Research Article

Text Feature Extraction for Public English Vocabulary Based on Wavelet Transform

Di Ye and Xiaojing Shi

School of Foreign Studies, Tangshan Normal University, Tangshan, Hebei 063000, China

Correspondence should be addressed to Di Ye; yedi@tstc.edu.cn

Received 26 April 2022; Revised 21 May 2022; Accepted 24 May 2022; Published 11 June 2022

Academic Editor: Naeem Jan

Copyright © 2022 Di Ye and Xiaojing Shi. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Text interpretation of public English vocabulary is a critical task in the subject of natural language processing, which uses technology to allow humans and computers to communicate effectively using natural language. Text feature extraction is one of the most fundamental and crucial elements in allowing computers to effectively grasp and read text. This paper proposes a text feature extraction method based on wavelet analysis that performs fast discrete wavelet transform and inverse discrete wavelet transform on the feature vectors under the traditional TF-IDF vector space model to address the problem of low feature differentiation of high-dimensional data in text feature extraction. In particular, due to the design of the Mallat algorithm, there is frequency aliasing in the signal decomposition process. This phenomenon is a problem that cannot be ignored when using wavelet analysis for feature extraction. Therefore, this paper proposes an improved inverse discrete wavelet transform method, in which the signal is decomposed by Mallat algorithm to obtain wavelet coefficients at each scale and then reconstructed to the required wavelet space coefficients according to the reconstruction method, and the reconstructed coefficients are used to analyze the signal at that scale instead of the wavelet coefficients obtained at the corresponding scale. Experiments on the public English vocabulary dataset reveal that the wavelet transform-based strategy suggested in this research outperforms existing feature extraction methods while maintaining greater classification accuracy while reducing the dimensionality of the TF-IDF vector space model.

1. Introduction

With the development of the Internet and the continuous updating of computers and information technology, the information stored on the network is becoming more and more abundant. The number of texts, as an effective expression of information, is also growing rapidly. In recent years, with the rise of cloud computing and big data, it has enabled the effective organization and management of huge amount of public English vocabulary texts. How to obtain effective information efficiently and accurately has become the main purpose of text mining, information retrieval, and network opinion analysis. The diversity, complexity, redundancy, and irregularity of public English lexical text data pose a great challenge for text understanding. The core of text understanding is to convert text data into signals that can be perceived and analyzed by computers through mathematical operations and to process them automatically to feedback the results depending on the task. In text understanding, one of the most fundamental and critical steps is text feature extraction. Therefore, feature reduction of complex feature space with high dimensionality of text becomes the main key point for text classification. The purpose of feature extraction is to effectively downscale the initial high-dimensional features and select an optimal subset of features from the high-dimensional feature space [1].

The vast amount of public English lexical texts on the Internet has brought about a rich corpus of resources but at the same time has made text perception, analysis, and processing a huge challenge. The first challenge is that any user can generate and disseminate data, the majority of which is text, resulting in the rapid growth of the text corpus; the second challenge is that behind the big data lies a large amount of repetitive and meaningless data, which is of mixed quality
and low value density. Finally, data exists in a variety of platforms, including structured data, semistructured data, and unstructured data, so “high feature dimensionality and complex structure” are the third challenge. The third challenge is “high feature dimensionality and complex structure.” Addressing these challenges is the main obstacle for text data analysis [2]. The feature vector space is usually formed by using a set of words of a text as attribute vectors. The original feature vector space of text contains all the attributes of words, which is high-dimensional and sparse, but not all attributes contribute to the classification decision. Therefore, it is necessary to effectively reduce the dimensionality of the high-dimensional text feature space and extract the best set of classification feature attributes without degrading the system performance.

