Hindawi
Wireless Communications and Mobile Computing
Volume 2023, Article ID 9796719, 1 page
https://doi.org/10.1155/2023/9796719

Retraction

Retracted: Text Classification Based on Machine Learning and Natural Language Processing Algorithms

Wireless Communications and Mobile Computing

Received 17 October 2023; Accepted 17 October 2023; Published 18 October 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] H. Li and Z. Li, "Text Classification Based on Machine Learning and Natural Language Processing Algorithms," *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 3915491, 12 pages, 2022.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 3915491, 12 pages
https://doi.org/10.1155/2022/3915491

Research Article

Text Classification Based on Machine Learning and Natural Language Processing Algorithms

Hui Li[1,2] and Zeming Li[2]

[1]*School of Computer and Information Engineering, Harbin University of Commerce, Harbin, 150028 Heilongjiang, China*
[2]*Heilongjiang Provincial Key Laboratory of Electronic Commerce and Information Processing, Harbin, 150028 Heilongjiang, China*

Correspondence should be addressed to Zeming Li; 161849056@masu.edu.cn

Received 27 April 2022; Revised 30 May 2022; Accepted 24 June 2022; Published 19 July 2022

Academic Editor: Chia-Huei Wu

Nowadays, with the development of media technology, people receive more and more information, but the current classification methods have the disadvantages of low classification efficiency and inability to identify multiple languages. In view of this, this paper is aimed at improving the text classification method by using machine learning and natural language processing technology. For text classification technology, this paper combines the technical requirements and application scenarios of text classification with ML to optimize the classification. For the application of natural language processing (NLP) technology in text classification, this paper puts forward the Trusted Platform Module (TPM) text classification algorithm. In the experiment of distinguishing spam from legitimate mail by text recognition, all performance indexes of the TPM algorithm are superior to other algorithms, and the accuracy of the TPM algorithm on different datasets is above 95%.

1. Introduction

Although the representation of information is getting richer and richer, so far, the main representation of information is still text. On the one hand, because text is the most natural form of information representation, it is easily accepted by people. On the other hand, due to the low cost of text representation, driven by the advocacy of paperless office, a large number of electronic publications, digital libraries, e-commerce, etc. have appeared in the form of text. In addition, with the rapid development of the global Internet in recent years, a large number of social networking sites, mobile Internet, and other industries have emerged.

From a global perspective, the number of websites will continue to grow, which will inevitably generate an even greater amount of information. Because the amount of text data is so large, while providing people with more usable information, it also makes it more difficult for people to find the information that interests them most. That is to say, information explosion leads to information trek. Therefore, how to dig out important information from massive information has very high research value and practical significance. Due to the different needs of users, how to excavate the characteristics of different users and find exclusive information for them has become the main problem that should be solved in current information processing. The text classification technology using artificial intelligence algorithms can automatically and efficiently perform classification tasks, greatly reducing cost consumption. It plays an important role in many fields such as sentiment analysis, public opinion analysis, domain recognition, and intent recognition.

In this paper, the first chapter briefly describes the current situation of natural language processing and machine learning. The second chapter is the research of related work, summarizing the advantages and disadvantages of other scholars' natural language processing algorithms. The third chapter describes the text classification algorithm in detail, paving the way for the subsequent algorithm. In Chapter 4, aiming at the adaptive algorithm of deep learning and intelligent learning technology, the existing natural

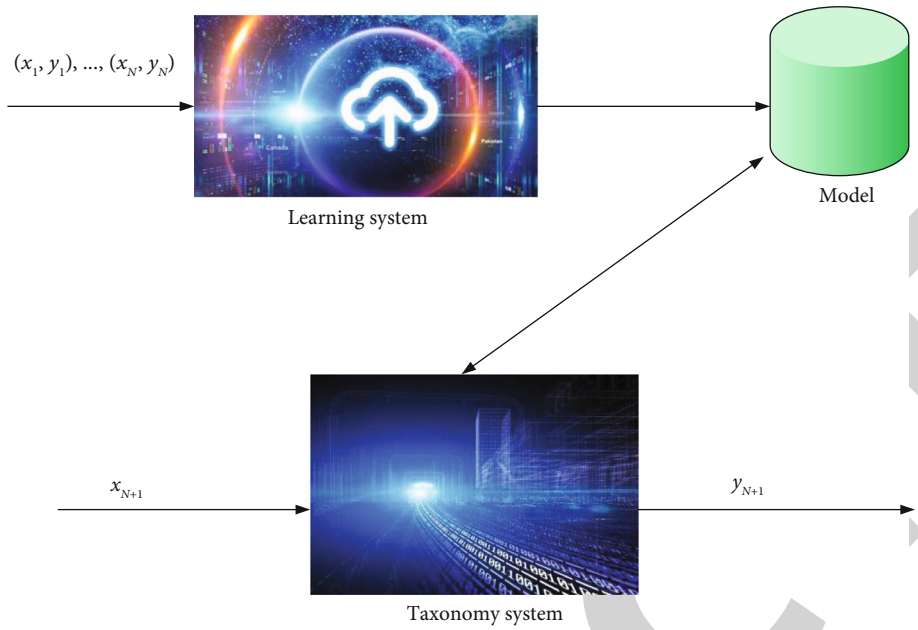

Figure 1: Description of the classification problem.

language algorithm is improved, and TPM algorithm is proposed and introduced. The fifth chapter analyzes the performance of the proposed algorithm. Finally, summarize the full text.

2. Related Work

So far, text classification technology has been widely used in information filtering, mail classification, search engines, query intent prediction, topic tracking, text corpus construction, and other fields. It can help users to accurately classify messy data to obtain classified text information and solve the problem of rapid positioning of information required by users. A large number of researchers in both academia and industry have begun to pay attention to this direction, which not only promotes academic development but also promotes the R&D and promotion of corresponding products. Mohamed et al. proposed a novel active learning method for text classification in order to solve the problem of manually labeling data samples during the training phase. The experimental results show that the proposed active learning method significantly reduces the labeling workload while improving the classification accuracy [1]. Mironczuk and Protasiewicz designed a semisupervised learning Universum algorithm based on boosting technology, mainly for the case of only a small number of labeled samples. Their experiments used four datasets and several combinations. Experimental results show that the algorithm can benefit from Universum samples and outperform several other methods, especially when the labeled samples are insufficient [2]. The purpose of Liu et al. was to extract state-of-the-art features for text classification. They believed that this study will help readers obtain the necessary information about these elements and their associated technologies [3]. Pavlinek and Podgorelec proposed a topic model-based approach for semisupervised text classification. The proposed method includes a self-training and model-based semisupervised text classification algorithm that determines parameter settings for any new collection of documents [4]. Kobayashi et al. are very interested in data mining technology; they believe that text classification technology can help in data mining technology [5]. He believes that today's athletes cannot avoid injuries during training, so he hopes to model the sports training of athletes and reduce sports injuries of athletes [6]. Shah et al. considered an improved version of a semisupervised learning algorithm for graph-structured data to address the problem of scaling deep learning methods to represent graph data [7]. Anoual and Zeroual's research on Arabic is very in-depth. They believe that Arabic is more complex than other languages on the Internet, and it is not so easy to accurately display and translate them on the Internet, so they study Arabic text classification technology and conduct research on Arabic word combinations [8]. However, through related research, it can also be found that although text is widely used by technology, it lacks real optimization, and in the era of big data, it is less combined with ML.

