Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 5840759, 10 pages
https://doi.org/10.1155/2021/5840759

Research Article

Effective Emotion Recognition Technique in NLP Task over Nonlinear Big Data Cluster

Woo Hyun Park,[1] Dong Ryeol Shin,[1] and Nawab Muhammad Faseeh Qureshi[2]

[1]*Department of Electrical and Computer Engineering, Sungkyunkwan University, 16419, Republic of Korea*
[2]*Department of Computer Education, Sungkyunkwan University, 03063, Republic of Korea*

Correspondence should be addressed to Nawab Muhammad Faseeh Qureshi; faseeh@skku.edu

Received 4 August 2021; Revised 9 November 2021; Accepted 16 November 2021; Published 14 December 2021

Academic Editor: Muhammad Shafiq

Human-to-human communication can be achieved not only by body language but also by high-level language. Moreover, information can be conveyed in writing. In particular, the high-level and specific process of logical thinking can be expressed in writing. Text is data that we encounter daily, and there are hidden patterns in it. A person's cognitive activity, that is, text data, contains the author's emotions. In the existing text analysis method, simply using the frequency of words has limited interpretability. The model proposed in this paper is a nonlinear emotion system based on emotion to increase document diversity. The purpose is to effectively converge features by assigning weights to a nonlinear function with existing training and learning methods. Our study used the confusion matrix, an area under the receiver operating characteristic curve, and F1-score as evaluation methods. This research created a new error function and measured emotions. The accuracy was 0.9447, and the model's receiver operating curve peak was 0.9845, which is somewhat similar to that of TF-IDF in the evaluation.

1. Introduction

What makes humans different from animals is that they have a high level of wisdom and cognitive activity. A typical example of this is linguistic elements. They can communicate with each other and also convey information through writing. We encounter textual data daily, and patterns are hidden in it. As a representative example, the writing tasks commonly used can extract SMS and spam mail. Text data contain the emotions of the author. It is unlikely for robots to have better cognitive abilities than humans in the future. However, with the advent of AI and the digital age, robots will analyze human texts accurately. Various natural language processing (NLP) techniques have been studied to analyze human-written documents. A typical example is the term frequency-inverse document frequency (TF-IDF) model. However, in this existing text analysis method, simply using the frequency of words has a limited meaning. Several NLP models have been studied to solve this problem. Reference [1] proposed using information from latent Dirichlet allocation (LDA) topic distribution and word frequency. In this paper, we propose the emotion nonlinear system model to solve the problem of increased performance. The proposed ENLS model contributes to the error by analyzing and transforming the sentiment of spam mail.

The contributions of the models proposed in this paper are as follows.

(i) Create a new error function

(ii) Measure the emotion included in the text

(iii) Narrow the gap performance between testing and training data with computational efficiency

2. Theory and Related Research

Various NLP techniques have been studied.

Reference [2] conducted a study on text similarity using Arabic. Correlations were calculated after collation with the original using support vector machine (SVM) and stochastic gradient descent (SGD), naïve Bayes, and 10 algorithms. The experimental results were good in Farasa and Qalsadi Lemmatizer.

Reference [3] used the TF-IDF model to analyze the Chinese Imperial Encyclopedia. In addition, detection algorithms were used to classify subjects, and Ochiai Coefficient and topological networks were also built. The results of the experiment confirmed that the distribution according to the subject was cyclical.

Reference [4] used machine learning and HHO to classify people who spread false information. Specifically, algorithms were compared and analyzed using user profiles, content, and word frequency. The experimental results showed that the logistic regression model had the best performance in classifying false information.

Reference [5] attempted to understand human emotions using emotional techniques. Features were extracted using TF-IDF, SIMON, affective space, and sentient information. Next, ranking statistics tests were conducted on five different datasets, and the proposed method effectively improved the performance.

Reference [6] attempted to analyze detailed emotional states using machine learning in Arabic. The techniques used in the study were TF-IDF and BERT, and the result of measuring the F1-score obtained the best performance.

Reference [7] classified texts on Urdu-speaking social media using convolutional, recurrent, and deep neural networks (CNN, RNN, and DNN, respectively). Experimental evaluation of TF-IDF and Wod2Vec resulted in over 80% accuracy using DNN.

Reference [8] proposed a system for automatically classifying Quranic passages. For classification, the Quran was preprocessed to train a corpus with ensemble multiple labels, followed by TF-IDF and Word2Vec and SVM, logistic regression, and random forest classifiers. The results of the experiment showed that the ensemble model performed best.

Reference [9] presented a method of combining a deep learning model with statistical methods to grasp the meaning of all languages in the world. Specifically, Jaccard coefficients were calculated using the TF-IDF and CNN algorithms to reduce the granularity of feature extraction.

Reference [10] proposed a model using GloVe and recurrent CNN to supplement the problems of the existing TF-IDF method and LDA used in plagiarism detection. This was effective in analyzing the sentence structure of Arabic.

Reference [11] used deep neural networks and trigrams to identify derogatory texts. Three language models and ten supervised and unsupervised learning and training functions were evaluated. The derogatory text was calculated with an accuracy of approximately 95% or more.

Reference [12] tracked attitudes and emotions by analyzing articles uploaded online to help the medical industry. The TFIDF was improved, and weights were given in the text filtering and emotion dictionaries. It showed better accuracy than the TF-IDF.

3. ENLS Model

3.1. Motivation. Documents contain people's emotions. It was intended to analyze the terms in documents to create a unique vector and calculate the emotion to create and analyze the emotion vector. By utilizing the identified features, the features corresponding to positive and negative emotions used in the document were quantified and weighted. Previous related studies are described in the following. Reference [13] investigated the effect of weights on the learning rate. As a result, it was confirmed that the parameter does not affect the normalization and is dependent on the y of the batch normalization. The regularization equation used in this paper was as follows [13]:

$$L_{\lambda}(\mathbf{w}) = L(\mathbf{w}) + \lambda \|\mathbf{w}\|_2^2. \tag{1}$$

In our study, we also conducted training using regularization.