In recent years, many scholars have proposed a large number of effective methods and techniques in text feature extraction to address these three challenges. Words in public English, as the most basic units in a text, are the smallest elements that constitute sentences and discourse. Feature extraction of words is usually called word-level representation, but the number of words in a text, no matter in English, is very large, and sequential coding of these words alone is not only labor-intensive but also difficult to reveal the semantic relationships between words, so a vectorized representation of word level with measurable semantic distance is very necessary. Specifically, given a semantic metric, each word or phrase is projected as a high-dimensional vector, and the space formed by these vectors is called the vector space at the word level, thus transforming the unstructured text into a manageable structured form. Public English lexical texts still have the common problems in text data analysis: high dimensionality and feature redundancy. Data compression has been one of the important application areas of wavelet analysis, and this has led to great social and economic benefits [3]. Compressed perception theory proves that a signal can be sampled at a lower frequency and reconstructed with high probability as long as the signal is sparse in some orthogonal space. Compressed perception theory can efficiently capture information from sparse signals and perceive measurements through noncorrelation, a property that makes compressed perception widely used in real-life applications. Compressed perception theory has brought a revolutionary breakthrough by solving the current bottleneck in information acquisition and processing technology and has received widespread attention from scholars in various countries, ranging from medical imaging and signal coding to astronomy and geophysics [4]. The Boolean model [5–7] is an early text representation paradigm that uses a collection of “1” and “0” variables to represent the feature items of the related text. However, regardless of semantic relevance, text features are always allocated equal weights by default, making the represented features unable to match the realistic meaning.

In this paper, to address the problems of high dimensionality and feature redundancy of public English lexical text, wavelet transform and inverse wavelet transform are performed on the feature vectors under the text vector space model, so that the text feature space dimensionality is reduced. In particular, for the phenomenon of frequency aliasing in the signal decomposition process of Mallat algorithm, an improved algorithm model that can effectively eliminate frequency aliasing is proposed, expecting to achieve the purpose of accurate extraction of public English vocabulary text features, and the proposed method is verified based on the text classification task.

The paper’s divisional layout is as follows: the related work is presented in Section 2. Section 3 analyzes the algorithm design of the proposed work. Section 4, discusses the experimentation and results. Finally, in Section 5, the research work is concluded.

2. Related Works

2.1. Text Feature Extraction. After text preprocessing and representation, text information still belongs to a high-dimensional and highly sparse vector matrix, which makes the computer’s computation and learning training process more difficult, and the classification effect is poor. Text feature selection is required to achieve even more dimensionality reduction. Finding the ideal subset of features in the solution space comprising all feature subsets and selecting the most representative combination of features with the least amount of time are the key to feature selection. Since computer devices can only recognize mathematical symbols such as binary, the original text language should be converted into a computer-recognizable representation before using computers to study text information. Among the many text representation methods, the most effective and accurate method is to build a text representation model. A good text representation model not only affects the accuracy of text detection but also is related to the relevance of the semantic connection between text data. Thus, text representation models have an irreplaceable position in text analysis. Boolean model [5] is an early text representation model, which represents the feature items of the corresponding text by the set of “1” and “0” variables. However, at this time, the text features are assigned with equal weights by default regardless of semantic relevance, which makes the represented features unable to match the realistic meaning. The vector space model (VSM) was proposed by Huang et al. [6] and is one of the classical models in text representation. This model is also a statistical model based on bag of words [7], which compares the words in a text to a bag of balls and represents the text ignoring the order of word occurrence, which has the feature of simplicity and speed. However, the individual unordered individuals alone are also unable to solve the relationship between words, which in turn causes the problem of losing semantic information. The above two models generally suffer from two main problems of high dimensionality and vector sparsity in the process of representing text, which eventually leads to a large amount of computer resources being spent but instead results in poor text clustering accuracy. In recent years, research scholars have focused on how to reduce the complexity of the text representation model on the basis of ensuring that the clustering accuracy is not affected, and the main solutions can be broadly divided into two categories: feature
selection and feature extraction. Among them, feature selection is used to obtain a subset of features from the initial text feature data as the text representation according to certain rules. For example, the term frequency and inverse document frequency (TF-IDF) for text feature selection [8, 9] can be used to solve the problem of equal weights. It gives feature words weight coefficients based on the number of occurrences and relative relevance of a word in the text, with the goal of filtering out feature words that highlight the text topic. Scholars have suggested feature selection techniques such as the chi-square test [10], information gain [11], and mutual information [2] since then, all of which have achieved the goal of decreasing the complexity of the text representation model without affecting the semantics of the text. In contrast, in the feature extraction approach, the extracted features are not derived from a subset of the initial text features, but a new set of text features created based on the initial text features is used as the text representation. The current common feature extraction methods tend to utilize the core idea of stitching features, where individuals with little difference in nature in the features are fused into new feature items. The Dirichlet allocation topic model [12] and the random mapping model [13] have made significant contributions to the research development of feature extraction models. With the rapid development of deep learning in recent years, Mikolov et al. [14] proposed the Word2Vec model, which has made great progress in feature extraction techniques. The model represents text by converting words in text space to vector space and using low-dimensional numerical vectors. It overcomes the text vector dimensionality explosion and increases the accuracy of conveying the semantics of the original text by representing words as word vectors by exploiting semantic linkages between contexts. After that, Kim et al. [15] proposed the Doc2vec model for text semantic feature extraction. The above mathematical models based on deep learning reflect the current situation that deep learning knowledge is gradually integrated into text representation as well as feature extraction; however, the problem of large-scale data and computational hardware loss required by deep learning models is also more significant.