3. Text Classification Technology

The classification problem includes two processes: learning and classification. The goal of the learning process is to build a classification model based on the known training data to obtain a classifier. The task of the classification process is to use the learned classifier to predict the class label of a new data instance. Figure 1 is a descriptive diagram of the classification problem.

In the figure, $(x_1, y_1), \cdots, (x_N, y_N)$ represents the training dataset that has been labeled with classes, x_i represents the data instance, and y_i represents the class label corresponding

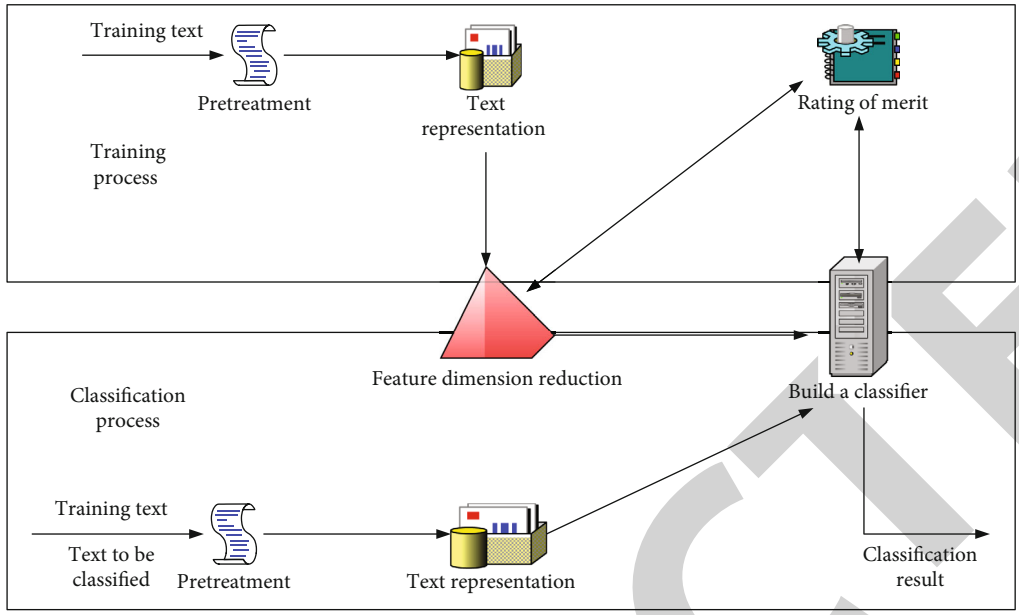

Figure 2: Text classification framework.

to x_i. The learning system is based on the training data, from which it learns a classifier $P(Y \mid X)$ or $Y = f(X)$. The classification system classifies a new input instance x_{N+1} with the already obtained classifier to predict the class label y_{N+1} of its output [9, 10].

A text classification problem is a guided learning process where the object is text and the task is to automatically classify new input text into one or more predefined categories; each text object may belong to one or more categories.

There is an unknown mapping function $\Phi : D \times C \longrightarrow \{T, F\}$ between the text set and the category set, where $D = \{d_1, d_2, \cdots, d_{|D|}\}$ represents the document set to be classified, and $C = \{c_1, c_2, \cdots, c_{|C|}\}$ represents the predefined category set. For each given data pair $< d_j, c_i >$, there are two values, a value of T indicates that document d_j belongs to category c_i, and a value of F indicates that d_j does not belong to c_i. That is to say, through the learning process, obtaining the optimal estimation of the target mapping function is what should be considered in the text classification task, which is also called the classifier.

The text classification framework is shown in Figure 2, which includes the basic problems that need to be solved.

As shown in Figure 2, the main functional modules of the text classification system are briefly described as follows: *Preprocessing*: in order to improve the quality of text representation and facilitate subsequent processing, preprocessing operations such as formatting are required for the original text corpus. *Text representation*: the problems that need to be solved in text representation include the following: First, what language elements should be selected as text features, most of which are words or phrases. The second is to choose what model to quantify text objects. *Feature dimensionality reduction*: in order to achieve text classification, it is necessary to select features from the text that can best reflect the subject of the document. *Building a classifier*: how to design a text classifier is the main research content of text classification methods. First, the text that can represent each category in the classification system is selected as the training set, the classifier is learned from the training set, and the classification of new objects is realized. *Performance evaluation*: the purpose of this step is to evaluate the pros and cons of the classification method and system performance. Different evaluation parameters can be used for different classification problems; for example, single-label classification and multi-label classification problems will use different parameters. Text classifier performance evaluation methods include recall rate, accuracy rate, F-value, microaverage, and macroaverage, so as to improve the performance of the classification system.

4. Improved Text Classification Algorithm Based on ML and NLP Algorithms

4.1. TMP Text Classification Algorithm. LSTM and CNN models are more commonly used neural network models. Their combination can create many possibilities [11, 12]. Based on this, this paper proposes a text classification algorithm model as shown in Figure 3.

As shown in Figure 3, the vector x can be obtained at the embedding layer:

$$X = \{x_0, \cdots, x_i, \cdots, x_{n-1}\} = \text{Embedding}(a_0, \cdots, a_{i+1}, \cdots, a_n). \quad (1)$$

The second step is to transmit the data of the input layer down, such as the following formula:

$$H = \left[\overrightarrow{\text{LSTM}}(X), \overleftarrow{\text{LSTM}}(X)\right]. \quad (2)$$

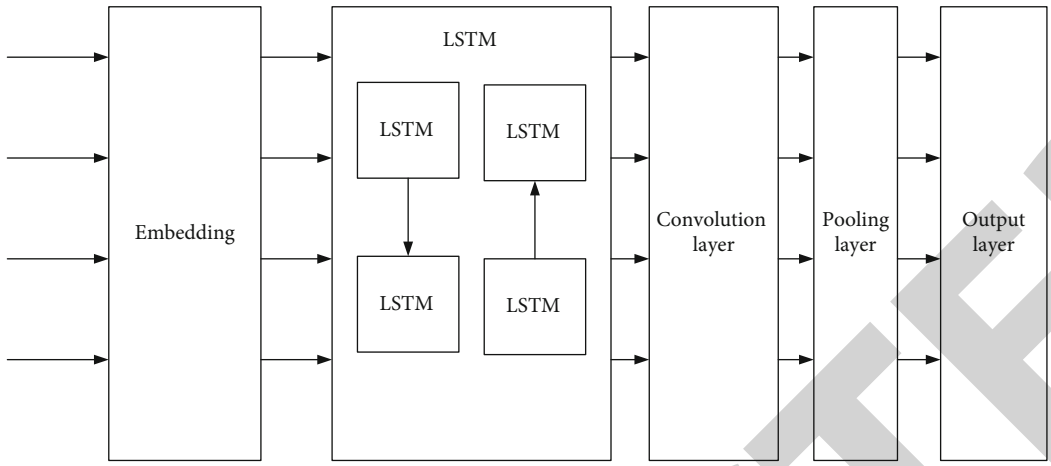

Figure 3: MS-KNN model classification algorithm.