Reference [14] used k-nearest neighbors and TF-IDF for text classification to determine k influencing classification performance and show the advantages and disadvantages and reliably obtained good results, emphasizing the importance of preprocessing the data.

The expression for TF-IDF is given by [14]

$$a_{ij} = tf_{ij} idf_i = tf_{ij} \times \log_2\left(\frac{N}{df_i}\right). \tag{2}$$

Reference [15] compared the linear discriminant analysis with logistic regression and found that logistic regression was superior, given as [15]

$$\frac{p(\mathbf{x}) = \Pr\left\{E\mathbf{x}\right\} = 1}{\left[1 + \exp\left\{-\alpha - \boldsymbol{\beta}'\mathbf{x}\right\}\right]}. \tag{3}$$

Therefore, in our study, logistic regression was used.

Reference [16] emphasized the importance of statistical error functions when evaluating models. The results of its experiments with the Black–Scholes model were as follows. It argued that the evaluation function used to evaluate the model should be the same. It was found that PBS had a greater effect on the performance of the model than root mean squared error (RMSE) and was more generalized.

Reference [17] was aimed at understanding the housing market situation by collecting real estate news articles corresponding to housing market prices. To analyze the articles, the TF-IDF, a weight model, and n-gram were applied.

Reference [18] proposed various methods for finding weights using terms, documents, and class frequencies to automatically classify posts on blog sites without the hassle of manual work.

Reference [19] conducted a study to measure the degree of similarity with summary texts for biomedical papers by effectively assigning document weights. A higher level of performance was derived than when using the TF-IDF model.

Reference [20] proposed a weighting model based on whether keywords reappeared and changed to improve search performance for patent papers. Better performance than simple frequency and IDF was derived.

Reference [21] applied negative word processing using *n* -gram and corpus-specific term weighting to analyze the sentiment of blog documents. As a result of experiments on both EMFA and SVMMC, *n*-gram and CSTW showed superior classification performance compared to the TF-IDF model.

Reference [22] analyzed comments and judged comments that slander others. In other words, a system for determining whether a comment is malicious or not was implemented. Weights were calculated using TF.

Reference [23] divided the population into subgroups using classification and regression tree analysis. Specifically, it investigated the method of calculating the criteria using the Gini impurity function for the prevalence data of influenza vaccination and the error techniques accordingly. The classification results were validated using various versions of the logistic regression model.

Reference [24] used data from 72 medical papers to examine logistic regression models and artificial neural networks originating from different fields and compared and analyzed similarities and differences through KNN, decision trees, and machine learning algorithms of SVM.

Reference [25] compared the performances of a decision tree, SVM, and ANN using naive Bayes to correctly classify documents. As a result, it was found that naive Bayes performed the best.

However, it was confirmed that the performance of logistic regression was much better in the spam document classification performed in this paper.

Reference [26] measured the classification accuracy and training time while growing an ensemble of a random forest and compared the performance with that of the SVM algorithm using ETM+data. As a result of the comparison, it was confirmed that the same performance was obtained and it was easier to define than SVM.

Reference [27] considered the difference in emotional levels in Korean and quantified emotional intensity using the chi-square statistic. Weight was calculated by dividing the emotional intensity of each sentence by the highest emotion. Learning was performed using SVM, and as a result of the evaluation, the performance was superior to that of the conventional method.

Reference [28] inferred the relationship between diseases and symptoms using medical information data. To solve the problem of finding meaningless association rules in medical data, a meaningful association rule was constructed by assigning weights based on TF-IDF. By comparing the rules of medical knowledge and data, the relationship between diseases and symptoms was more effectively identified with the frequent pattern growth algorithm, helping medical staff in decision-making.

Reference [29] constructed a recommendation system to help customers choose their book products. Rather than using the existing bibliographic information and user information, meaningful information was extracted from the text content using the TF-IDF model and subject words of the text. The sentence structure was classified into four levels, and weights were calculated. The subject words were extracted more effectively than when simply using the content.

Reference [30] classified and ranked socially occurring problems using a machine learning algorithm. It concluded that machine learning algorithms and functions could not be interpreted naturally by humans and could not solve all social problems.

Reference [31] predicted consumer opinion for positive feedback on consumer reviews. It used TF-IDF, Word2Vec, BOW, GloVe, etc. to conduct logistic regression, gradient descent, and neural network research. The use of Word2Vec and CNNs produced good classification results for consumer reviews.

Reference [32] detected erotic/sexual content by analyzing comments using NLP technology. The technologies used in this research were BOW, TF-IDF, Word2Vec, SVM, and logistic regression. The classification accuracy was 97% using TF-IDF and SVM.

Reference [33] improved the DID classifier, a method that uses Gaussian mixture modeling. To classify spoken dialects, it used an *n*-gram model and LSA and TF-IDF and logistic regression classification to classify dialect-like noise and achieve similar performance to acoustic systems.

Reference [34] achieved 93% for KNN, 89% for naive Bayes, 94% for the decision tree, 95% for AdaBoost, 94% for the random forest, and 94% for SVM in the performance evaluation of random spam email classification. Among them, AdaBoost had the highest performance. It has a recall, precision, and F1-measure of 95%, 96%, 96%, and 96%, respectively. The accuracy of the ENLS proposed for classifying spam emails in this paper is 94%.

Reference [35] uses Bayesian logistic regression to achieve 93%, hidden naive Bayes to achieve 90%, RBFNetwork to achieve 89%, VotedPerceptron to achieve 92%, LazyBayesianRules to achieve 92%, LogitBoost to achieve 89%, and Nnge to achieve 91% accuracy. The model tree achieved 93% accuracy, and in J48, 92% accuracy was achieved and studied. Among them, the highest accuracy of 94% was achieved using the rotation forest algorithm. Its recall, precision, and F1-measure are 94%, 94%, and 94%, respectively.

