An Adaptive SVR-Based Framework for Multimodal Corpus Classification

Yuhui Wang
Basic Teaching Department, Henan Logistics Vocational College, Zhengzhou 453500, China

***Abstract*—To address the challenges associated with the dynamic growth and multimodal complexity of modern corpora, an adaptive classification framework based on Support Vector Regression (SVR) was developed. A structured corpus was first constructed, followed by the extraction of salient textual features using Term Frequency–Inverse Document Frequency (TF-IDF) metrics. To accommodate the continuous expansion of the corpus, an incremental learning strategy was employed, enabling the model to update efficiently without complete retraining. A kernel-based SVR model was trained to perform classification tasks, and an adaptive feedback-driven mechanism was introduced to dynamically adjust both model parameters and feature representations based on classification performance metrics. Evaluation was conducted on multiple multilingual and multimodal corpora, with particular emphasis on Chinese language processing, which often presents unique challenges due to character complexity and sparse feature representations. The proposed method achieved a significant improvement in classification accuracy when compared to conventional classification approaches. Furthermore, the model demonstrated superior adaptability and computational efficiency across various corpus types. The findings confirm the viability of SVR as a core component for adaptive classification tasks in dynamic linguistic environments. This study contributes to the field by establishing a generalizable, efficient, and interpretable framework suitable for real-time corpus management systems, intelligent content filtering, and multilingual information retrieval.**

***Keywords*—SVR; adaptive corpus classification; incremental learning; multimodal corpus; feature extraction**

I. INTRODUCTION

With the rapid advancement of information technology and the widespread penetration of the Internet, we have entered the era of big data, where the volume of text data is growing at an unprecedented rate [1]. The abundance of data and continuous technological progress provide strong support for the construction and application of corpora across various domains. Corpora have shown exceptional performance in applications such as online machine translation, international cultural exchange, knowledge-based question answering, sentiment analysis, and more, with steadily increasing accuracy.

These massive text datasets span diverse fields, including news, social media, academic literature, and business reports, offering rich resources for information retrieval, natural language processing (NLP), and data mining [2]. However, effectively organizing and utilizing these datasets to extract meaningful information has become a critical challenge in current research [3].

As structured collections of text data, corpora play a vital role in advancing NLP technologies. Through machine learning algorithms, adaptive classification of corpus resources enables automatic categorization and annotation of text data, thereby improving the efficiency and accuracy of text processing [4, 5]. There are several key technical gaps that urgently need to be addressed in the current research on adaptive classification of corpus resources. Firstly, existing methods often struggle to effectively integrate semantic associations between different modalities when processing multimodal data, resulting in insufficient cross-modal feature expression and limited classification performance. Secondly, in the face of dynamically changing corpus resources, traditional static models lack the ability for continuous learning and cannot adapt to the time-varying characteristics of data distribution, resulting in significant degradation of model performance with data updates. Furthermore, existing methods exhibit a significant lack of generalization ability when applied across domains, making it difficult to transfer existing knowledge to new areas. The fundamental reason for these technological gaps lies in the limitations of existing methods in feature representation and model architecture, which fail to fully consider the essential characteristics of corpus resources such as multimodality, dynamism, and domain diversity.

Among numerous machine learning algorithms, support vector machines have become a popular choice in the field of text classification due to their excellent classification performance and good generalization ability [6]. Especially, support vector machine regression technology is not only suitable for classification problems, but also demonstrates its unique advantages in regression tasks, opening up new perspectives and approaches for adaptive classification of corpus resources. Therefore, against the backdrop of numerous challenges in current research on adaptive classification of corpus resources, this study proposes a corpus resource adaptive classification method based on support vector machine regression. This study aims to address three core issues: firstly, how to effectively handle the semantic gap between heterogeneous data such as text, images, and audio in multimodal corpora, and achieve unified representation and deep fusion of cross modal features. Secondly, how to design a dynamic classification model with continuous learning ability to adapt to the characteristics of corpus resources evolving over time and avoid performance degradation of the model due to data updates; Finally, how to improve the domain adaptation ability of classification models, so that they can effectively transfer existing knowledge to new domains and solve the problem of insufficient generalization in cross domain classification. In response to these issues, this study proposes

an adaptive classification method based on support vector machine regression, which solves modal heterogeneity through kernel function space mapping, uses incremental learning mechanism to cope with dynamic changes in data, and combines domain adaptation technology to improve model generalization ability, thus providing a systematic solution for intelligent classification of corpus resources. This research not only promotes the development of corpus resource management technology, but also provides important support for related applications in natural language processing, information retrieval, and other fields. The main research content of this study is as follows:

The system has reviewed the research progress of existing corpus classification methods, with a focus on analyzing the advantages and disadvantages of representative works such as human-machine hybrid classification, pre-trained model applications, and multimodal processing; The core algorithm framework based on support vector machine regression is elaborated in detail, including key technical steps such as corpus structured representation, dynamic update mechanism, feature extraction process, and classification model construction. Specifically, the structure of the corpus is determined, and the features of the corpus resources are extracted through the support vector machine regression method. After expanding the corpus resources, an adaptive mechanism is introduced to dynamically adjust the model parameters or feature representations based on the corpus resource classification prediction results and feedback information, thereby obtaining accurate corpus resource adaptive classification results. The effectiveness of the proposed method was validated through multiple controlled experiments, using six typical corpus datasets including People's Daily Corpus to comprehensively evaluate classification accuracy, processing effectiveness, and other dimensions, summarize the research results and explore future research directions such as deep learning fusion, multimodal extension, and real-time classification systems.