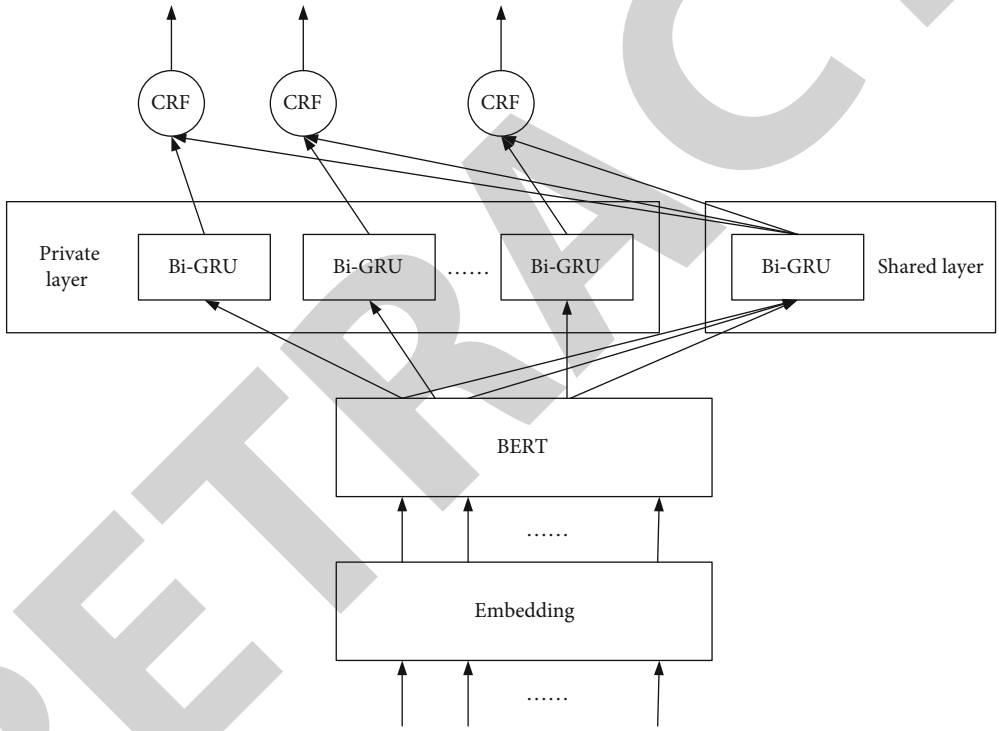

Figure 4: TPM Chinese word segmentation model.

The third step is to extract the corresponding features as in the following equations:

$$c_i = \mathrm{Conv}(W_i, H), \tag{3}$$

$$p_i = \mathrm{pooling}(C_i). \tag{4}$$

The final output data is shown in the following equation:

$$y_k = \text{soft max}\ (W_k P + b_k). \tag{5}$$

W_k represents the linear parameters of the fully connected layer, b represents the bias, and y_k represents the probability that the text belongs to a certain class. The specific model structure is shown in Figure 4.

As can be seen from Figure 4, the input of the TPM Chinese word segmentation model is still a piece of preprocessed Chinese text [13, 14]. First, through the embedding layer of the model, the natural language is converted into a text vector that can be recognized by the computer. Then, the powerful semantic feature extraction ability of the BERT model is used to extract semantic features, which is equivalent to reencoding the text according to the context semantics. Then, according to the original dataset where the input data is located, the semantic feature vector is inputted into the corresponding Bi-GRU model of the private layer. It is used to extract the unique features of the dataset

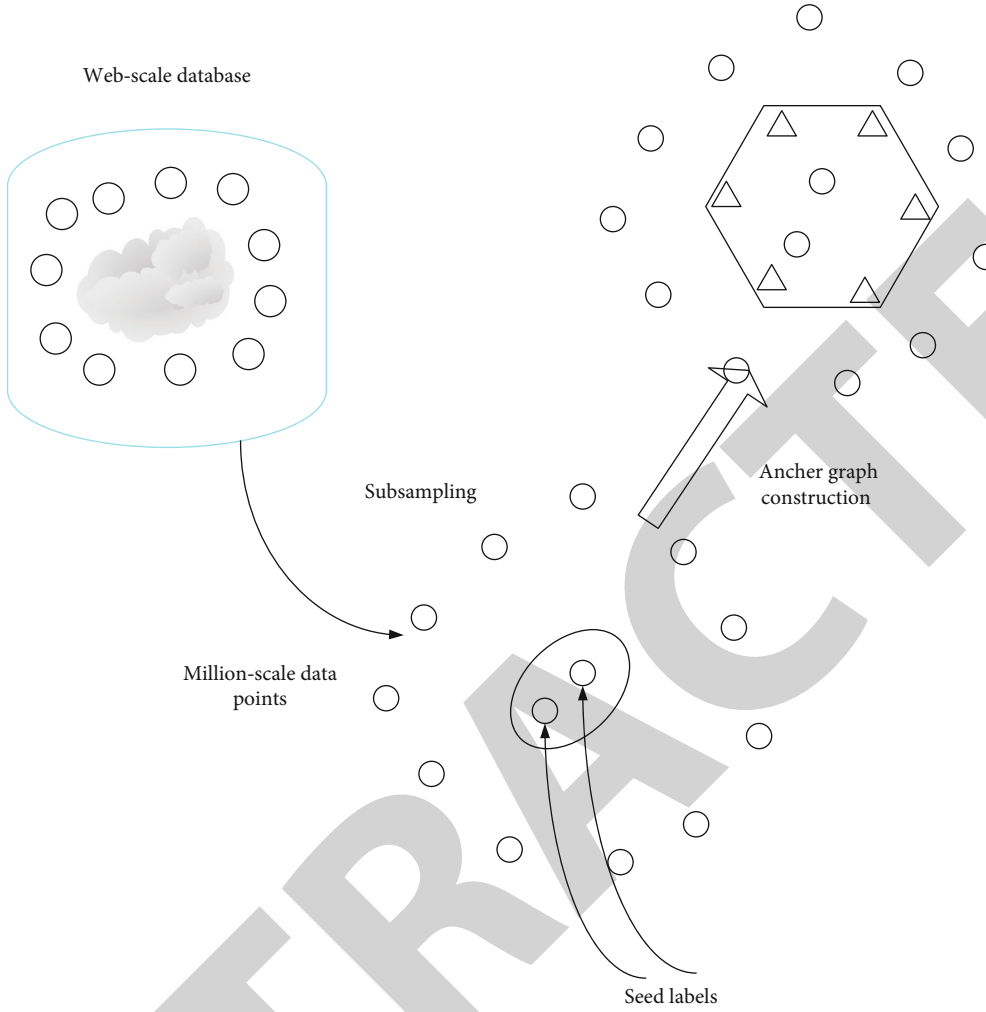

Figure 5: Schematic diagram of the anchor map-based label propagation method.

compared to other datasets. At the same time, the semantic feature vector is inputted into the Bi-GRU model of the shared layer, which is used to extract common features of multiple datasets. Finally, the private features of the data are combined with the public features and put into the corresponding CRF model of the inference layer to obtain the label of each character in the text. Finally, according to the label of each character, the input text is divided into a sequence of words and output, and the word segmentation operation of the data is completed by the model.