2.2. Wavelet Transform. The idea of wavelet transform comes from the method of stretching and translation, which is to compress and stretch the signal in the time axis; translation is to move the wavelet basis function in parallel with the waveform in the time auxiliary. In 1974, the French engineer Morlet first proposed wavelet transform, which was not available at that time. In 1974, the mathematician Meyer proved the existence of wavelet function and carried out an in-depth study in the theory. So far, wavelet transform theory has been widely used in signal analysis [16], image processing [17], fault diagnosis [18], and other engineering fields.

Since 1990, the wavelet transform and its engineering applications have gradually received the attention of scientists from all over the world, and it is considered a major breakthrough to the Fourier transform. Wavelet transform is developed on the basis of Fourier transform, and they are essentially different from each other in time and frequency domains. Wavelet transform provides an adaptive analysis method of simultaneous local changes in time and frequency domains, which can automatically adjust the time and frequency windows to meet the needs of practical engineering, no matter analyzing low-frequency or high-frequency local signals. Characterizing the singularity of the signal is another feature of wavelet transform, and the maximum value of wavelet transform modes of the fault signal at different decomposition scales can represent the sudden change of the signal. The application of wavelet transform theory to the field of signal processing has been developed rapidly in recent years, mainly including signal denoising, data compression, and fault diagnosis.

The wavelet is defined as follows: let \( \psi(t) \) be a square productable function, \( \psi(t) \in L^2(\mathbb{R}) \). Its Fourier transform \( \psi(\omega) \) is given by

\[
C_\psi = \int_{\mathbb{R}} \left| \frac{\psi(\omega)}{\omega} \right|^2 d\omega < \infty, \tag{1}
\]

calling \( \psi(t) \) as the base wavelet or mother wavelet; equation (1) is the tolerance condition of the wavelet function. Scaling and translating \( \psi(t) \), the function \( \psi_{a,t}(t) \) is obtained as

\[
\psi_{a,t}(t) = \frac{1}{\sqrt{|a|}} \psi\left(\frac{t - \tau}{a}\right), \quad a, \tau \in \mathbb{R}; a \neq 0. \tag{2}
\]

In formula (2), \( a \) and \( \tau \) are the scaling factor and translation factor, respectively. The wavelet transform is defined as follows: after translating the wavelet function \( \psi(t) \) by \( \tau \), the inner product operation is done with the signal to be analyzed \( x(t) \) at different scales \( a \), as shown in the following equation:

\[
WT_x(a, \tau) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t) \psi^\ast\left(\frac{t - \tau}{a}\right) dt, \quad a > 0. \tag{3}
\]

The equivalent frequency domain representation of equation (3) is shown in equation (4), where \( X(\omega) \) and \( \psi(\omega) \) are the Fourier transforms of \( x(t) \) and \( \psi(t) \), respectively.