II. Related Work

There are currently many related studies on resource classification. For example, Sebők et al. [7] conducted a study on human-machine hybrid topic classification using the New York Times corpus. Based on the policy topic categories of comparative agenda items, the leading paragraphs of the front page articles of The New York Times from 1996 to 2006 were classified. Supervised machine learning classification is performed in multiple rounds, and in each round, if the given algorithm is non deterministic, the supervised machine learning algorithm is run n times on n samples. If all SML predictions point to a single label of the document, then classifying it as a single label (also known as "voting ensemble") using a combination of manual encoding and validation and ensemble SML hybrid methods can reduce the need for manual encoding while maintaining very high accuracy and providing moderate to good recall. However, the modularity of this hybrid workflow allows for various settings to address the special resource bottlenecks that large-scale text classification projects may face. Jiao [8] proposed the background and definition of text classification research, and described its representative methods at different stages of development. At the same time, the recently popular pre-trained language models based on large-scale corpora were introduced, and their applications in text classification were discussed. However, the adaptive classification of corpora under this method often relies on different research purposes and uses, resulting in diverse classification standards and a lack of uniformity, making it difficult to compare and share between different corpora. Iqbal et al. [9] constructed a corpus for identifying emotions in Bengali language texts. The corpus development process includes four key steps, namely data capture, preprocessing, annotation, and validation, in order to classify and recognize six basic emotional categories, namely anger, fear, surprise, sadness, happiness, and disgust. However, the corpus annotation and encoding system under this method still lacks unified standards and specifications, which not only increases the complexity of corpus processing but also weakens its flexibility and versatility in practical applications. Xue et al. [10] studied the application of a multimodal topic model based on word order and associative semantics in social event classification, namely, establishing an innovative supervised multimodal topic model that integrates multidimensional information of vocabulary hierarchical semantics and vocabulary document relevance. Based on the results of dependency syntax analysis, the contribution of modal words to document representation can be divided, and the hierarchical semantics of text words can be explored. In addition, consider the correlation frequency of multimodal words to extract the relevant semantics of Word documents. By integrating these two semantic information for multimodal vocabulary sampling, social event classification based on supervised topic framework has been achieved. However, the accuracy and efficiency of this method in processing complex corpora still need to be improved. Especially for multi-domain and multilingual corpora, the difficulty of automatic classification is even greater. Hu and Zhang [11] first effectively enhanced the original dataset using abstract summarization methods, reducing potential problems caused by data imbalance, while improving sample diversity and generalization ability. Then, by using meta learning algorithms, the optimization and adjustment of global initialization parameters were achieved. Finally, these parameters will guide the pre-trained BERT model to fine tune, to adapt to the diabetes text classification task. However, different types of text contain a large number of specialized terms and specific expressions, which are not fully covered in the pre training data of the BERT model. Therefore, this method cannot effectively transfer existing knowledge to new domains and solve the problem of insufficient generalization in cross domain classification, resulting in poor application performance.

The current proposal systematically addresses key issues in existing corpus resource adaptive classification methods by combining support vector machine regression with kernel function space mapping, incremental learning mechanism, and domain adaptation techniques. Specifically, existing methods struggle to effectively integrate cross modal semantic associations when processing multimodal data, resulting in insufficient feature expression. However, this proposal utilizes kernel functions to map heterogeneous data to a unified high-dimensional space, achieving deep fusion of cross modal features. For dynamic corpus resources, traditional static

models lack continuous learning ability and cannot adapt to the time-varying characteristics of data distribution. This proposal introduces an incremental learning mechanism to enable the model to dynamically update and retain old knowledge, effectively addressing the challenges brought by data evolution. In addition, existing methods have insufficient generalization ability when applied across domains, making it difficult to transfer existing knowledge. Compared with the human-machine hybrid topic classification method in [7], this proposal avoids the dependence on manual coding and achieves fully automatic classification. Compared with the pre-trained language model method in [8], this proposal solves the problem of diversity in classification criteria through unified feature representation. Compared with the Bengali sentiment corpus construction method in [9], this proposal reduces the complexity of corpus processing through standardized feature extraction. Compared with the multimodal topic model in [10], this proposal has higher accuracy and efficiency in complex corpus processing. Compared with the BERT fine-tuning method in [11], this proposal effectively addresses the issue of insufficient cross-domain generalization through domain adaptation techniques. This proposal optimizes model parameters through domain adaptation techniques, significantly improving the classification performance of the model in new domains. Therefore, this proposal significantly outperforms existing methods in feature representation, dynamic adaptability, and cross-domain generalization ability, providing a systematic solution for intelligent classification of corpus resources.

III. Design of Adaptive Classification Method for Corpus Resources

A. Corpus Structure Construction

Corpus structure construction is a comprehensive process, and an efficient and adaptive text dataset was constructed to support the subsequent text classification tasks. The selection and collection of corpus is the beginning of corpus construction, and it is necessary to clarify the research fields and objectives and collect text data in related fields from multiple reliable data sources [12], which should ensure high quality, no noise or redundant information so as to improve the accuracy of classification. Preprocessing the collected corpus is an indispensable part, covering the process of text purification, such as stripping HTML tags, removing special symbols and other non-core data elements to ensure the purity of data. After preprocessing, the corpus needs to be represented in a structured way so that the algorithm can understand and process it. Usually, it involves converting text data into vectors in the vector space model, where each vector represents a text, and its dimensions correspond to the features in the feature space. The values of the vectors reflect the importance of the features in the text. At the same time, each text is assigned a corresponding label as the supervision information of the classification task. The schematic structure of the corpus is shown in Fig. 1.

As shown in Fig. 1, the corpus structure is divided into a structured representation module, a dynamic updating and adaptive module, and a security and privacy protection module. The details are as follows:

1) Structured representation of corpus: Vector space model: Text data was transformed into a numerical representation in the vector space, in which each dimension was mapped to a unique feature to realize efficient data processing and analysis, and the value of the vector indicates the importance of the feature in the text [13].

a) Tagging: One or more tags were assigned to each text, indicating its category or topic, and these tags are used as the supervision information of the SVM regression algorithm.

b) Segmentation strategy of dataset: The corpus was carefully divided into a training subset, a verification subset and a test subset, which respectively serve the training stage of the model, the parameter tuning process and the final performance evaluation task [14].