3.2. Overall Structure. In our study, UCI spam data were collected and used as input.

The system for classifying texts by learning the weight error function along with emotions is as follows (see Figure 1). The system consists of five modules. The first is a bag-of-words (BOW) module. Spam data were retrieved and preprocessed to classify normal and abnormal mail. It preprocesses unnecessary parts of the research data and makes it a form of learning. Subsequently, the one-hot encoding processes and frequency counts were calculated. After that, the feature was created using equation (2). The second is an emotion-based analysis module. This module analyzes the sentiment of spam data. Third, the results of these preprocessing and emotion analyses were characterized without using them as they were. The fourth was an

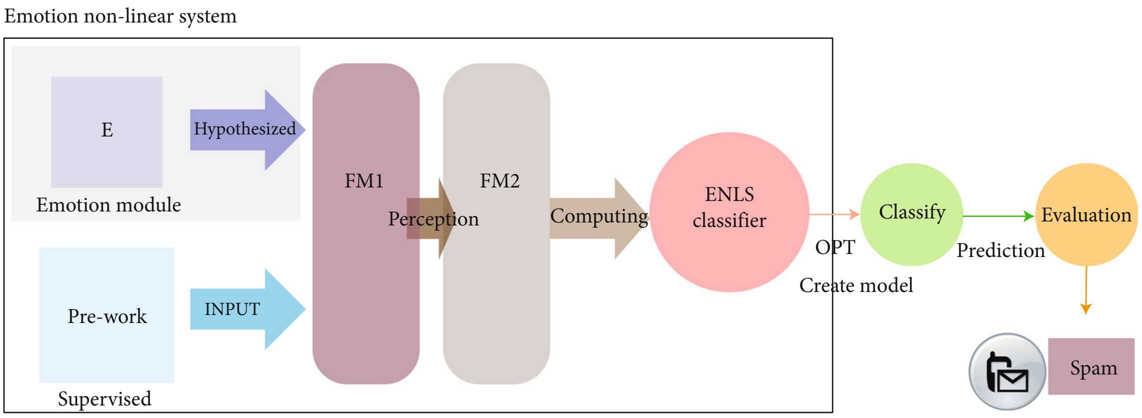


Figure 1: Illustration of the overall system.

emotion-based error function that uses ENLS. Finally, classification was performed. Here, E_n stands for the emotion module. It extracts sentiment from the mail. Prework preprocesses the data to create word frequency features. Alpha represents the weighted value of the emotion. The F1 module vectorizes the emotional words extracted as positive and negative and converts the word frequency feature into a vector. k is defined as a real number. w stands for the weight of the emotion word. The F2 module constrains a limit between the word frequency feature and the sentiment vector, passing a valid value through it. i represents the transformed value into a vector. The ENLS (E, R) module serves to optimize the transformed matrix.

3.3. ENLS (Emotional Nonlinear System). We hypothesized that positive words have a positive effect on spam and negative words have a negative effect.

The first processing of ENLS was performed through emotion analysis as input. By analyzing the emotion of the spam text data received as input, the emotion value was calculated using the positive and negative words of good and bad, respectively. The emotional expression for this is as follows. The formula for this is given as

$$E_a = \begin{cases} \alpha, & \text{if Pos}, \\ \log(\alpha), & \text{if Neg}. \end{cases} \tag{4}$$

First, if it is positive, it is initialized to the value of alpha. Alpha takes a real value. If it is negative, the value of alpha is logged.

3.4. Mixture Emotional Spam Identification. The ENLS process proceeds in the F1 module. The frequency count is preprocessed, and the emotions are weighted in space. This function is given as

$$F_1(x) = \begin{cases} \text{TF}, \\ E_0(\omega_1, \omega_2, \cdots, \omega_v), \end{cases} \tag{5}$$

When the result value of F1 enters the input of module F2, module F2 passes it through using a nonlinear function. The nonlinear function is given by

$$F_2(i) = \begin{cases} k * \mathbf{i}, & \text{if } 0\rangle \mathbf{k} * i\rangle \mathbf{i}, \\ i, & \text{if } i > 0, \end{cases} \tag{6}$$

$$F_2'(i) = \begin{cases} \mathbf{k}, & \text{if } 0\rangle \mathbf{k} * \mathbf{i}\rangle \mathbf{i}, \\ 0, & \text{if } i > 0. \end{cases} \tag{7}$$

An influence of this characteristic function is that if i is positive, it has a value of 0 concerning emotion vectors, while if it is negative, it has value k. k is a constant and acts as a parameter. Before entering the input of the ENLS ends, matrices E and R of size $m \times n$ are created, and then, the parameters k and i are solved.

$$R = AB. \tag{8}$$

A vector space is defined to make vectors independent of each other as the dimension of R vector space R. This is a matrix randomly generated based on the number of ranks of the matrix. The formula for calculating the rank of R is as follows:

$$\rho\kappa(R). \tag{9}$$

The expression ENLS (E, R) is computed by

$$\sum_{\upsilon,\iota\in\Sigma} E_{\upsilon\iota}\left(R - R_{\upsilon\iota}'\right)^2. \tag{10}$$

u and i represent the dimensions of the real space and are calculated as the sum of the squares of the error between the predicted value and this matrix by assigning weights based on the emotion vector.

To evaluate this, MSE was used alongside the L2 regularization of equation (1).

Let us call these p and q, as per equation (11).

$$p = BE_uB^T + \lambda I, \tag{11}$$

$$q = BE_uR_u^T. \tag{12}$$

```
    Input: Data Set
Initialization:
        1.$\{x_1, y_1, \cdots, x_n, y_n\}$
        2.$a \in R$
Procedure:
for $u, i \longleftarrow 0$ to $k - 1$ do
            1. $\mathrm{E}_{n(a)} \longleftarrow Pos$ in $Eq.$ (4)
            2. $\mathrm{E}_{n(a)} \longleftarrow Neg$ in $Eq.$ (4)
            3. Feature $\longleftarrow$ in $Eq.$ (3)
            $\mathrm{F}_1 \longleftarrow BOW, E_n$
end for
Creates $(\mathrm{M} \times \mathrm{N})$ matrix from $\mathrm{F}_1$
According to K, $\alpha$
if $i \in R > 0$ then
            $F_2 \longleftarrow$ using Eq. (6), Eq. (7)
else
            $F_2 \longleftarrow$ using Eq. (6), Eq. (7)
end if
1. Calculate $rk(R_{mn})$
2. $R = A_m B_n$
Repeat
            1. $EALS_{ui} \longleftarrow$ in $Eq.$(10)
            2. Find $p, q \longleftarrow$ in $Eq.$ (11) and $Eq.$ (12)
            3. Solve $\longleftarrow$ in $Eq.$ (11) and $Eq.$ (12)
            4. Convergence
            5. $EALS\,(E, R) \longleftarrow$ result
```

Algorithm 1: Algorithm of the ENLS model.