\[
WT_x(a, \tau) = \frac{\sqrt{a}}{2\pi} \int_{-\infty}^{+\infty} X(\omega)\psi^\ast(\omega)e^{i\omega t} d\omega. \tag{4}
\]

The wavelet transform can obtain a multiresolution description of the signal, and this description conforms to the general laws of human observation of the world. At the same time, wavelet transform has rich wavelet basis that can be adapted to signals with different characteristics. The features of wavelet and wavelet transform are (1) joint local analysis function in time domain and frequency domain; (2) multiresolution and multiscale analysis function; (3) a good local approximation basis for nonlinear system; and (4) fast algorithm based on conjugate mirror filter bank. At present, wavelet transform is widely used in industry, medical treatment, military, and other fields.
3. Algorithm Design

The vector space model simplifies text processing, and combinations between different words may affect the effect of their alignment. However, its disadvantage is that the vector dimensions increase rapidly as the text set expands and the dictionary words increase. Because it is difficult for a single text vector to include all of the words in the dictionary, there are many dimensions with weight values of 0, resulting in high dimensionality and vector sparsity. Therefore, we need to reduce the dimensionality of the vectors in the traditional vector space model, which can be viewed as digital signals for text vectors. And wavelet analysis theory has strong advantages for digital signal processing. Existing theories and practices show that the transformed digital signal can be highly restored to the original signal, and wavelet analysis can uniquely capture the localization details, which makes it possible to operate in the wavelet transform space for this paper. In this paper, the DWT and IDWT methods used in this paper are performed on the vector space model to test the effectiveness of this text feature extraction method with the accuracy of text classification. The flow chart of the algorithm is shown in Figure 1. Firstly, the dataset is divided into words and a dictionary is constructed to obtain the TF-IDF feature space vectors; then, the obtained TF-IDF vectors are subjected to a one-dimensional discrete wavelet transform (DWT) to obtain the scale coefficients and wavelet coefficients; the scale coefficients and wavelet coefficients are summed up in the corresponding components to obtain the wavelet space vectors proposed in this paper. Finally, the scale coefficients and wavelet coefficients obtained in the previous step are summed up again by reducing the corresponding scale functions, and several of their middle dimensions are extracted from the obtained vectors to obtain the inverse discrete wavelet transform (IDWT) space vectors proposed in this paper.

3.1. TF-IDF Feature Extraction. Term frequency-inverse document frequency (TF-IDF) is one of the popular algorithms in the field of text mining, which is obtained by multiplying TF (word frequency) and IDF (inverse document rate) [19]. The actual meaning of TF is the probability of occurrence of a keyword in a document, as shown in equation (5). The more the keyword appears in a single text, the larger the TF value is, where \( n_{i,j} \) denotes the number of occurrences of word \( t_i \) in text \( d_j \) and \( \sum_k n_{k,j} \) denotes the number of occurrences of all words in document \( d_j \).

\[
TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}. \tag{5}
\]

The emergence of TF has led many scholars to use TF as a method for judging text similarity. Although this method can reflect the similarity of high-frequency words in each text, it can provide theoretical support for judging text similarity to a certain extent. However, there are many popular or basic words in all kinds of texts, so it is obvious that the TF frequency alone is not scientifically supported. Therefore, the IDF variable is introduced as the weight factor of TF, which is used to regulate the problem that the TF value is generally high among the texts of multiple categories. The IDF is calculated by taking the number of texts containing a certain keyword as the denominator and the number of texts in the corpus as the numerator, and the result of this fractional formula is taken as the logarithm to obtain the result of IDF, as shown in the following equation:

\[
IDF = \log \frac{|D|}{1 + |\{ j : t_i \in d_j \}|}. \tag{6}
\]

In equation (6), \( D \) is the whole corpus and \( \{ j : t_i \in d_j \} \) is the number of texts in which a word appears in the corpus. In order to avoid the situation that the denominator is 0 because the word does not exist in the corpus, the denominator is increased by 1. If a high-frequency word appears in a large number of texts in the corpus, it means that the word is less important for a single text and is not a keyword to be extracted. In this case, the IDF value of this word will be smaller, and the weight of this keyword will be reduced. For example, if all the texts in a corpus are about a person, the name of the person may appear in all the texts, and then, the IDF value will be very small. In summary, the formula of TF-IDF is shown in equation (7), which eliminates a large number of common words in the text while retaining high-frequency words and extracts words with a high degree of importance.