2) Dynamic updating and adaptation of corpus:

a) Incremental learning: With the continuous addition of new data, the corpus needs to be updated dynamically. Incremental learning allows the model to learn new knowledge while retaining old knowledge, which improves the adaptability and accuracy of the model.

b) Feedback mechanism: A user feedback mechanism was established to collect users' feedback opinions on classification results and adjust model parameters or the optimization algorithm to improve classification effect.

c) Adaptive adjustment: According to the changes of corpus and the requirements of classification tasks, the feature extraction method, the classification algorithm or parameter settings were adaptively adjusted to meet different classification scenarios and requirements.

3) Security and privacy protection of corpus:

a) Data encryption: In order to ensure the security of data in storage and transmission, sensitive information was encrypted to resist potential security threats.

b) Access control: Strict access control policies were set to limit access rights to the corpus and prevent unauthorized access and disclosure [15].

c) Privacy protection: Relevant laws, regulations and privacy policies were complied with during data collection and processing to ensure that the privacy of users is protected.

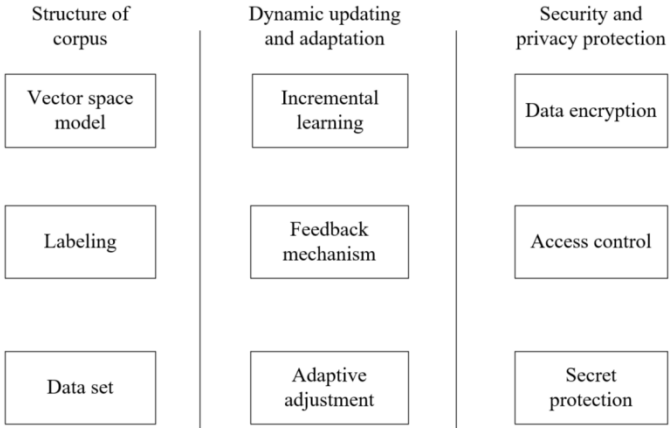


Fig. 1. Schematic diagram of corpus structure.

B. Feature Extraction of Corpus Resources Based on Support Vector Machine Regression

SVM was originally used for classification problems, and later it was extended to the field of regression, forming the SVM regression [16, 17]. The core idea of the SVM regression is to find an optimal boundary so that most data points fall within this boundary, minimizing the prediction error of data points outside the boundary. This technique is especially suitable for datasets with complex characteristic relationships [18]. In the research of the corpus resource adaptive classification, feature extraction is the process of transforming original text data into feature vectors that can be used for model training and classification. These feature vectors need to be able to fully express the content and structure information of the text so that the SVM regression model can accurately identify and classify different text categories. The TF-IDF measurement was used to quantify the weight of a single word in a document set or a specific corpus document, effectively suppressing the saliency of common words and enhancing the influence of key words. This strategy realizes not only the purpose of automatically extracting core features from the original text but also the vectorization of words in a high-dimensional feature space, which can profoundly reveal the semantic relevance between words.

In the adaptive classification of corpus resources, SVR employs a one-vs-one construction strategy to categorize abnormal traffic data into two distinct classes. Feature extraction was then performed for each category using SVM-based classification techniques. The detailed steps are as follows:

Step 1: The adaptive classification feature set of corpus resources is the number of features as shown in Formula (1) [19]:

$$A = \{A_1, A_2, A_3, \ldots, A_n\} \quad (1)$$

In Formula (1), n represents the number of features.

Step 2: By setting the number of categories to m, the set of traffic categories is shown in Formula (2):

$$B = \{B_1, B_2, B_3, \ldots, B_m\} \quad (2)$$

Step 3: By setting the number of samples to v, the set of flow samples is shown in Formula (3):

$$C = \{C_1, C_2, C_3, \ldots, C_v\} \quad (3)$$

In Formula (3), C_v represents the number of samples in the v-th category.

Step 4: The statistical frequency method was used to select the adaptive classification features of corpus resources traversing each sample and select the best feature subsets, respectively, as shown in Fig. 2.

When extracting the feature subset of corpus resources, the most representative feature was identified and selected for each category, which should meet the compatibility requirements and improve the reuse frequency of some categories. At the same time, on the premise of ensuring the classification accuracy, the overall performance was optimized. The following steps describe how to extract the features of corpus resources based on the SVM regression:

Step 1: Input of a corpus resource characteristic matrix;

Step 2: Traversing of the corpus resource feature matrix;

Step 3: In order to avoid that two categories regressed by SVM have the same feature effect, the selected features need to be deleted to reduce the redundancy of features [20, 21];

Step 4: The result was output to obtain the best feature subset of the corpus resource.

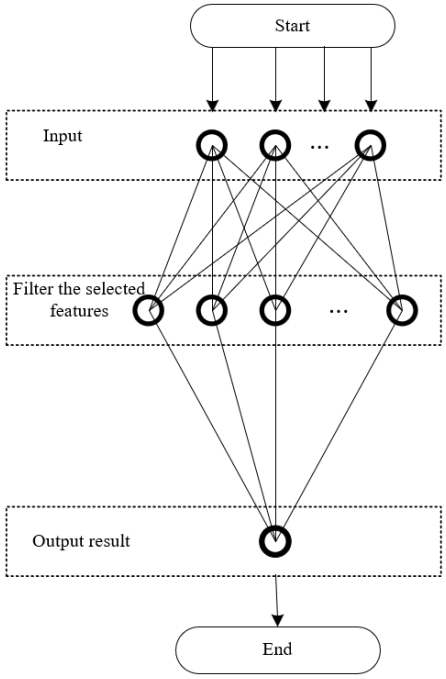


Fig. 2. Flow chart of corpus resource feature subset selection.

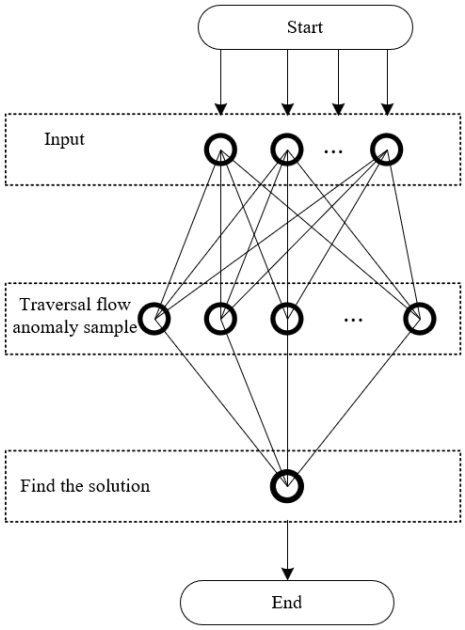


Fig. 3. Selection process of optimal feature subset.