The text classification algorithm includes active learning stage and mainstream active learning methods. Among the pool-based active learning methods, uncertainty sampling is one of the simplest and most commonly used query frameworks. Typical uncertain sampling methods include least confident (LC), margin sampling (MS), entropy sampling (ES), and centroid sampling (CS). In this paper, Edge MS is chosen as the active learning algorithm because of its excellent performance in mail classification.

The second stage is the text classification stage. Traditional text classification algorithms include naive Bayes, k -nearest neighbor, and support vector machine. In this paper, k-nearest neighbor (KNN) is selected to compare with support vector machine (SVM), among which k-nearest neighbor (KNN) [105] is simple and intuitive, without explicit learning process and offline training of classification models. Its basic idea is as follows: given the training set, in which the data category has been determined, when a new sample to be classified is inputted, the similarity measurement method is used to determine the similarity between the new sample and the training data, and then, the nearest K samples are found from the training set, and the prediction is made by majority voting. Support vector machine (SVM) is a widely used text classification method. It is a machine learning method based on statistical learning theory. It was first proposed for binary classification problems. For multiclassification problems, it is necessary to build multiple classifiers. When constructing a binary SVM classifier, its core task is to find an optimal hyperplane from countless classification interfaces, which is also called the decision plane. It can best distinguish the samples in two categories, and the distance between different categories and this plane is the largest. From the geometric point of view, this hyperplane divides the input space into positive and negative spaces, which is a line in two-dimensional space and a surface in three-dimensional space.

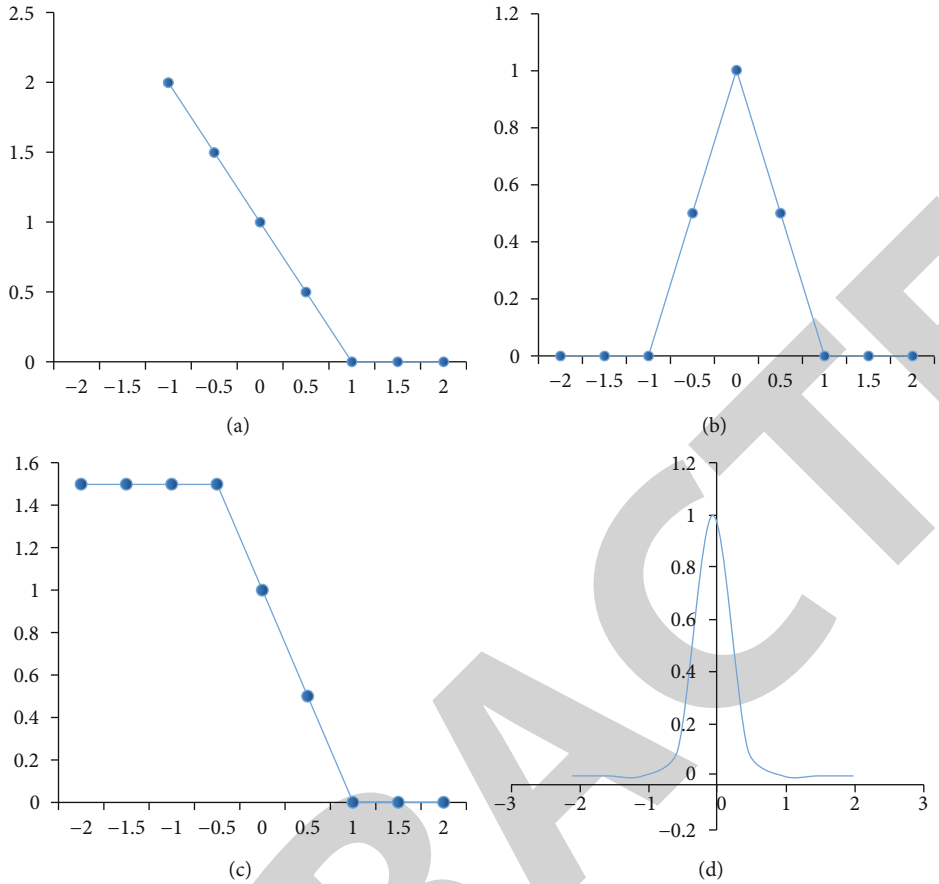

Figure 6: Schematic diagram of loss function.

The TPM algorithm in this paper is applied to the research of Chinese word segmentation in multitask learning. In order to speed up the training and further enhance the extraction of text semantic information, based on the original multistandard word segmentation model, a new multistandard word segmentation model TPM based on POS_LSTM (Particle Swarm Optimization, Long-term and Short-term Network Memory Model) is improved. In the active learning stage and text classification stage, the TMP algorithm and boundary sampling method proposed in this paper are compared with the MS_KNN and MS_SVM algorithms combined with k-nearest neighbor and support vector machine.

4.2. Application of NLP in Text Classification. Graph-based semisupervised learning algorithms build a graph of all data samples (labeled and unlabeled) based on their similarity, and each point on the graph represents a data sample. The edge between two nodes is generally defined by a certain similarity measure, reflecting the connection between samples. There are usually two ways to define similarity: K -adjacent and Gaussian kernel.

When building a graph, the similarity between two vertices can be defined by itself. It can be assumed that the Gaussian kernel in formula (6) is used to define it. In the process of label transmission, a probability matrix of label transmission needs to be established; the size is $(L+U)\times(L+U)$, as shown in the following formula:

$$T_{ij} + P(i \longrightarrow j) = \frac{w_{ij}}{\sum_{k=1}^{L+U} w_{ik}}. \tag{6}$$

The time complexity of the algorithm has reached a high level. The label propagation algorithm is a method of transduction learning. Every time the test set is changed, the algorithm must be run again. Therefore, for large-scale tasks, time overhead is a key factor like application promotion. Figure 5 is a schematic diagram of the anchor map-based label propagation method.