\[
TF_{i,j} - IDF_i = TF_{i,j} \times IDF_{i,j}. \tag{7}
\]

3.2. Improved Mallat Algorithm

3.2.1. Traditional Mallat Algorithm. The Mallat algorithm is a fast algorithm for wavelet analysis that introduces the idea of multiresolution analysis in the field of digital video image wavelet analysis. However, the Mallat algorithm is implemented by using wavelet low-pass filters \( H \) and \( h \) and...
wavelet high-pass filters $G$ and $g$ corresponding to the scale function and wavelet function to filter the signal in low pass and high pass. For convenience, the scale function is referred to as the low-frequency subband and the wavelet function is referred to as the sub-high-frequency subband. The Mallat decomposition algorithm is as follows:

$$
A_0[f(t)] = f(t),
$$
$$
A_j[f(t)] = \sum_k H(2t - k)A_{j-1}[f(t)], \tag{8}
$$
$$
D_j[f(t)] = \sum_k G(2t - k)A_{j-1}[f(t)],
$$

where $t$ is the discrete time series number, $t = 1, 2, \cdots, N$; $f(t)$ is the original signal; $j$ is the number of layers, $j = 1, 2, \cdots, J$, $J$ takes $[\log_2 N]$; $H$ and $G$ are wavelet decomposition filters; $A_j$ is the wavelet coefficient of the approximate part of the $j$th layer of the signal $f(t)$; and $D_j$ is the wavelet coefficient of the detailed part of the signal $f(t)$ in layer $j$. The decomposition process of Mallat two-dimensional tower wavelet transform is shown in Figure 2.

Refactoring algorithm

$$
A_j[f(t)] = 2\left\{ \sum_k h(t - 2k)A_{j+1}[f(t)] + \sum_k g(t - 2k)D_{j+1}[f(t)] \right\}, \tag{9}
$$

where $j, J$ is the same as equation (8), $j = J - 1, J - 2, \cdots, 0$; $h, g$ are wavelet reconstruction; and $A_j$ and $D_j$ have the same meaning as equation (2). The reconstruction process of Mallat two-dimensional tower wavelet transform is shown in Figure 3.

3.2.2. Mallat Algorithm Analysis. From Mallat’s decomposition algorithm, it follows that the approximate part of the signal $f(t)$ is at the 2\textsuperscript{th} scale (jth layer). The wavelet coefficients $A_j$ of the low-frequency part are obtained by convolving the wavelet coefficients $A_{j-1}$ of the approximate part of the 2\textsuperscript{j-1} scale (the $j$ - 1th layer) with the decomposition filter $H$, and then, the convolution result is sampled at intervals. The wavelet coefficients $D_j$ of the high-frequency part of the signal $f(t)$ at the 2\textsuperscript{j}th scale (the $j$th layer) are obtained by convolving the wavelet coefficients $A_{j-1}$ of the
approximate part of the $2^{j-1}$ scale (the $j-1$ th layer) with the decomposition filter $G$ and then sampling the convolution result at intervals. It can be said that the Mallat algorithm is basically done in 3 steps:

1. Wavelet filtering (convolution)
2. Interval sampling
3. Interval zero interpolation

Figure 5: Simulation effect of the improved Mallat algorithm.
Therefore, the frequency overlap phenomenon is definitely generated during the above 3 steps. The interpolation of zeroes is a technique used in wavelet reconstruction, which does not affect the wavelet decomposition stage; i.e., it is not relevant to the problem discussed in this paper. In fact, the wavelet filter is not an ideal filter, and the nonideal frequency domain characteristics of the filter make it possible for the signal to be filtered with each bandlimited sub-band containing its neighboring subband components, thus generating the frequency aliasing phenomenon. However, considering the unevenness of frequency aliasing in each band-limited subband (approximate and detailed), there is almost no aliasing in the approximate band-limited sub-band. In fact, interval sampling is the root cause of frequency aliasing, because it violates Shannon’s sampling theorem. Therefore, if the original signal contains frequency components close to the sampling edge, the decomposition according to Mallat’s algorithm will definitely generate frequency aliasing imagination.