Table 1: Settings.

Dataset	Environment	Models
Data at UCI University	scikit-learn/Python 3.4 in Windows 10-home 64-bit	ENLS TF-IDF

Table 2: TF-IDF and ENLS 1 ($a = 0.5$, $k = 7$), 2 ($a = 0.5$, $k = 0.4$), and 3 ($a = 0.5$, $k = 0.07$).

(a)

LR/model type		TF-IDF
Evaluation	Train ACC	0.9659
	Test ACC	0.9480
	Matrix	[[2404 2] [93 288]]
	Recall	1.00/0.76
	Precision	0.96/0.99
	F1-val	0.98/0.86

(b)

LR/model type		ENLS 1	ENLS 2	ENLS 3
Evaluation	Train ACC	0.9320	0.9361	0.9377
	Test ACC	0.9318	0.9358	0.9372
	Matrix	[[2415 6] [184 182]]	[[2418 3] [176 190]]	[[2418 3] [172 194]]
	Recall	1.00/0.5	1.00/0.52	1.00/0.53
	Precision	0.93/0.97	0.93/0.98	0.93/0.98
	F1-val	0.66/0.98	0.96/0.68	0.97/0.69
	ROC curve	Figure 2	Figure 2	Figure 2

Table 3: ENLS 1 ($a = 0.6$, $k = 3$), 2 ($a = 0.6$, $k = 0.3$), and 3 ($a = 0.6$, $k = 0.04$).

LR/model type		ENLS 1	ENLS 2	ENLS 3
Evaluation	Train ACC	0.9421	0.9536	0.9279
	Test ACC	0.9386	0.9447	0.9214
	Matrix	[[2371 50] [314 52]]	[[2414 7] [147 219]]	[[2420 1] [218 148]]
	Recall	1.00/0.55	1.00/0.6	1.00/0.40
	Precision	0.94/0.98	0.94/0.97	0.92/1.00
	F1-val	0.97/0.70	0.97/0.74	0.96/0.57
	ROC curve	Figure 2	Figure 2	Figure 2

p is the value obtained by multiplying B by the emotion vector, which is an orthogonal matrix, by the product of the transpose matrix of B, and then by lambda multiplied by the identity matrix. The emotion vector was obtained through the F1 module, and the TF vector was obtained. F2 created E and R matrices based on this. Optimization was performed by calculating the number of ranks from the original matrix of the feature, calculating the square error between the calculated and generated matrices from the number of ranks, multiplying the emotion weights, and performing expression learning. Here, lambda is 0.001. q is calculated by multiplying B by the emotion vector, which is an orthogonal matrix, and the transpose matrix of R. While solving p and q, total learning was performed 15 times using the training data and training the feature representation learning in the ENLS system until it converged. The following shows the ENLS algorithm process. To classify spam data, it starts initialization in the form of supervised learning and contains alpha, which is a real value depending on the positive and negative values. It quantifies emotions in the emotional module as a feature of spam. To do this, it vectorizes the bag of words and positive and negative emotions of the preprocessed values. After that, emotions were processed based on thresholds using nonlinear relationships. After that, expression learning was performed by calculating the number of ranks from the original matrix of features created by word frequency, calculating the square error between the calculated and generated matrices from the number of ranks and multiplying by emotion weight. Optimization was performed iteratively until convergence.

4. Results and Discussion

4.1. Baseline. UCI storage was used to use spam data, and Python and scikit-learn were used in Windows environments for implementation and result simulation. In detail, as the input of the ENLS model, spam data [36, 37, 38, 39, 40] provided by UCI University were used. The ratio of training to test data was 1 : 1. Specific setting values are listed in Table 1.