The calculation steps of the algorithm are as follows:

(1) Use the K-means algorithm to select m anchor points

(2) Use the formula to calculate the mapping relationship between sample points and anchor points (data2anchor) matrix $Z(x)$:

$$Z(x)\frac{\left[\delta_1 \exp\left(-D^2(x,u_1)/t\right),\cdots,\delta_m \exp\left(-D^2(x,u_m)/t\right)\right]^T}{\sum_{j=1}^{m}\delta_j \exp\left(-D^2\left(x,u_j\right)/t\right)}. \tag{7}$$

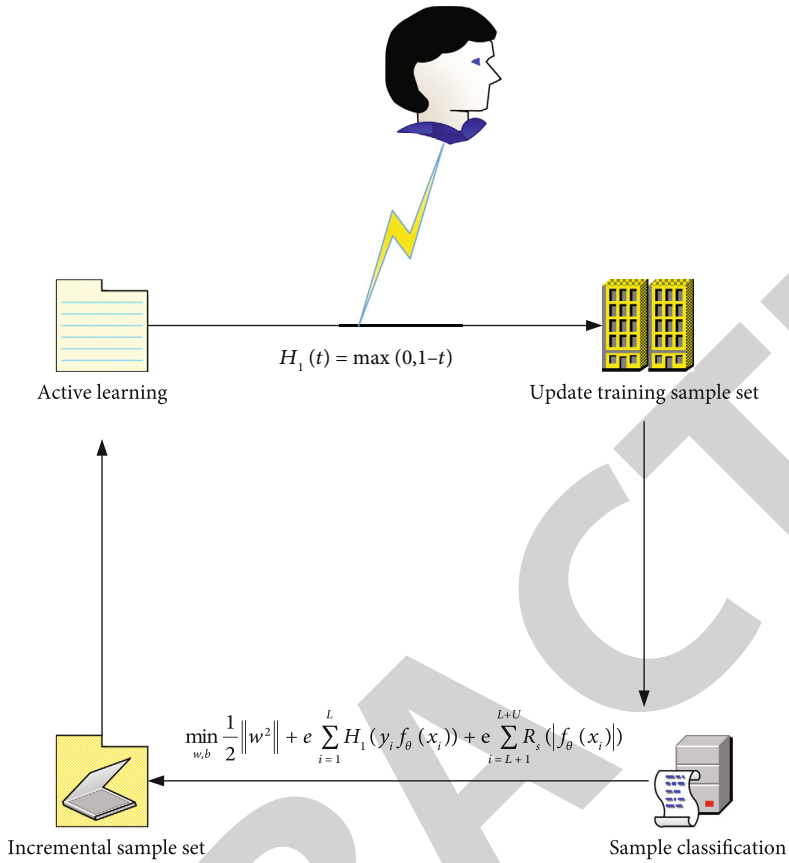

Figure 7: Schematic diagram of email recognition process.

(3) Using the AGR algorithm, the soft label matrix A^* of the anchor point is solved by the following formula:

$$A^* = \left(Z_1^T Z_1 + \gamma Z^T Z - \gamma Z^T Z \Lambda^{-1} \left(Z^T Z\right)^{-1} Z_1^T Y_1\right). \tag{8}$$

(4) According to the decision function, the formula calculates the label of the unlabeled sample:

$$\hat{y}_1 = \arg \max_{j \in (1,\cdots,C)} \frac{Z_i a_j}{\lambda_j}, \quad i = 1, \cdots, n, \tag{9}$$

where δ is the indicated value, $\delta \in (0, 1)$. D is the distance function, which can you define by yourself.

Although the algorithms reduce the time complexity of graph-based algorithms to linear, the problem of data sparseness has not been properly solved. Therefore, the algorithm has only achieved application progress in the field of image classification. We believe that if the sparsity of the task is solved, the anchor graph-based label propagation algorithm can be extended to the field of natural language processing. We take the part-of-speech tagging task as an example and try to generalize the algorithm to NLP [15, 16].

Table 1: TR07 and ES datasets.

Dataset	TR07	ES
Spam quantity	50199	17171
Legal mail quantity	25220	16545
Amount to	75419	33716

For labeled data, according to the traditional support vector machine (SVM) theory, the loss function is the hinge loss, formula (10), as shown in Figure 6(a).

$$H_1(t) = \max(0, 1 - t). \tag{10}$$

Part of the improvement proposes a smoother version of the loss function, such as formula (11), as shown in Figure 6(c):

$$S(t) = \exp\left(-3t^2\right). \tag{11}$$

In the subsequent experiments, we use HingeLoss, which has the best overall effect of efficiency and accuracy, as the loss function for labeled data, and RampLoss as the loss function for unlabeled data, as shown in Figure 6(b). Then, in the following chapters [17, 18], our optimization objective