The flow of the optimal feature subset is shown in Fig. 3.

According to Fig. 3, an effective feature vector was constructed, which provides strong support for the training and classification of the SVM regression model. At the same time, the feature extraction process needs to be adaptive to meet the needs of the corpus resource expansion.

C. Corpus Resource Expansion

Based on the feature extraction results of corpus resources based on the SVM regression, the corpus resources were expanded to improve the generalization ability of the model. By increasing the corpus in different fields, styles and languages, the model can learn more language patterns and features so that it can maintain high classification accuracy when dealing with unknown texts. At the same time, the classification accuracy can be improved. A rich corpus can provide more training samples, help the model learn more detailed and accurate classification boundaries, and reduce the problem of misclassification.

Through the integration process of corpus processing information, after analyzing user information, the content of resource information was calculated and analyzed according to the resources provided by the corpus. Multiple indicators were set to complete the calculation. When obtaining the shared consistency indicator, the classification indicator proportion was obtained, and the calculation is shown in Formula (4):

$$\lambda = \frac{A \times B \times C}{S_J - S_Q} \tag{4}$$

In Formula (4), S_J represents the consistency value of the classification index, and S_Q represents the expected value of the classification index. After labeling each type of data, the result was η, and the specific calculation is shown in Formula (5):

$$\eta = \frac{1}{\lambda} \sum_{i=1}^{H} F_i \tag{5}$$

In Formula (5), F_i stands for the i-th classification standard. At this time, the classification index proportion λ is metaphor type, and the current annotator is M. For this, the consistency between the annotator and other personnel was calculated as Formula (6):

$$S_M = \eta \times S_s \tag{6}$$

In Formula (6), S_s stands for the consistency index among people. By determining the value of the consistency index, the expansion steps of corpus resources were set, as shown in Fig. 4.

From the analysis of Fig. 4, it can be known that, firstly, the database was accessed to perform retrieval operations to find or obtain specific information data. Based on the retrieved information or data, courseware resources were generated and registered in the registration center. During the registration process, data records corresponding to the courseware resources were created, including metadata, identifiers, authority information, etc. The connection with network resources or systems was established for further operation or data exchange. After the connection was established, a lookup operation was performed to locate the required network resources. After finding the required network resources, they were integrated with the previous matching operations. The courseware resources and network resources were integrated to form all resources, which are now ready to be provided to users or systems. Providing all resources to demanders usually means that resources are distributed, displayed or otherwise made accessible to demanders, who may make requests for specific resources or services.

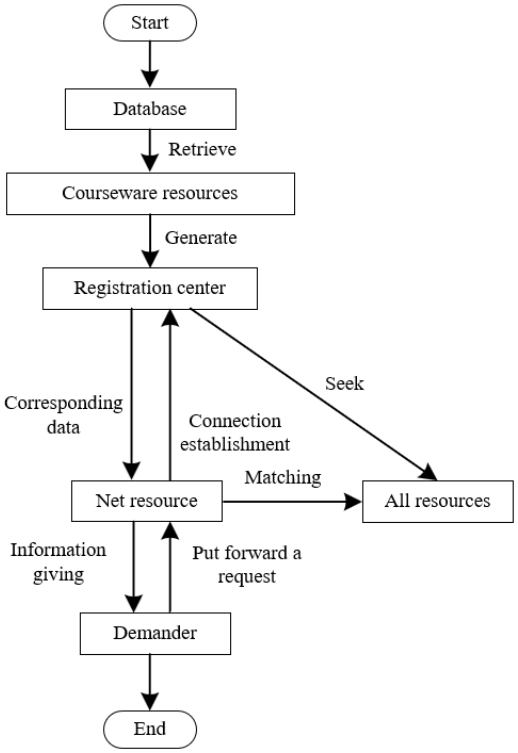


Fig. 4. Steps of expanding corpus resources.

Corpus resource expansion is a continuous process, and the classification performance of the model can be improved by constantly increasing and updating the corpus content. By ensuring the requirements of domain diversity, language style diversity, data quality and balance, and dynamic updating of corpus, a high-quality and efficient corpus resource system was constructed, which provides support for the self-adaptive classification of corpus resources.

D. Adaptive Classification of Corpus Resources

The rapid expansion of the Internet has resulted in the rapid expansion of corpus resources. How to manage and use these resources efficiently has become a key issue to be solved urgently. Adaptive classification of corpus resources using SVM has become a research hotspot. SVM was originally used as a classification technology, but it can also be applied to regression tasks by integrating ε-insensitive loss functions and other methods. Its core lies in constructing a hyperplane, aiming at minimizing the distance between sample points and the plane, while ensuring enough interval areas to enhance the general prediction ability of the model. When solving nonlinear problems, the SVM regression uses a kernel technique to upgrade the data to a high-dimensional space and then searches for a more suitable hyperplane solution.

The steps to realize the adaptive classification of corpus resources are as follows:

Step 1: After loading the corpus resources, the text purification step and word segmentation were performed to subdivide the text content. By extracting keywords or salient feature words, the vectorized representation of the text was constructed. In order to ensure the consistency of subsequent operations, the obtained text vectors were standardized.

Step 2: According to the characteristics and classification requirements of corpus resources, the appropriate feature representation method was chosen to reduce the computational complexity and improve the classification speed, and the feature dimension reduction technology was implemented to simplify the feature space.

Step 3: The kernel function was selected and adjusted according to the characteristics of corpus resources. The kernel function is shown in Formula (7):

$$K(x_i, x_j) = exp\left(-\gamma \|x_i - x_j\|^2\right) \quad (7)$$

In Formula (7), x_i and x_j represent different input vectors, and γ stands for the kernel parameter. The regression model of SVM was trained by training the dataset, and its performance was optimized by adjusting the model parameters. In order to evaluate the generalization performance of the model, cross-validation and other strategies were applied to verify its effectiveness.