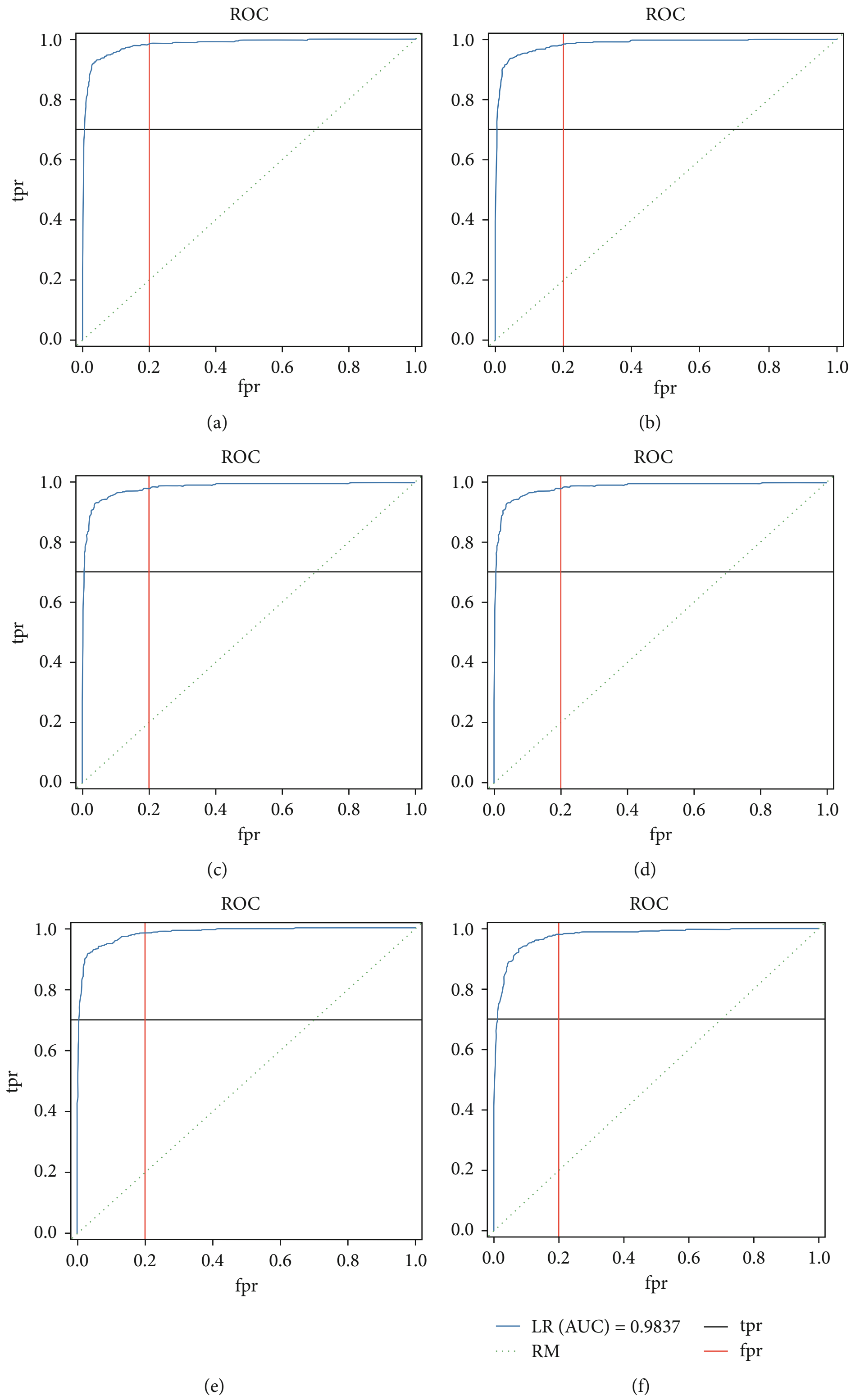
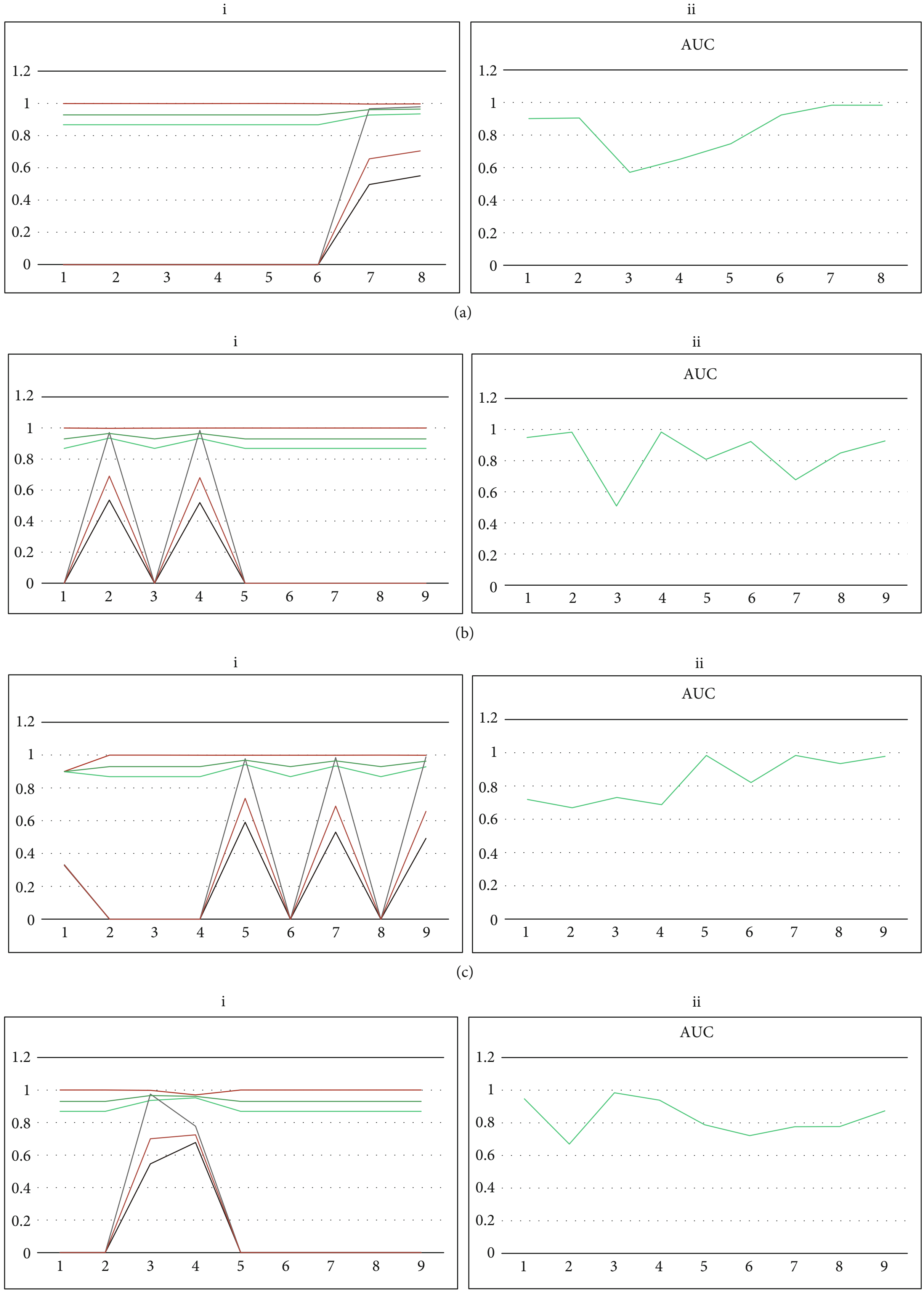


Figure 2: ROC curves: (a) ENLS ($a = 0.5$, $k = 7$), (b) ENLS ($a = 0.5$, $k = 0.4$), (c) ENLS ($a = 0.5$, $k = 0.07$), (d) ENLS ($a = 0.6$, $k = 3$), (e) ENLS ($a = 0.6$, $k = 0.3$), and (f) ENLS ($a = 0.6$, $k = 0.04$).