3.2.3. Improved Mallat Algorithm. After a thorough study of Mallat algorithm, it is clear that the implementation of Mallat algorithm inevitably generates frequency aliasing. We can see from the wavelet decomposition and reconstruction algorithms that the wavelet decomposition process is sampling at intervals and the reconstruction process is interpolating zeros at intervals; both processes cause frequency aliasing, but in different directions. That is to say, the decomposition process generated by the overlap in the reconstruction process has been corrected. However, this also provides an idea of how to solve the frequency aliasing phenomenon generated by Mallat decomposition algorithm: after decomposing the signal by Mallat algorithm to obtain wavelet coefficients at each scale and then reconstructing to the required wavelet space coefficients \( c_d \) according to the reconstruction method and using the reconstructed coefficients \( d_j \) to analyze the signal at that scale instead of the wavelet coefficients \( d_j \) obtained at the corresponding scale. In this way, the effects of frequency aliasing can be better resolved to achieve the desired goal. In this paper, this algorithm is called subband signal reconstruction algorithm. The improved algorithm model is shown in Figure 4. Using this model, the TF-IDF features of the equation input are decomposed and the simulation results show that the algorithm is practical and effective. Figure 5 shows the analysis of the TF-IDF feature vector using this algorithm and the spectral effect of the detailed signal \( d_j \).

4. Experiments

4.1. Experiment Preparation. The text dataset of public English vocabulary used in this paper is a corpus retrieved from the web. It contains a total of 10,505 samples from eight categories: education, culture, finance and economics, science and technology, sports, military, agriculture, and politics. This data source is rich and suitable for the study of text classification. The number of samples in each category is shown in Table 1. Two major groups of experiments are conducted in this paper. The first set of experiments tests the classification performance of each training space when targeting internal samples (the test set and training set are from the same distribution), and 80% of the data in the experiments are used for the training set and the rest for testing; the second set of experiments verifies the superiority of the wavelet analysis method proposed in this paper for exotic samples. The text data with the categories of education, culture, finance and economics, science and technology, and sports are used as the training set, and the text data with the categories of military, agriculture, and politics are used as the test set (the training set and the test set are from different distributions). Singular value decomposition (SVD) [20], independent component analysis (ICA) [21], principal component analysis (PCA) [22] method, and the modified Mallat method in this paper were used to extract the feature vectors under the vector space (TF-IDF). The feature vectors (TF-IDF) of the vector space are extracted by dimensionality reduction. The accuracy of each space is measured using the KNN method (cosine distance as the similarity measure), and the dimensionality of SVD space, ICA space, and PCA space is consistent. In this paper, the precision, recall, and F1-score are used to evaluate the model performance. The accuracy rate indicates the percentage of correctly predicted results to the total samples; the precision rate indicates the number of correctly predicted samples among the samples predicted to be of that class; the recall rate is the probability of being predicted to be of that class among the samples that are actually of that class; both the precision rate and recall rate are considered simultaneously, so that both are simultaneously maximized and a balance is achieved.

### Table 1: Text dataset of public English vocabulary.

| Text category | Sample size |
|---------------|-------------|
| Education     | 1021        |
| Culture       | 1568        |
| Economic      | 1392        |
| Science       | 1077        |
| Sports        | 1408        |
| Military      | 1390        |
| Agricultural  | 1296        |
| Political     | 1353        |
| Total number  | 10505       |