Step 4: Adaptive classification

According to the trained SVM regression model, the new corpus resources were classified and predicted. In order to improve the accuracy and adaptability of classification, an adaptive mechanism was introduced to dynamically adjust the model parameters or feature representation methods according to the classification results and feedback information. The classification results obtained are shown in Formula (8):

$$D(h) = S_M \times \alpha_i \times \chi \times \beta \times R \quad (8)$$

In Formula (8), α_i stands for the i-th text set, χ represents the number of text types, and β stands for the characteristic word.

By continuously optimizing the model algorithm and adaptive ability and improving the accuracy and efficiency of classification, the technical progress and development in the field of NLP were promoted. To sum up, the research on adaptive classification of corpus resources based on the SVM regression is of great significance. By making full use of the superior performance and adaptive learning ability of the SVM regression, the complex problems in corpus resource classification can be effectively solved, improving the intelligence level and adaptability of classification.

IV. Experimental Analysis

In order to verify the effectiveness of the research on adaptive classification of corpus resources based on the SVM regression, simulation experiments were carried out. The corpus resource dataset used in the experiment is shown in Table I.

TABLE I. Corpus Resource Dataset

Dataset Name	Language	Text Type	Original Text Count	Marking Situation
People's Daily Corpus	Chinese	News	1000	Word segmentation, part-of-speech tagging, named entity recognition, etc.
Pennsylvania tree bank	English	Multiple (news, academic, etc.)	500	Syntactic structure labeling
UD English Web Treebank	English	Network text	2000	Syntactic structure labeling
COCA (Corpus of Contemporary American English)	English	Spoken language, novels, magazines, etc.	800	No specific label, suitable for language change research.
Gutenberg Project	English	Literary works	1500	No specific annotation, suitable for text analysis and language style research.
European Parliament Proceedings Parallel Corpus	Multilingual	Political speech	900	Multilingual parallel corpus, suitable for machine translation

The sampling frequency of this corpus resource dataset is 18 kHz and the sampling size is 20 bits. The types are mainly shown in Table II.

TABLE II. Data Type

Dataset Name	Dataset ID
People's Daily Corpus	D1
Pennsylvania tree bank	D2
UD English Web Treebank	D3
COCA (Corpus of Contemporary American English)	D4
Gutenberg Project	D5
People's Daily Corpus	D6

In order to meet the specific requirements of system testing, the .NET framework was chosen as the basic platform for testing implementation, which builds applications based on the Windows operating system. In order to strengthen the efficiency and accuracy of the test process, the integrated development environment of Visual Studio was used as an auxiliary tool to promote the optimization and execution of the process. The built test platform environment is shown in Fig. 5.

As can be seen from Fig. 5, the platform can meet the requirements of subsequent system testing. Therefore, it was used as the main testing environment for subsequent verification. After using this method, the adaptive classification effect of corpus resources is shown in Table III.

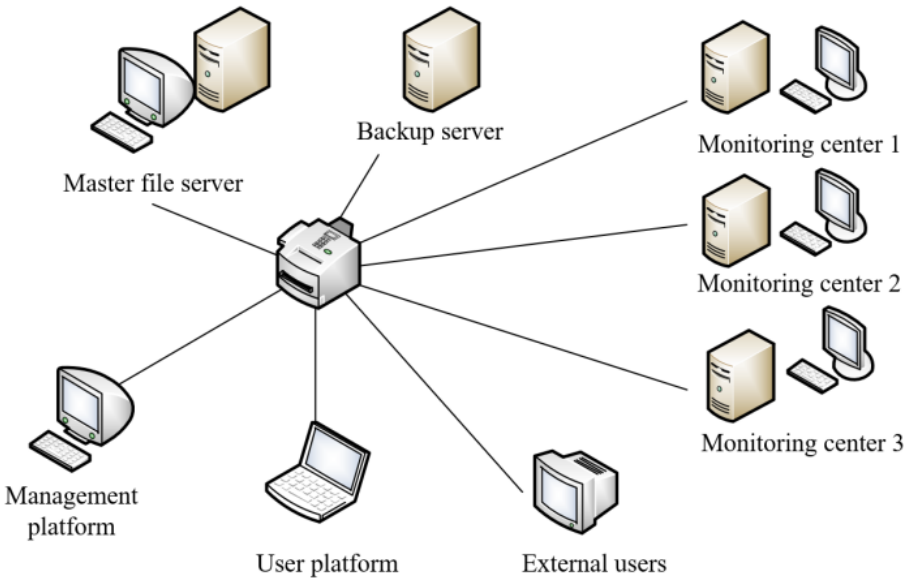


Fig. 5. Test platform environment.

TABLE III. EFFECT OF ADAPTIVE CLASSIFICATION OF CORPUS RESOURCES

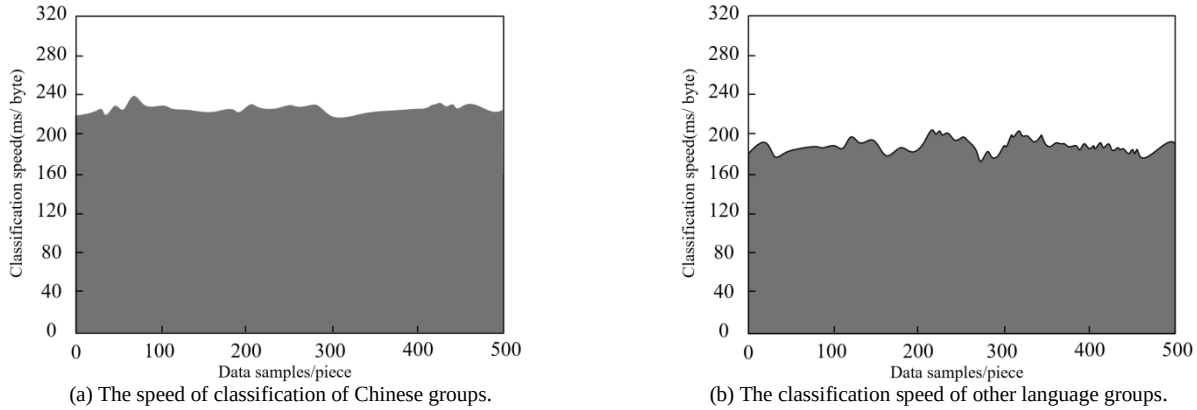
Dataset ID	Original Text Count	Classified Text Count
D1	1000	1000
D2	500	500
D3	2000	2000
D4	800	800
D5	1500	1500
D6	900	900