4.2. ENLS vs. TF-IDF. To evaluate ENLS, we objectively compared its performance with TF-IDF. The specific performance measurement results are presented in Tables 2 and 3. First of all, the training and testing accuracies of TF-IDF are about 0.965 and 0.948, respectively. The recall of the model was 1/0.76. The precision of the model was 0.96/0.99. F1 of the model was 0.98. Table 2 provides a summary table. The performance of ENLS was 0.9536 and 0.9447 for



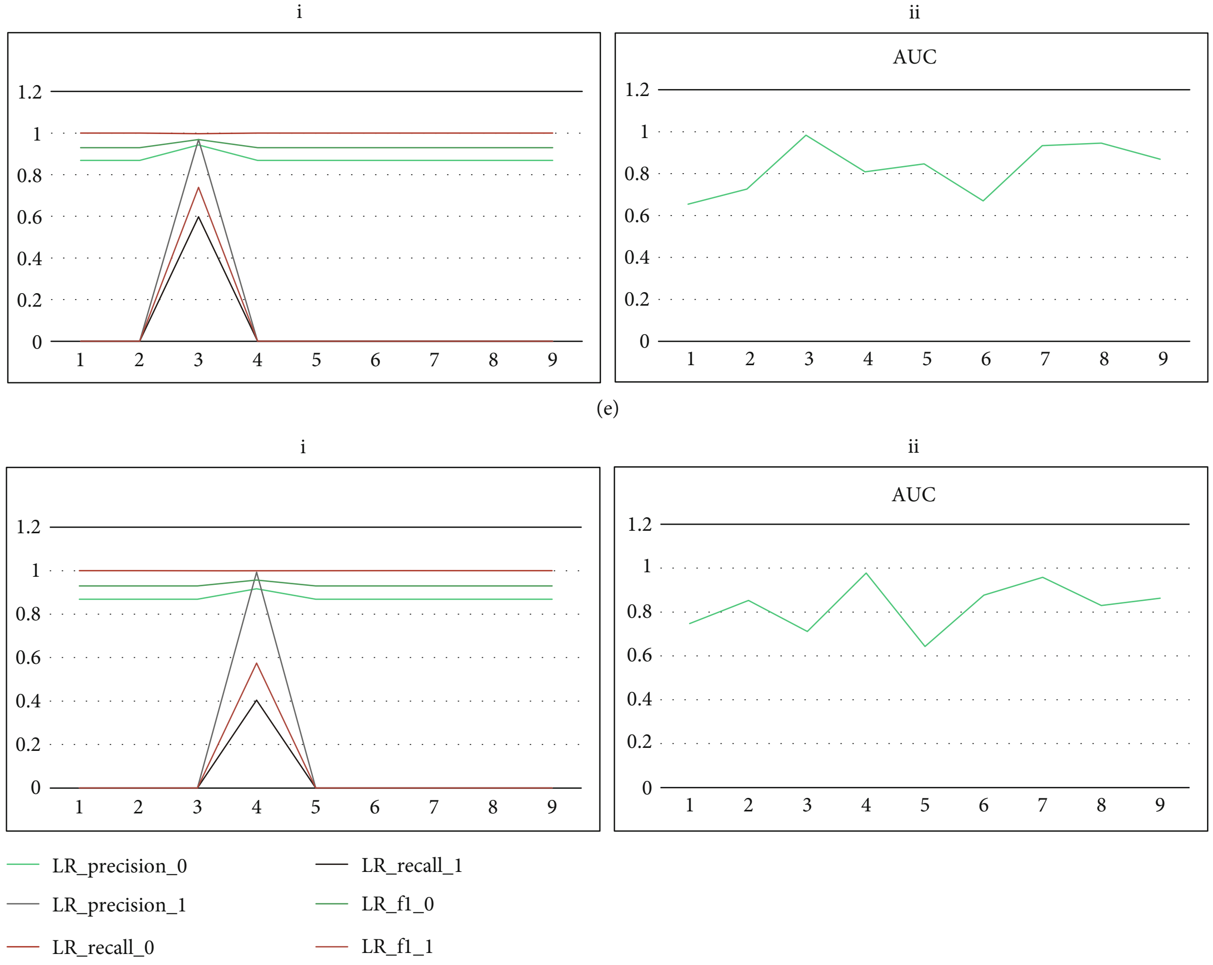
(a)

(b)

(c)

(d)

Figure 3: Continued.



(e)

(f)

Figure 3: Comparison with different parameters. (a) ENLS ($a = 0.5$, $k = 7$): (i) precision, recall, and F1-score and (ii) AUC. (b) ENLS ($a = 0.5$, $k = 0.4$): (i) precision, recall, and F1-score and (ii) AUC. (c) ENLS ($a = 0.5$, $k = 0.07$): (i) precision, recall, and F1-score and (ii) AUC. (d) ENLS ($a = 0.6$, $k = 3$): (i) precision, recall, and F1-score and (ii) AUC. (e) ENLS ($a = 0.6$, $k = 0.3$): (i) precision, recall, and F1-score and (ii) AUC. (f) ENLS ($a = 0.6$, $k = 0.04$): (i) precision, recall, and F1-score and (ii) AUC.

training and testing, respectively. The recall values were 1 and 0.6. The precision values were 0.94 and 0.97. F-val values of 0.97 and 0.74 were measured. The ROC curve is shown in Figure 2. When the performance of the ENLS model was compared with that of the TF-IDF, the results were -0.0123 (train) and -0.0033 (test). The ROC curve was approximately -0.0059. In TF-IDF, the difference between the training and testing data was 0.0179, whereas the difference in the predicted values between the training and testing data in ENLS was 0.0089 on average. Therefore, it can be seen that, although ENLS had slightly lower accuracy than TF-IDF, the ENLS system performed emotion-based loss function evaluation with L2 appropriately.