### Table 2: Average experimental results of internal category text classification.

| Methodology | Precision | Recall | F1-score |
|-------------|-----------|--------|----------|
| SVD         | 0.875     | 0.859  | 0.867    |
| ICA         | 0.826     | 0.847  | 0.836    |
| PCA         | 0.883     | 0.862  | 0.872    |
| Proposed    | 0.904     | 0.911  | 0.907    |
4.2. Internal Category Text Classification. In this section, we conduct classification experiments on internal categories of public English lexical texts to verify the feature extraction effect. Since a total of 8 categories of training samples are used in this experiment, the average of precision rate, recall rate, and F1-score is taken as indicators to evaluate the effectiveness of each method. The classification results are shown in Table 2, and the F1-score of the method in this paper is improved by 04%, 7.10%, and 3.50% compared with that of SVD, ICA, and PCA, respectively. The results show that the wavelet transform-based feature extraction methods proposed in this paper can all obtain robust feature representations, which confirms the stability of the method.

In order to show more intuitively and clearly the effect of the improved feature extraction algorithm on the index improvement of classifier classification, the F1-score of each algorithm under different categories is plotted in this paper based on the data in Table 2 as shown in Figure 6. It can be observed that the ICA and PCA methods have similar feature extraction ability for public English lexical texts, while the wavelet transform-based method proposed in this paper has the best feature extraction ability for each category. This is because the wavelet transform inherits and develops the localization idea of the short-time Fourier transform, overcomes the shortcomings such as the window size does not change with frequency, and can fully highlight the features of certain aspects of the problem through the transform. It can localize the analysis of time (space) frequency and gradually carry out multiscale refinement of the signal (function) through the telescopic translation.
operation, finally achieving time subdivision at high frequencies and frequency subdivision at low frequencies, which can automatically adapt to the requirements of time-frequency signal analysis and thus can focus on any details of the signal.

4.3. External Category Text Classification. The second set of experiments is aimed at verifying the feature extraction ability of the algorithm for samples with different distributions from the training set. Table 3 shows the detailed classification accuracy of the three categories in the test set. From the table, it can be concluded that the proposed method based on the improved inverse discrete wavelet transform can have better feature extraction ability for unseen categories of text, while the SVD, ICA, and PCA algorithms depend on the characteristics of the training set.

5. Conclusions

In this paper, an improved inverse discrete wavelet transformation text feature extraction method based on the Mallat algorithm is proposed for the problem that public English lexical text features have high dimensionality and the features contain redundant information. In this work, the wavelet transform is simply the sum of the respective components of the low-frequency and high-frequency vectors created by the one-dimensional discrete wavelet transform. After the one-dimensional discrete wavelet transform, the dimensionality of the low-frequency and high-frequency vectors is the same and around half of the original vectors, meeting the dimensionality reduction goal. We discover the primary reasons of frequency mixing in the Mallat algorithm in this work, based on an in-depth examination of the algorithm, and suggest a better model to solve the frequency mixing. Under certain circumstances, the proposed inverse wavelet space may have a higher accuracy for a specific classification category. Many low-dimensional feature extraction approaches lose crucial classification features; however, the highly sparse orthogonal wavelet space vector in this study, according to compressed perception theory, can reliably maintain the important properties of the most original feature vector. The next work is to check whether the feature extraction efficiency of wavelet analysis method has certain advantages in experiments and how to expand the specific conditions of this paper’s inverse wavelet space to make its specific advantages greater.

Data Availability

The datasets used during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare that they have no conflict of interest.