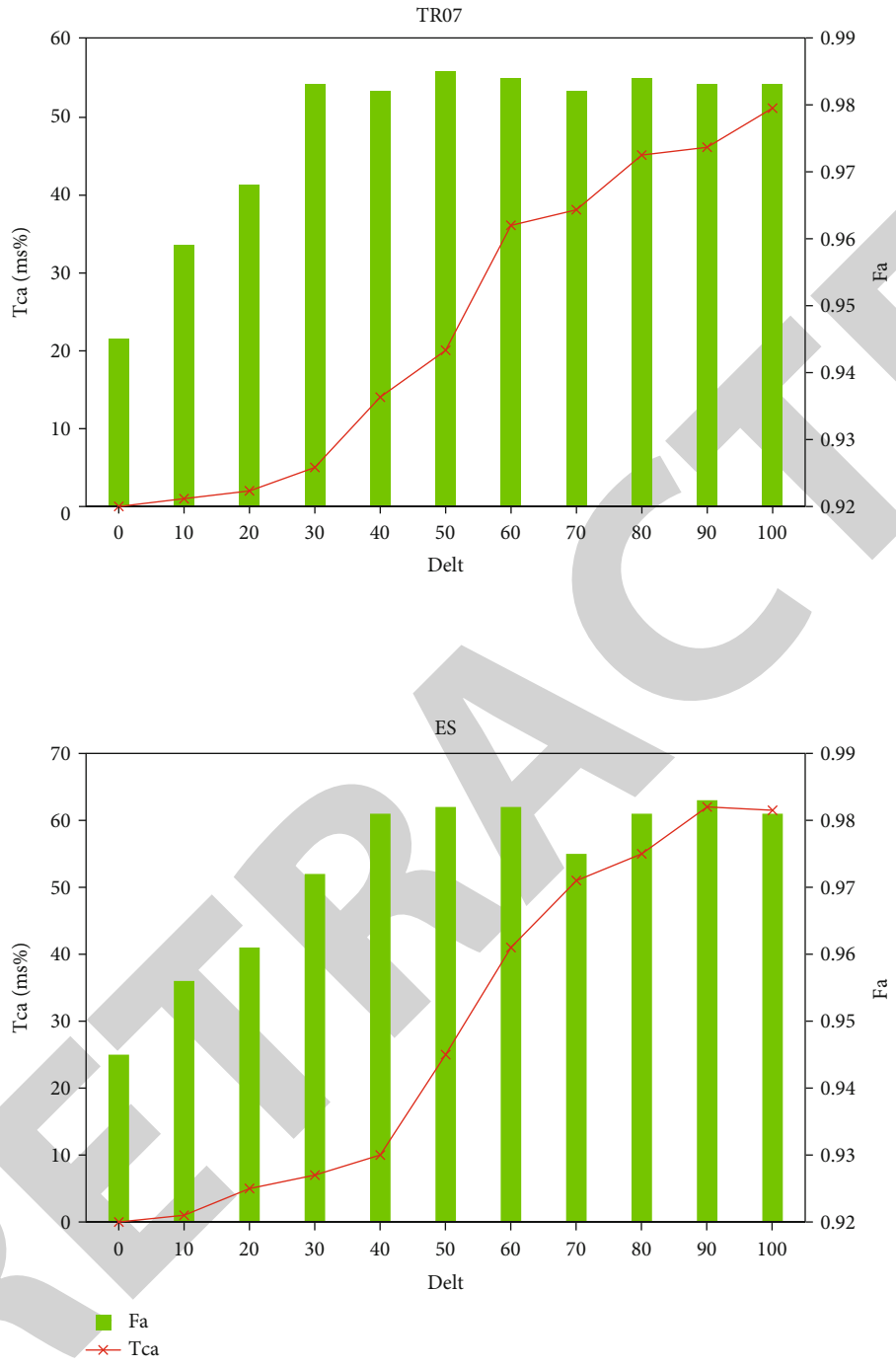

Figure 8: Fa and Tca values for TR07 and ES datasets when the value range is [0, 100].

is formula (12), as shown in Figure 6(d):

$$\min_{w,b} \frac{1}{2}\left\|w^2\right\| + e\sum_{i=1}^{L} H_1(y_i f_\theta(x_i)) + e\sum_{i=L+1}^{L+U} R_S(|f_\theta(x_i)|). \quad (12)$$

5. Text Classification Performance Test

5.1. Experimental Results and Analysis. Figure 7 is a schematic flow chart of the method for testing the text recognition and classification of emails.

To verify the effectiveness of our method, we use and conduct experiments on two benchmark datasets: TREC2007 (TR07) and Enron-spam (ES) [19, 20]. It contains spam and legitimate mail as shown in Table 1.

5.1.1. Selection of Threshold ∇. The time overhead required for classification is actually related to the value of the parameter ∇. Wanting to obtain the optimized parameter ∇, when the value of ∇ varies between 0 and 100, we conduct the corresponding statistical experiments [21, 22]. Statistical experiments were performed on the TR07 and ES datasets, and

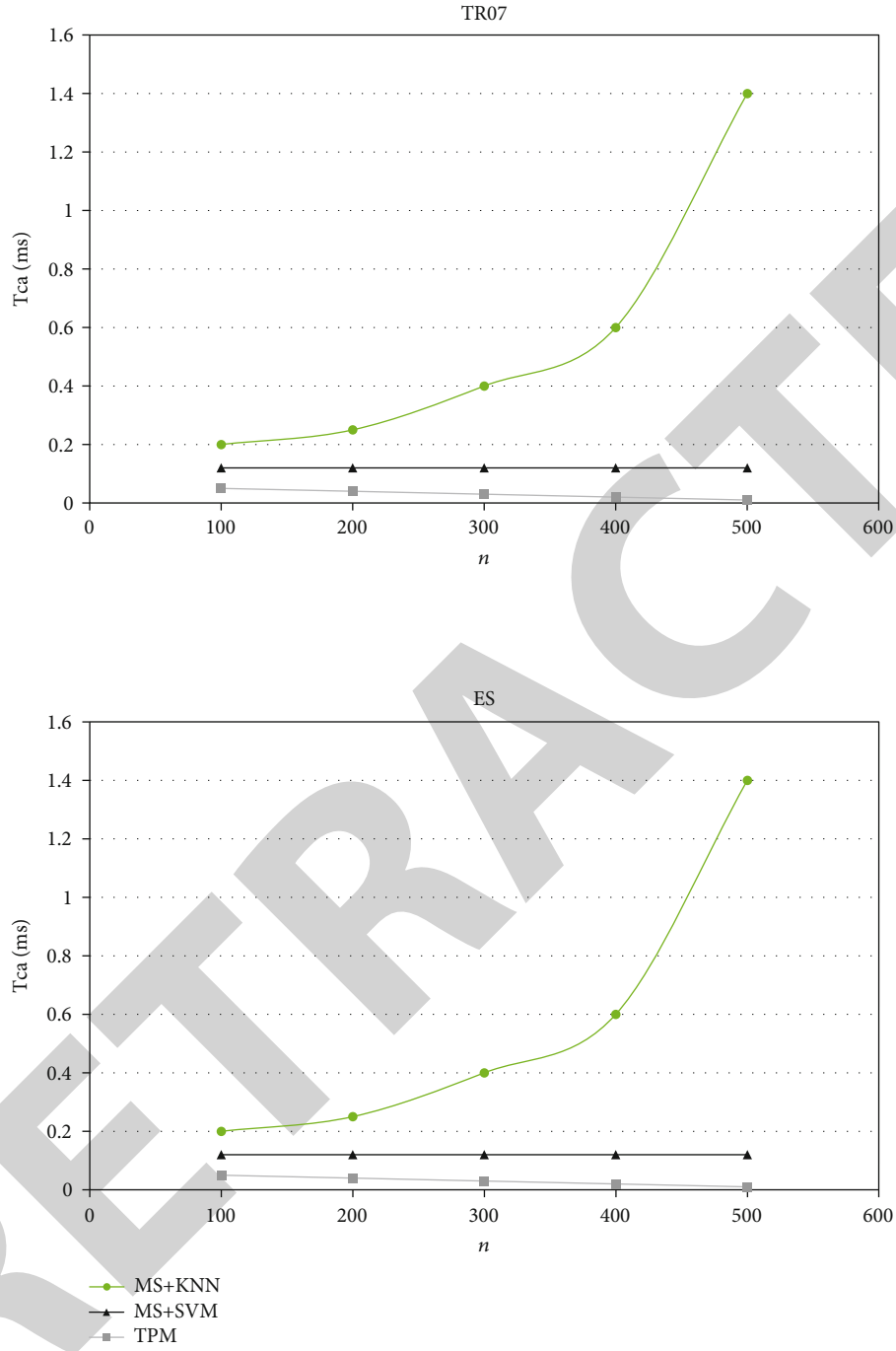

Figure 9: Tca values for TR07 and ES datasets when using different methods.

the corresponding calculated values of Fa and Tca are shown in Figure 7. ∇ is denoted as delt in Figure 8.

From Figure 8, we can see that on dataset TR07, when the value of parameter ∇ varies between the interval [0, 30], the value of Fa grows rapidly. As the value of ∇ increases further, the value of Fa tends to stabilize. In the dataset ES, when the value of parameter ∇ varies between the interval [0, 40], the value of Fa increases rapidly, and as the value of ∇ increases further, the value of Fa tends to be stable. Therefore, in order to reduce the time overhead in the sample classification process as much as possible, when using the TR07 dataset, set ∇ = 30, and when using the ES dataset, set ∇ = 40.

5.1.2. Comparative Analysis of Time Cost. Suppose that |A0 | = 200; the value of ns is set to 100, 200, 300, 400, and 500, and the value of |Si| is set to 60 and 120 [23, 24]. Figure 9 shows the value of the time cost Tca obtained when experiments are performed on the TR07 and ES datasets.