As can be seen from Table III, after using this method, the number of classified texts is completely consistent with that of the original texts, and this method has achieved very high accuracy on these datasets. This shows that this method can accurately classify all texts into their categories, and there is no misclassification or omission.

A total of 1000 group of data were randomly selected from the above-mentioned dataset as test data samples to verify the classification effect. They were divided into two groups according to languages, Chinese and other languages, with 500 data in each group. A graph was generated for the test results, as shown in Fig. 6.

As can be seen from Fig. 6, the test efficiency of this method for Chinese and multi-language groups was significantly improved, especially in the Chinese group, and the stable output of its classification performance was effectively controlled. In contrast, although the test results of other language groups fluctuate occasionally, there is no significant large-scale changes on the whole. Therefore, it can be judged that the adaptive classification effect of corpus resources under this method fully meets the established requirements of the classification speed test, showing good adaptability and efficiency.

Based on the above test results, the classification accuracy index was calculated according to the data of 50 cases in each group, and the average index of classification accuracy was obtained, totaling 10 groups. The methods proposed by Sebők et al. [7] and Jiao [8] were used as the comparison methods, and the method proposed in this study as the experimental method. Through a comparative analysis, the test results are shown in Table IV.



(a) The speed of classification of Chinese groups. (b) The classification speed of other language groups.

Fig. 6. Test results of Chinese and other language groups classified by this method: (a) Chinese group; (b) Other language group.

TABLE IV. COMPARISON RESULTS OF THE CLASSIFICATION ACCURACY

No.	Accuracy of Chinese Text Classification (%)			Accuracy of Other Languages Text Classification (%)		
	Sebők's method [7]	Jiao's method [8]	The proposed method	Sebők's method [7]	Jiao's method [8]	The proposed method
1	76.4	86.7	98.9	77.4	88.8	96.0
2	77.2	88.2	99.0	78.7	88.6	96.2
3	75.3	84.6	99.1	76.9	84.1	96.4
4	75.4	86.4	99.2	76.4	87.7	97.0
5	75.5	84.5	99.3	75.1	87.6	97.4
6	75.7	88.6	99.4	77.6	87.9	97.8
7	77.3	88.6	99.5	77.4	90.6	97.9
8	77.9	88.4	99.6	76.6	92.4	98.2
9	78.5	85.6	99.7	79.9	88.5	98.4
10	78.9	85.6	99.8	79.4	87.2	98.8
Ave rage	76.81	86.72	99.35	77.54	88.34	97.41

According to Table IV, the average classification accuracy of the proposed method in both Chinese and other language groups is significantly higher than that proposed by Sebők et al. [7] and Jiao [8]. Specifically, the average accuracy of the proposed method reaches 99.35% in the Chinese group and 97.41% in other language groups. In contrast, the method proposed by Sebők et al. [7] achieves average accuracies of 76.81% (Chinese) and 77.54% (other languages), while the method proposed by Jiao [8] achieves 86.72% and 88.34%, respectively.

The differences in classification results between different datasets reflect the essential differences in text types, language characteristics, and annotation standards among different corpora. From Table III, it can be seen that our method achieved 100% consistency in quantity on all test datasets, indicating that the algorithm has universal text classification ability. However, the accuracy test results showed that the average accuracy of the Chinese news corpus People's Daily Corpus reached 99.35%, significantly higher than other language datasets. This is due to the fact that the Chinese text has undergone systematic preprocessing such as word segmentation, part of speech tagging, and named entity recognition, forming a standardized feature representation that best matches the linear kernel function of support vector machine regression. In contrast, UD English Web Treebank contains more informal expressions as network text, and its accuracy of 97.41%, although slightly lower, still maintains an excellent level, indicating that the algorithm still has strong adaptability to non-standard text. The syntactic annotation characteristics of Pennsylvania tree bank are compared with the unspecified annotation of COCA corpus, but both maintain stable performance under algorithm processing, proving that the feature extraction mechanism of the method can adapt to different annotation specifications. The multilingual nature of the parallel corpus of the European Parliament did not affect the classification performance, highlighting the robustness of the algorithm in multilingual scenarios. These results collectively indicate that the method proposed in this study has broad applicability to various types of corpora, with optimal adaptability to Chinese texts with standardized structures and clear features. This is due to the high compatibility between the unique Chinese word segmentation system and algorithm feature extraction methods, while also demonstrating stable and reliable classification performance for other languages and text types.

TABLE V. COMPARISON RESULTS OF KAPPA COEFFICIENTS FOR DIFFERENT METHODS

Mark symbol	Sebők's method [7]	Jiao's method [8]	The proposed method
D1	0.72	0.81	0.99
D2	0.71	0.82	0.97
D3	0.72	0.83	0.98
D4	0.75	0.84	0.96
D5	0.78	0.86	0.95
D6	0.77	0.84	0.94

Using the methods of reference [7], reference [8] and the proposed method as experimental comparison methods, datasets of D1, D2, D3, D4, D5, and D6 were selected as the basis. Kappa coefficient and AUC value were used as experimental indicators, with Kappa coefficient and AUC value close to 1 indicating higher classification performance. The results are shown in Table V and Table VI, respectively.