References

[1] K. Thirumoorthy and K. Muneeswaran, “Feature selection using hybrid poor and rich optimization algorithm for text classification,” Pattern Recognition Letters, vol. 147, pp. 63–70, 2021.
[2] X. Tang, Y. Dai, and Y. Xiang, “Feature selection based on feature interactions with application to text categorization,” Expert Systems with Applications, vol. 120, pp. 207–216, 2019.
[3] M. Andrecut, “Wavelet lossy compression of random data,” International Journal of Modern Physics C, vol. 20, no. 1, pp. 109–116, 2009.
[4] L. Li, Y. Fang, L. Liu, H. Peng, J. Kurths, and Y. Yang, “Overview of compressed sensing: sensing model, reconstruction algorithm, and its applications,” Applied Sciences, vol. 10, no. 17, p. 5909, 2020.
[5] B. Menauwer, S. Mohammed, and M. Nada, “Towards a model to improve Boolean knowledge mapping by using text mining and its applications,” International Journal of Information Retrieval Research (IJIRR), vol. 10, no. 3, pp. 35–56, 2020.
[6] L. B. Huang, V. Balakrishnan, and R. G. Raj, “Improving the relevancy of document search using the multi-term adjacency keyword-order model,” Malaysian Journal of Computer Science, vol. 25, no. 1, pp. 1–10, 2012.
[7] D. Yan, K. Li, S. Gu, and L. Yang, “Network-based bag-of-words model for text classification,” IEEE Access, vol. 8, pp. 82641–82652, 2020.
[8] N. Azam and J. T. Yao, “Comparison of term frequency and document frequency based feature selection metrics in text categorization,” Expert Systems with Applications, vol. 39, no. 5, pp. 4760–4768, 2012.
[9] T. Peng, L. Liu, and W. Zuo, “PU text classification enhanced by term frequency-inverse document frequency-improved weighting,” Concurrency and Computation: Practice and Experience, vol. 26, no. 3, pp. 728–741, 2014.
[10] Y. Lu, M. Liang, Z. Ye, and L. Cao, “Improved particle swarm optimization algorithm and its application in text feature selection,” Applied Soft Computing, vol. 35, pp. 629–636, 2015.
[11] C. Shang, M. Li, S. Feng, Q. Jiang, and J. Fan, “Feature selection via maximizing global information gain for text classification,” Knowledge-Based Systems, vol. 54, pp. 298–309, 2013.
[12] N. Liu, X. J. Tang, Y. Lu, M. X. Li, H. W. Wang, and P. Xiao, “Topic-sensitive multi-document summarization algorithm,” in 2014 sixth international symposium on parallel architectures, algorithms and programming, pp. 69–74, Beijing China, 2014.
[13] C. B. Do and A. Y. Ng, “Transfer learning for text classification,” Advances in Neural Information Processing Systems, vol. 18, 2005.
[14] S. Ji, N. Satish, S. Li, and P. K. Dubey, “Parallelizing word2vec in shared and distributed memory,” IEEE Transactions on Parallel and Distributed Systems, vol. 30, no. 9, pp. 2090–2100, 2019.
[15] D. Kim, D. Seo, S. Cho, and P. Kang, “Multi-co-training for document classification using various document representations: TF-IDF, LDA, and Doc2Vec,” Information Sciences, vol. 477, pp. 15–29, 2019.
[16] D. A. Slesarev and V. A. Barat, “Application of wavelet transform to the analysis of signals with pulsed components,” Measurement Techniques, vol. 43, no. 8, pp. 696–699, 2000.
[17] G. B. Gebremeskel, “A critical analysis of the multi-focus image fusion using discrete wavelet transform and computer vision,” Soft Computing, vol. 26, pp. 5209–5225, 2022.

[18] M. H. Wang, S. D. Lu, and R. M. Liao, “Fault diagnosis for power cables based on convolutional neural network with chaotic system and discrete wavelet transform,” IEEE Transactions on Power Delivery, vol. 37, no. 1, pp. 582–590, 2022.

[19] J. Yi, G. Yang, and J. Wan, “Category discrimination based feature selection algorithm in Chinese text classification,” Journal of Information Science & Engineering, vol. 32, no. 5, 2016.

[20] D. G. Tadesse and M. Carpenter, “A method for selecting the relevant dimensions for high-dimensional classification in singular vector spaces,” Advances in Data Analysis and Classification, vol. 13, no. 2, pp. 405–426, 2019.

[21] L. C. Yu and W. N. Chien, “Independent component analysis for near-synonym choice,” Decision Support Systems, vol. 55, no. 1, pp. 146–155, 2013.

[22] Y. Zhang, G. Li, and H. Zong, “A method of dimensionality reduction by selection of components in principal component analysis for text classification,” Univerzitet u Nišu, vol. 32, no. 5, pp. 1499–1506, 2018.