The upper part of Figure 9 corresponds to the dataset TR07, while the lower part of Figure 9 corresponds to the

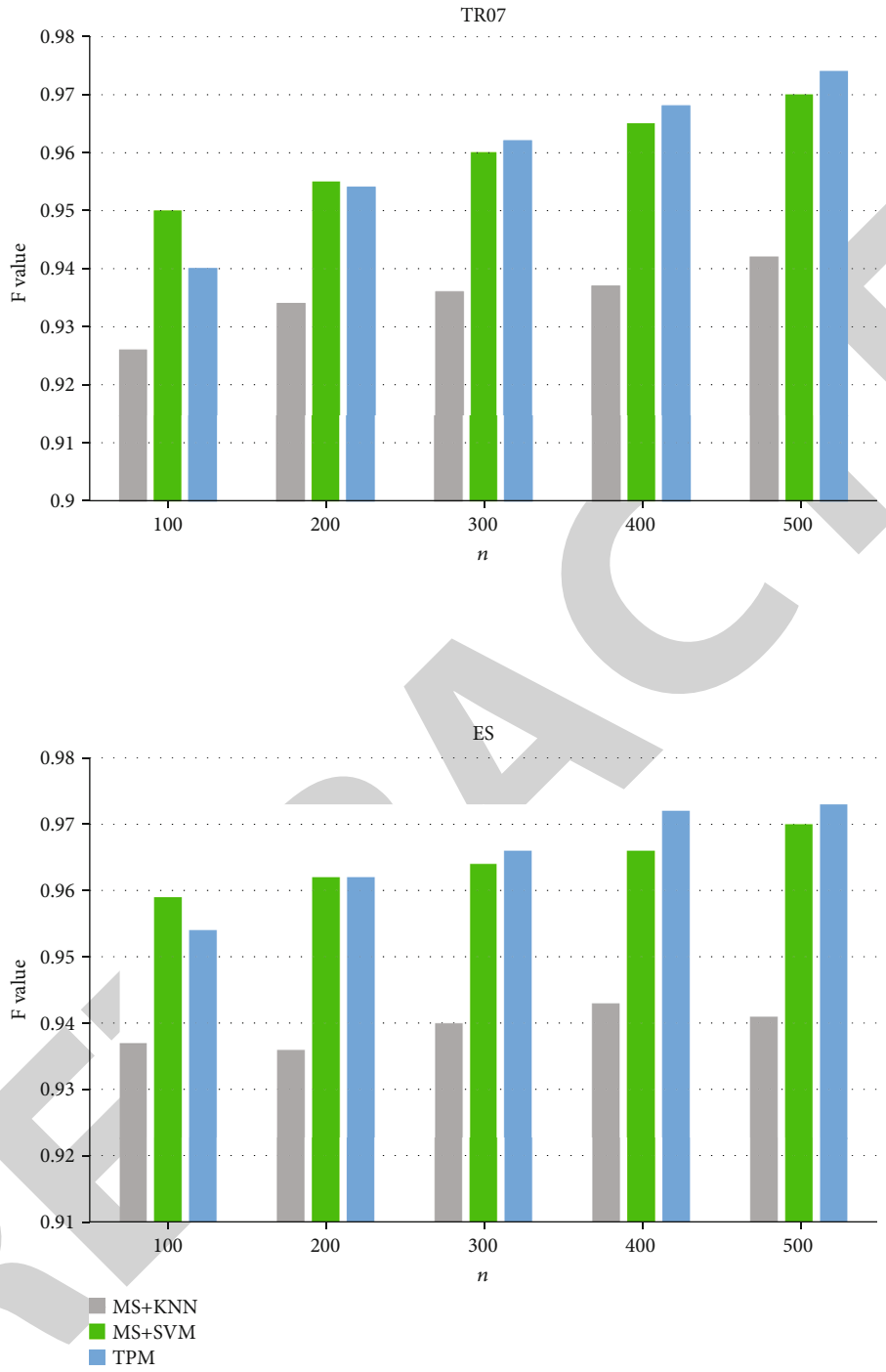

Figure 10: Fa values for TR07 and ES datasets when using different methods.

Table 2: Corresponding FM and n values when using different methods.

Dataset	ES		TR07	
Algorithm	FM	n	FM	n
MS+KNN	0.965	1825	0.967	1682
MS+SVM	0.976	1156	0.981	1215
TPM	0.979	563	0.983	697

dataset ES. So when the value of ns varies between the interval [100, 500], the value of Tca produced by the MS+KNN and ES+KNN methods combined with the KNN classification algorithm increases significantly. At the same time, we also noticed that the MS+SVM and ES+SVM methods combined with the SVM classifier have better performance in terms of computational complexity than those combined with the KNN classification algorithm [25, 26]. Likewise, in Figure 9, we can also observe that the MS+NB and ES+NB methods combined with the NB classifier have smaller Tca

values relative to the method combined with the SVM classifier. This is because the computational complexity of the NB classifier is only related to the vector dimension of the feature space. Compared with the MS+NB and ES+NB methods combined with the NB classifier, when the value of ns is greater than 300, the method in this paper obviously has the best performance. This is mainly because, in the word frequency-based user interest set method proposed in this chapter, the direct use of the SVM classifier is avoided. Therefore, it effectively reduces the average time overhead of the sample classification generated in the classification process.

5.1.3. Accuracy Comparison between Different Methods. Figure 10 shows the average values of Fa obtained by the method in this paper when experiments are performed on the TR07 dataset and the ES dataset.

It can be seen from Figure 10 that compared with other methods, the method combined with the KNN classifier performs the worst. This is due to the fact that during active learning and classification, as the value of n increases, the Fa value of TPM is getting closer and closer to those of the MS+SVM and ES+SVM methods combined with the SVM classifier, and it can be seen that the value is significantly higher than in the other methods.

5.1.4. Comparative Analysis of Sample Labeling Burden. In order to facilitate the calculation, the initialization parameters for sample labeling are given, $|A0|$ is set to 300, and $|Si|$ is set to 300. For dataset TR07 and dataset ES, the maximum value achieved by F1 in the experiment is defined as FM [27, 28], as shown in Table 2.

From the experimental results in Table 2, it can be seen that when using dataset TR07 and dataset ES, the values of the minimum FM produced by all methods on these two datasets are 0.961 and 0.964, respectively. Corresponding to different FM values, the calculated total number of samples recommended to users for labeling is defined as n in Table 2. From Table 2, we can also find that when the value of FM is not less than 0.96, compared with other methods, the n value of our method is relatively low [29].

6. Conclusion

This paper is an optimization and improvement study of the text classification algorithm. The datasets used in the experiment are the TREC2007 and Enron-spam datasets, and the classification process adopts support vector machine, naive Bayes classifier, and k-nearest neighbor classifier. The experimental results show that, for the TREC2007 and Enron-spam datasets, under the premise of less burden of sample annotation, when F1 value is used for evaluation, the proposed method also shows relatively better performance than other methods.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

This research is supported by the Natural Science Foundation of Heilongjiang Province of China (No. YQ2020G002), University Nursing Program for Young Scholar with Creative Talents in Heilongjiang Province (No. UNPYSCT-2020212), and Science Foundation of Harbin Commerce University (No. 2019CX22 and No. 18XN064).