4.3. Parametric Analytics. K is for the positive and negative components of the word and nonlinear functions. The experimental results are presented in Tables 2 and 3. First, when the parameters were digitized in the ENLS system, the data measured based on the optimal alpha value of 0.5 were 0.9320 (training) and 0.9318 (testing). Recall values were 1 and 0.5. The precision values were 0.93 and 0.97. F-val values of 0.66 and 0.98 were measured. The change in value is shown (see Figure 3). In general, all values, including the AUC value, showed an increasing trend as the value of k increased (see Table 2, ENLS 1). Second, their results were 0.9361 (training) and 0.9358 (testing). The recall values were 1 and 0.52. The precision values were 0.93 and 0.98. F-val values of 0.96 and 0.68 were measured. The change in value is shown in Figure 3. All values, including the AUC value, were recorded with the highest k values at 2 and 4 (see Table 2, ENLS 2). Third, they were 0.9377 (training) and 0.9372 (testing). The recall values were 1 and 0.53. The precision values were 0.93 and 0.98. F-val values of 0.97 and 0.69 were measured. The change in value is shown in Figure 3. All values, including the AUC value, increased proportionally as k increased (see Table 2, ENLS 3). Fourth, the data measured based on the optimal alpha value of 0.6 were 0.9421 (training) and 0.9386 (testing). The recall values were 1 and 0.53. The precision values were 0.94 and 0.98. F-val values of 0.97 and 0.70 were measured. The change in value is shown in Figure 3. All values, including the AUC value, recorded the highest point at k of 3 (see Table 3, ENLS 1). Fifth, the data measured based on the optimal alpha value of 0.6 were 0.9536 (training) and 0.9447 (testing). The recall values were 1 and 0.6. The precision values were 0.94 and 0.97. F-val values of 0.97 and 0.74 were measured. The change in value is shown in Figure 3. All values, including the AUC value, peaked at k of 3 (see Table 3, ENLS 2). Sixth, the data measured based on the optimal alpha value of 0.6 were 0.9279 (training) and 0.9214 (testing). The recall values were 1 and 0.4. Precision values of 0.92 and 1 were measured. F-val values of 0.96 and 0.57 were measured. The change in value is shown in Figure 3. All values, including the AUC value, recorded the highest point at k of 4 (see Table 3, ENLS 3).

5. Conclusions

In this paper, a feature creation ENLS model was presented. Specifically, we used positive and negative vectors in spam data and proposed a nonlinear function transformation and error function system. We also found that emotions influence spam data. To evaluate the performance, a comparison with the well-known TF-IDF model was performed. Recall values were 1.00/0.6, and precision was 0.94/0.97. The F1-val value was 0.97/0.74. The model's ROC peak was 98.45%. As a result of the experiment, the values of training data and evaluation data of the ENLS model were 0.0123/0.0033 lower than those of the TF-IDF model. However, as shown in Tables 2 and 3, the ENLS model generalizes well between the training and testing data without the risk of overfitting. In the result, the distribution of ENLS can be concluded that when $a = 0.5$, it was generally between 0 and 1. When the parameters are prime, noise in unnecessary parts has been reduced. It is believed that this initial study could further improve the accuracy of document classification in the future. Future work will be aimed at increasing the accuracy more effectively.

Data Availability

Spam data for the structures reported in this manuscript have been deposited with UCI University. Copies of these data can be obtained free of charge from https://archive.ics.uci.edu/ml/datasets/sms+spam+collection.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] J. Qin, Z. Zhou, Y. Tan, X. Xiang, and Z. He, "A big data text coverless information hiding based on topic distribution and TF-IDF," *International Journal of Digital Crime and Forensics*, vol. 13, no. 4, pp. 40–56, 2021.

[2] M. O. Alhawarat, H. Abdeljaber, and A. Hilal, "Effect of stemming on text similarity for Arabic language at sentence level," *PeerJ Computer Science*, vol. 7, p. e530, 2021.

[3] L. Qi, Y. Wang, J. Chen, M. Liao, and J. Zhang, "Culture under complex perspective: a classification for traditional Chinese cultural elements based on NLP and complex networks," *Complexity*, vol. 2021, Article ID 6693753, 15 pages, 2021.

[4] T. Thaher, M. Saheb, H. Turabieh, and H. Chantar, "Intelligent detection of false information in Arabic tweets utilizing hybrid Harris Hawks based feature selection and machine learning models," *Symmetry*, vol. 13, no. 4, p. 556, 2021.

[5] O. Araque and C. A. Iglesias, "An ensemble method for radicalization and hate speech detection online empowered by sentic computing," *Cognitive Computation*, pp. 1–14, 2021.

[6] A. T. Nora, "The evolution of language models applied to emotion analysis of Arabic tweets," *Information*, vol. 12, no. 2, p. 84, 2021.

[7] D. Ali, M. M. S. Missen, and M. Husnain, "Multiclass event classification from text," *Scientific Programming*, vol. 2021, Article ID 6660651, 15 pages, 2021.

[8] E. H. Mohamed and H. E. B. Wessam, "An ensemble multi-label themes-based classification for holy Qur'an verses using Word2Vec embedding," *Arabian Journal for Science and Engineering*, vol. 46, no. 4, pp. 3519–3529, 2021.

[9] P. Zhang, X. Huang, Y. Wang, C. Jiang, S. He, and H. Wang, "Semantic similarity computing model based on multi model fine-grained nonlinear fusion," *IEEE Access*, vol. 9, pp. 8433–8443, 2021.

[10] A. Mahmoud and Z. Mounir, "Semantic similarity analysis for corpus development and paraphrase detection in Arabic," *International Arab Journal of Information Technology*, vol. 18, no. 1, pp. 1–7, 2020.

[11] A. Onan and A. T. Mansur, "A term weighted neural language model and stacked bidirectional LSTM based framework for sarcasm identification," *IEEE Access*, vol. 9, pp. 7701–7722, 2021.

[12] A. M. Abirami and A. Abdulkhader, "Feature based sentiment analysis for service reviews," *Journal of Universal Computer Science*, vol. 22, no. 5, pp. 650–670, 2016.

[13] V. L. Twan, "L2 regularization versus batch and weight normalization," 2017, https://arxiv.org/abs/1706.05350.

