions. Future research may integrate lightweight contextual embeddings, ensemble response-surface structures, and online parameter updating to improve semantic recognition ability and robustness under changing spam patterns.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Asim S, Mohd N N, Zubair M R, et al. An Improved Framework for Content-and-Link-Based Web-Spam Detection: A Combined Approach. Complexity, 2021. DOI: 10.1155/2021/6625739.
[2] Shan C L, Zhang X Y, Xing H L, et al. A Chinese Spam Recognition Method Based on Content and ERNIE3.0-CapsNet. Journal of Information Security Research, 2024, 10(03): 233-240.
[3] Gui J, Zhou Y, Yu K, et al. PSC-BERT: A spam identification and classification algorithm via prompt learning and spell check. Knowledge-Based Systems, 2024, 301: 112266.
[4] Dawei Z, Yanyu L. Identification and Filtering of Web Spams Using a Machine Learning Method. International Journal of Computational Intelligence and Applications, 2022, 21(04).
[5] Zengle G. A fusion algorithm model based on KNN-SVM to classify and recognize spam. Journal of Physics: Conference Series, 2021, 1982(1).
[6] Zhang Y Q, Wang W. Spam Classification Using Support Vector Machine Optimized by Genetic Algorithm. Journal of Computer Applications, 2009, 29(10): 2755-2757.
[7] Sharma P K, Lal G, Shukla M, et al. Quantum behaved binary gravitational search algorithm with random forest for twitter spammer detection. Results in Engineering, 2025, 25: 103993.
[8] Ömer A, Serkant S A, Merve O, et al. A Comprehensive Review of Cyber Security Vulnerabilities, Threats, Attacks, and Solutions. Electronics, 2023, 12(6): 1333.
[9] Faris H, Al-Zoubi M A, Heidari A A, et al. An Intelligent System for Spam Detection and Identification of the Most Relevant Features Based on Evolutionary Random Weight Networks. Information Fusion, 2018, 48: 67-83.
[10] Wang Z Y, Xu W R, Guo J. New Technology for Intelligent Text Search. CAAI Transactions on Intelligent Systems, 2012, 7(01): 40-49.
[11] Shang T, Guo Z Y, Wang Y S. A High-Accuracy Spam Recognition Method. Journal of Huazhong University of Science and Technology (Natural Science Edition), 2011, 39(S2): 287-290.
[12] Xu Y B, Li Z, Dong Y. Fast Spam Filtering Based on Social Computing and Machine Learning. Systems Engineering — Theory & Practice, 2014, 34(S1): 179-186.
[13] Chen J Y, Zhou S F, Min H Q. A Hybrid "Word Frequency-Filter" Feature Selection Method for Spam Recognition. Journal of South China University of Technology (Natural Science Edition), 2017, 45(03): 82-88.
[14] Zengle G. A fusion algorithm model based on KNN-SVM to classify and recognize spam. Journal of Physics: Conference Series, 2021, 1982(1).
[15] Dawei Z, Yanyu L. Identification and Filtering of Web Spams Using a Machine Learning Method. International Journal of Computational Intelligence and Applications, 2022, 21(04).