References

[1] M. Goudjil, M. Koudil, M. Bedda, and N. Ghoggali, "A novel active learning method using SVM for text classification," *International Journal of Automation and Computing*, vol. 15, no. 3, pp. 290–298, 2018.

[2] M. M. Mironczuk and J. Protasiewicz, "A recent overview of the state-of-the-art elements of text classification," *Expert Systems with Applications*, vol. 106, pp. 36–54, 2018.

[3] C. L. Liu, W. H. Hsaio, C. H. Lee, T. H. Chang, and T. H. Kuo, "Semi-supervised text classification with Universum learning," *IEEE Transactions on Cybernetics*, vol. 46, no. 2, pp. 462–473, 2016.

[4] M. Pavlinek and V. Podgorelec, "Text classification method based on self-training and LDA topic models," *Expert Systems with Applications*, vol. 80, pp. 83–93, 2017.

[5] V. B. Kobayashi, S. T. Mol, H. A. Berkers, G. Kismihók, and D. N. den Hartog, "Text classification for organizational researchers: a tutorial," *Organizational Research Methods*, vol. 21, no. 3, pp. 766–799, 2018.

[6] K. He, "Prediction model of juvenile football players' sports injury based on text classification technology of ML," *Mobile Information Systems*, vol. 2021, Article ID 2955215, 10 pages, 2021.

[7] S. M. Shah, H. Ge, S. A. Haider et al., "A quantum spatial graph convolutional network for text classification," *Computer Systems Science and Engineering*, vol. 36, no. 2, pp. 369–382, 2021.

[8] E. K. Anoual and I. Zeroual, "The effects of pre-processing techniques on Arabic text classification," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 10, no. 1, pp. 41–48, 2021.

[9] J. Atwan, M. Wedyan, Q. Bsoul, A. Hamadeen, R. Alturki, and M. Ikram, "The effect of using light stemming for Arabic text classification," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 5, pp. 768–773, 2021.

[10] H. Amazal and M. Kissi, "A new big data feature selection approach for text classification," *Scientific Programming*, vol. 2021, no. 2, 10 pages, 2021.

[11] Q. Wang, W. Li, and Z. Jin, "Review of text classification in deep learning," *Open Access Library Journal*, vol. 8, no. 3, pp. 1–8, 2021.

[12] X. Luo, "Efficient English text classification using selected machine learning techniques," *AEJ-Alexandria Engineering Journal*, vol. 60, no. 3, pp. 3401–3409, 2021.

[13] T. Salles, M. Goncalves, V. Rodrigues, and L. Rocha, "Improving random forests by neighborhood projection for effective

text classification," *Information Systems*, vol. 77, pp. 1–21, 2018.

[14] S. F. Yin, H. Zheng, S. H. Xu, H. Rong, and N. Zhang, "A text classification algorithm based on feature library projection," *Journal of Central South University*, vol. 48, no. 7, pp. 1782–1789, 2017.

[15] F. A. Wenando, T. B. Adji, and I. Ardiyanto, "Text classification to detect student level of understanding in prior knowledge activation process," *Advanced Science Letters*, vol. 23, no. 3, pp. 2285–2287, 2017.

[16] W. Cao, A. Song, and J. Hu, "Stacked residual recurrent neural network with word weight for text classification," *IAENG International Journal of Computer Science*, vol. 44, no. 3, pp. 277–284, 2017.

[17] S. Bahassine, A. Madani, and M. Kissi, "Arabic text classification using new stemmer for feature selection and decision trees," *Journal of Engineering Science and Technology*, vol. 12, no. 126, pp. 1475–1487, 2017.

[18] S. Yu, D. Liu, W. Zhu, Y. Zhang, and S. Zhao, "Attention-based LSTM, GRU and CNN for short text classification," *Journal of Intelligent and Fuzzy Systems*, vol. 39, no. 1, pp. 333–340, 2020.

[19] T. Hernandez-Boussard, P. Kourdis, R. Dulal et al., "A natural language processing algorithm to measure quality prostate cancer care," *Journal of Clinical Oncology*, vol. 35, Supplement_8, pp. 232–232, 2017.

[20] Z. Kong, C. Yue, Y. Shi, J. Yu, C. Xie, and L. Xie, "Entity extraction of electrical equipment malfunction text by a hybrid natural language processing algorithm," *IEEE Access*, vol. 9, no. 99, pp. 40216–40226, 2021.

[21] Y. Gong, N. Lu, and J. Zhang, "Application of deep learning fusion algorithm in natural language processing in emotional semantic analysis," *Concurrency & Computation Practice & Experience*, vol. 31, no. 10, pp. e4779.1–e4779.9, 2019.

[22] W. H. Weng, K. B. Wagholikar, A. T. McCray, P. Szolovits, and H. C. Chueh, "Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach," *Bmc Medical Informatics & Decision Making*, vol. 17, no. 1, pp. 155–167, 2017.

[23] D. G. Morgan, K. Chorneyko, D. Swain, B. Bowes, V. Lee, and J. Tinmouth, "279 - validation of a natural language processing algorithm to identify colonic adenomas across a health system," *Gastroenterology*, vol. 156, no. 6, p. S-56, 2019.

[24] J. M. Ehrenfeld, K. G. Gottlieb, L. B. Beach, S. E. Monahan, and D. Fabbri, "Development of a natural language processing algorithm to identify and evaluate transgender patients in electronic health record system," *Ethnicity & Disease*, vol. 29, Supplement 2, pp. 441–450, 2019.

[25] C. L. Wi, S. Sohn, M. C. Rolfes et al., "Application of a natural language processing algorithm to asthma ascertainment: an automated chart review," *American Journal of Respiratory and Critical Care Medicine*, vol. 196, no. 4, pp. 430–437, 2017.

[26] S. Triputra and F. Atqiya, "Implementation of natural language processing in seller-bot for SMEs," *Journal of Physics Conference Series*, vol. 1764, no. 1, pp. 012069–012075, 2021.

[27] J. S. Kim, V. Arvind, J. T. Schwartz et al., "P72. Natural language processing of operative note dictations to automatically generate CPT codes for billing," *The Spine Journal*, vol. 20, no. 9, pp. S181–S182, 2020.

[28] R. W. Chang, L. Y. Tucker, K. A. Rothenberg et al., "Establishing a carotid artery stenosis disease cohort for comparative effectiveness research using natural language processing," *Journal of Vascular Surgery*, vol. 68, no. 3, pp. e32–e33, 2018.

[29] N. Afzal, V. P. Mallipeddi, S. Sohn et al., "Natural language processing of clinical notes for identification of critical limb ischemia," *International Journal of Medical Informatics*, vol. 111, pp. 83–89, 2018.