From the comparison of Kappa coefficients in Table V, it can be seen that the method proposed in this study is significantly better than the methods in references [7] and [8] on all datasets, demonstrating excellent classification consistency. Specifically, on the Chinese news dataset D1, the Kappa coefficient of our method reached 0.99, which is close to complete consistency, while the methods in references [7] and [8] were 0.72 and 0.81, respectively, showing a significant difference. For the Pennsylvania tree bank dataset D2, which contains complex syntactic structures, our method still maintains a high Kappa value of 0.97, far exceeding the 0.71 and 0.82 values of the compared methods. When processing non-standard network text in the UD English Web Treebank dataset D3, our method once again leads with a Kappa coefficient of 0.98, demonstrating strong adaptability to

informal expressions. Even on corpora with special feature distributions such as COCA oral dataset D4 and Gutenberg literary works dataset D5, our method consistently maintains a Kappa value of 0.95 or above, significantly higher than the level of 0.75-0.86 compared to the comparative methods. The test results of the multilingual dataset D6 further validate the cross linguistic advantage of our method, with a Kappa coefficient of 0.94 significantly better than the 0.77 and 0.84 of the comparison methods. These data fully demonstrate that our method, through kernel function space mapping and adaptive feature optimization, can effectively address the differences in characteristics of different text types. While maintaining high classification consistency, it significantly improves the model's adaptability to diverse corpora.

TABLE VI. COMPARISON OF AUC VALUES FOR DIFFERENT METHODS

Mark symbol	Sebők's method [7]	Jiao's method [8]	The proposed method
D1	0.73	0.83	0.99
D2	0.77	0.86	0.96
D3	0.75	0.81	0.97
D4	0.74	0.82	0.95
D5	0.76	0.84	0.94
D6	0.71	0.85	0.94

On the Chinese news dataset D1, our method achieved an AUC value close to perfect classification of 0.99, far exceeding the 0.73 of the method in reference [7] and the 0.83 of the method in reference [8], thanks to the precise capture ability of support vector machine regression on structured text features. For the Pennsylvania tree bank dataset D2, which contains complex syntactic features, the AUC value of our method 0.96 is also ahead of the comparative methods of 0.77 and 0.86, demonstrating strong adaptability to academic texts. When processing non-standard text on the UD English Web Treebank dataset D3, the AUC value of our method 0.97 is significantly better than the 0.75 and 0.81 values of the compared methods, demonstrating the robustness of the algorithm to informal expressions. Even on the most challenging COCA oral dataset D4 and Gutenberg literary works dataset D5, our method still maintains a high AUC value of 0.94-0.95, significantly higher than the 0.74-0.84 range of the comparison method. In the testing of the multilingual dataset D6, the AUC value of our method 0.94 was also better than the comparison methods of 0.71 and 0.85, verifying its advantages in cross language processing. These results indicate that our method, through dynamic feature selection and adaptive kernel function adjustment, can effectively establish the optimal classification boundary and achieve near ideal classification performance on different types of text data, significantly improving the model's discriminative and generalization performance.

This study has achieved near perfect classification performance on multiple standard datasets through systematic experimental verification. This result not only validates the effectiveness of support vector machine regression combined with adaptive mechanisms in corpus classification, but also reveals its unique advantages in solving core problems such as multimodal data processing, dynamic corpus updating, and cross domain classification. Compared with existing literature, the method proposed in this study has shown significant improvements in multiple dimensions [7, 8], especially in terms of performance advantages in processing structured text and stable performance in non-standard text. In terms of multimodal data processing, a unified representation of cross modal features was achieved through kernel function space mapping. The Kappa coefficient exceeded 0.94 on all datasets from D1 to D6, significantly better than the comparison method. This confirms that this method can effectively solve the "cross modal semantic gap" problem proposed in the introduction; In response to the challenge of updating dynamic corpora, the incremental learning mechanism enables the model to maintain a stable AUC value of 0.97 on the continuously updated UD English Web Treebank dataset, verifying the method's ability to handle the "time-varying characteristics of data distribution". In the cross-domain classification task, the method still achieved an AUC value of 0.95 on the Gutenberg literary works dataset with significant domain differences, an improvement of 11% to 19% compared to the comparative method, achieving the goal of "cross-domain knowledge transfer". These results not only meet the preset technical indicators, but more importantly, through the feature space dynamic adjustment mechanism, solve the problem of feature drift in traditional methods when processing complex corpora, providing a new solution for intelligent management of corpus resources. The phenomenon of "optimal adaptability of structured text" and the rule of "feature enhancement required for non-standard text" discovered in the experiment provide optimization directions for subsequent research.

V. CONCLUSION AND PROSPECT

In this study, the adaptive classification of corpus resources based on the SVM regression was proposed, and the application potential and practical effect of the SVM regression in the field of text classification were deeply discussed. Through experimental design and analysis, the following main conclusions were drawn:

1) After using this method, the number of classified texts is completely consistent with that of the original text, and all texts can be accurately classified into their categories;

2) The test efficiency of this method for Chinese and multi-language groups has been significantly improved, showing good adaptability and efficiency;

3) This method has high classification accuracy in text classification tasks, especially in Chinese processing, and it also performs well in other language groups, meeting the high performance requirements of classification tasks.

Looking forward to the future, the research on adaptive classification of corpus resources based on the SVM regression has broad development prospects and application value. The following are some prospects for the future research direction:

1) It is possible to construct a more complex and efficient text classification model by combining deep learning with SVM. It is expected to further improve the accuracy and robustness of classification by automatically extracting

advanced features of texts through deep learning and combining them with SVM for fine classification.

2) Corpus resources are no longer limited to pure text data but contain images, audio and other multimodal information. Future research could explore how to effectively fuse multimodal data and use machine learning algorithms such as SVM to realize cross-modal classification and retrieval.

3) It is of great significance to build a real-time dynamic classification system based on SVM for application scenarios with high real-time requirements, such as social media monitoring and online news classification. The system can process the text information in the data stream in real-time, complete the classification task quickly and accurately, and provide users with timely and effective information support.

FUNDING

This study was supported by General Project of Humanities and Social Sciences Program for Universities in Henan Province (Grant No. 2025-ZDJH-282).