[14] T. Bruno, M. Sasa, and D. Donko, "KNN with TF-IDF based framework for text categorization," *Procedia Engineering*, vol. 69, pp. 1356–1364, 2014.

[15] S. J. Press and S. Wilson, "Choosing between logistic regression and discriminant analysis," *Journal of the American Statistical Association*, vol. 73, no. 364, pp. 699–705, 1978.

[16] C. Peter and K. Jacobs, "The importance of the loss function in option valuation," *Journal of Financial Economics*, vol. 72, no. 2, pp. 291–318, 2004.

[17] J. Y. Park and C. W. Seo, "Analysis of change characteristics of housing market using TF-IDF weight model," *Real Estate Gazette*, vol. 63, no. 63, pp. 46–58, 2015.

[18] S. A. Kim, H. S. Jo, and H. A. Lee, "Automatic classification of blog posts using various term weighting," *Korean Journal of Marine Engineering*, vol. 39, no. 1, pp. 58–62, 2015.

[19] J. Hoon, "Analysis of the effect of weighting considering document characteristics based on MLP on document summary performance," *Korean Society of Electronics Engineers Conference*, vol. 1330, no. 1333, 2013.

[20] K. J. Son and S. J. Lee, "A method for weighting compound nouns in patent literature searches," *Proceedings of the Korean Society of Information Sciences*, vol. 31, no. 1B, pp. 895–897, 2004.

[21] Y. C. Jeong, Y. J. Choi, and S. H. Maeng, "A study on negation handling and term weighting schemes and their effects on mood-based text classification," *Cognitive Science*, vol. 19, no. 4, pp. 477–497, 2008.

[22] M. S. Kim and S. S. Kang, "Design and implementation of malicious comment identification system using SVM," in *Annual Conference on Human and Language Technology. Human and Language Technology, Pohang University*pp. 285–289, Pohang University of Science and Technology in Korea, 2006.

[23] S. C. Lemon, J. Roy, M. A. Clark, P. D. Friedmann, and W. Rakowski, "Classification and regression tree analysis in public health: methodological review and comparison with logistic regression," *Annals of Behavioral Medicine*, vol. 26, no. 3, pp. 172–181, 2003.

[24] S. Dreiseitl and O. M. Lucila, "Logistic regression and artificial neural network classification models: a methodology review," *Journal of Biomedical Informatics*, vol. 35, no. 5-6, pp. 352–359, 2002.

[25] S. L. Ting, W. H. Ip, and H. T. Albert, "Is Naive Bayes a good classifier for document classification," *International Journal of Software Engineering and Its Applications*, vol. 5, no. 3, pp. 37–46, 2011.

[26] M. Pal, "Random forest classifier for remote sensing classification," *International Journal of Remote Sensing*, vol. 26, no. 1, pp. 217–222, 2005.

[27] J. W. Hwang and Y. J. Koh, "A document sentiment classification system based on an improved feature weighting technique that reflects the sentence sentiment intensity," *Journal of the Information Science Society: Software and Applications*, vol. 36, no. 6, pp. 491–497, 2009.

[28] H. S. Park, M. Lee, S. Hwang, and S. Oh, "TF-IDF based association rule analysis system for medical data," *KIPS Transactions on Software and Data Engineering*, vol. 5, no. 3, pp. 145–154, 2016.

[29] E. S. Yoo, K. H. Choi, and S. H. Kim, "Study on extraction of keywords using TF-IDF and text structure of novels," *Journal of the Korean Society for Computer and Information Sciences*, vol. 20, no. 2, pp. 121–129, 2015.

[30] B. Jenna, "How the machine 'thinks': understanding opacity in machine learning algorithms," *Big Data & Society*, vol. 3, no. 1, article 205395171562251, 2016.

[31] S. A. Aljuhani and N. S. Alghamdi, "A comparison of sentiment analysis methods on Amazon reviews of mobile phones," *International Journal of Advanced Computer Science & Applications*, vol. 10, no. 6, pp. 608–617, 2019.

[32] G. M. Barrientos, R. Alaiz-Rodríguez, V. González-Castro, and A. C. Parnell, "Machine learning techniques for the detection of inappropriate erotic content in text," *International Journal of Computational Intelligence Systems*, vol. 13, no. 1, pp. 591–603, 2020.

[33] J. H. Hansen and G. Liu, "Unsupervised accent classification for deep data fusion of accent and language information," *Speech Communication.*, vol. 78, pp. 19–33, 2016.

[34] N. Saidani, A. Kamel, and S. A. Mohand, "A semantic-based classification approach for an enhanced spam detection," *Computers & Security*, vol. 94, p. 101716, 2020.

[35] S. M. Abdulhamid, S. Maryam, and O. Oluwafemi, "Comparative analysis of classification algorithms for email spam detection," *International Journal of Computer Network & Information Security*, vol. 10, no. 1, pp. 60–67, 2018.

[36] M. G. H. José, A. A. Tiago, and Y. Akebo, "On the validity of a new SMS spam collection," *11th International Conference on Machine Learning and Applications*, vol. 2, pp. 240–245, 2012.

[37] W. H. Park, N. M. F. Qureshi, and D. R. Shin, "Pseudo NLP joint spam classification technique for big data cluster," *CMC-Computers, Materials & Continua*, vol. 71, no. 1, pp. 517–535, 2022.

[38] I. F. Siddiqui, N. M. F. Qureshi, B. S. Chowdhry, and M. A. Uqaili, "Pseudo-cache-based IoT small files management framework in HDFS cluster," *Wireless Personal Communications*, vol. 113, no. 3, pp. 1495–1522, 2020.

[39] N. M. F. Qureshi, I. F. Siddiqui, A. Abbas et al., "Stream-based authentication strategy using IoT sensor data in multi-homing sub-aqueous big data network," *Wireless Personal Communications*, vol. 116, no. 2, pp. 1217–1229, 2021.

[40] H. W. Choi, N. M. F. Qureshi, and D. R. Shin, "Comparative analysis of electricity consumption at home through a Silhouette-score prospective," in *2019 21st International Conference on Advanced Communication Technology (ICACT)*, pp. 589–591, IEEE, PyeongChang, Korea (South), 2019.