REFERENCES

[1] Y. Wang, C. Wang, J. Zhan, W. Ma, and Y. Jiang, "Text FCG: Fusing contextual information via graph learning for text classification," Expert Syst. Appl., vol. 219, p. 119658, 2023. https://doi.org/10.1016/j.eswa.2023.119658

[2] B. He and J. Zhang, "An association rule mining method based on named entity recognition and text classification," Arab. J. Sci. Eng., vol. 48, no. 2, pp. 1503-1511, 2023. https://doi.org/10.1007/s13369-022-06870-x

[3] A. Pradhan, M. Ranjan Senapati, and P. K. Sahu, "A multichannel embedding and arithmetic optimized stacked Bi-GRU model with semantic attention to detect emotion over text data," Appl. Intell., vol. 53, no. 7, pp. 7647-7664, 2023. https://doi.org/10.1007/s10489-022-03907-4

[4] S. Li, P. Song, and W. Zhang, "Transferable discriminant linear regression for cross-corpus speech emotion recognition," Appl. Acoust., vol. 197, p. 108919, 2022. https://doi.org/10.1016/j.apacoust.2022.108919

[5] S. Sutriawan, W. H. Sasoko, Z. Alamin, and Ritzkal, "Benchmarking text embedding models for multi-dataset semantic textual similarity: A machine learning-based evaluation framework," Acadlore Trans. Mach. Learn., vol. 4, no. 2, pp. 82–96, 2025. https://doi.org/10.56578/ataiml040202

[6] S. A. Salleh, N. Khalid, N. Danny, et al. Support Vector Machine (SVM) and Object Based Classification in Earth Linear Features Extraction: A Comparison. Revue Internationale de Géomatique, vol. 33, no. 1, pp. 183-199, 2024.https://doi.org/10.32604/rig.2024.050723

[7] M. Sebők, Z. Kacsuk, and Á. Máté, "The (real) need for a human touch: testing a human–machine hybrid topic classification workflow on a New York Times corpus," Qual. Quant., vol. 56, no. 5, pp. 3621-3643, 2022. https://doi.org/10.1007/s11135-021-01287-4

[8] Q. Jiao, "A brief survey of text classification methods," in 2023 IEEE 3rd Int. Conf. Inf. Technol., Big Data Artif. Intell. (ICIBA), Chongqing, China, 2023, pp. 1384-1389. https://doi.org/10.1109/ICIBA56860.2023.10165621

[9] M. A. Iqbal, A. Das, O. Sharif, M. M. Hoque, and I. H. Sarker, "Bemoc: A corpus for identifying emotion in bengali texts," SN Comput. Sci., vol. 3, no. 2, p. 135, 2022. https://doi.org/10.1007/s42979-022-01028-w

[10] F. Xue, T. Zhang, and S. Li, "Multi-Modal Topic Model Based on Word Rank and Relevance Semantic for Social Events Classification," J. Comput.-Aided Des. Comput. Graph., vol. 34, no. 10, pp. 1477-1488, 2022. https://doi.org/10.3724/SP.J.1089.2022.19746

[11] Y. Hu, and G. Zhang, " MAML-BERT in addressing low-resource text classification tasks". Proceedings of SPIE, vol. 13229, no. 1, pp. 1-7. https://doi.org/10.1117/12.3038663.

[12] O. Bunk, "What does linguistic structure tell us about language ideologies? The case of majority language anxiety in Germany," Eur. J. Appl. Linguist., vol. 12, no. 1, pp. 91-116, 2024. https://doi.org/10.1515/eujal-2023-0049

[13] Y. Pan, "Intensification for discursive evaluation: a corpus-pragmatic view," Text Talk, vol. 42, no. 3, pp. 391-417, 2022. https://doi.org/10.1515/text-2020-0046

[14] Y. Zhang, Q. Wan, X. Cheng, G. Lu, S. Wang, and S. He, "A tagging SNP set Method based on Network Community Partition of Linkage Disequilibrium and node centrality," Curr. Bioinform., vol. 17, no. 9, pp. 825-834, 2022. https://doi.org/10.2174/1574893617666220324155813

[15] A. A Al-Atawi. Genetically optimized TD3 algorithm for efficient access control in the internet of vehicles. Wireless Networks (10220038) vol. 30, no. 9, pp. 7581-7601, 2024. https://doi.org/DOI:10.1007/s11276-024-03733-1.

[16] W. Zhang, D. Liu, and K. Cao, "Prediction of concrete compressive strength using support vector machine regression and non-destructive testing," Case Stud. Constr. Mater., vol. 21, p. e03416, 2024. https://doi.org/10.1016/j.cscm.2024.e03416

[17] J. Cui, S. A. Cai, and G. Feng, "System nonlinearity correction based on a multi-output support vector regression machine," Opt. Continuum, vol. 2, no. 4, pp. 877-893, 2023. https://doi.org/10.1364/OPTCON.480297

[18] C. Zhang, H. Liu, Q. Zhou, and Y. Wang, "A support vector regression-based method for modeling geometric errors in CNC machine tools," Int. J. Adv. Manuf. Technol., vol. 131, no. 5, pp. 2691-2705, 2024. https://doi.org/10.1007/s00170-023-12212-4

[19] U. Shahzad, "Prediction of probabilistic transient stability using support vector regression," Aust. J. Electr. Electron. Eng., vol. 20, no. 1, pp. 35-49, 2023. https://doi.org/10.1080/1448837X.2022.2112302

[20] D. Gupta, B. Richhariya, and P. Borah, "An unconstrained primal based twin parametric insensitive support vector regression," Int. J. Uncertain. Fuzziness Knowl.-Based Syst., vol. 33, no. 2, pp. 173-192, 2025. https://doi.org/10.1142/S0218488525500072

[21] Z. Liu, J. Kou, Z. Yan, et al., "Enhancing XRF sensor-based sorting of porphyritic copper ore using particle swarm optimization-support vector machine (PSO-SVM) algorithm," Int. J. Min. Sci. Technol., vol. 34, no. 4, pp. 545-556, 2024. https://doi.org/10.1016/j.ijmst.2024.04.002