International Journal of Information Security and Privacy
Volume 19 • Issue 1 • January-December 2025

Legal Privacy Protection Machine Learning Method Based on Word2Vec Algorithm

Rongrong Wang
https://orcid.org/0009-0009-8000-379X
Zhe Jiang J.R.C. Law Firm, China

ABSTRACT

This study uses Word2Vec's word vector representation technology to finely capture the semantic relationships of vocabulary in legal texts through the Skip-gram model. By introducing Hierarchical Softmax optimization, a legal privacy protection model based on Word2Vec algorithm is ultimately designed. The results showed that the model performed better than other comparative algorithms in both the macro classification performance (Fl_macro) and the micro classification performance (Fl_micro). In practical legal sensitive word recognition tasks, the accuracy, recall rate, and F1 score of the model reached 92.56%, 88.78%, and 90.62%, respectively. Therefore, the proposed model effectively improved the accuracy of identifying sensitive legal privacy words and providing new methods for the personal information security protection system.

KEYWORDS

Word2Vec, LSTM-B, Privacy Protection, Machine Learning, Text Recognition

INTRODUCTION

In consequence of the accelerated pace of technological advancement in the sphere of internet technology, the number of internet users in China has surpassed one billion, with the penetration rate of the internet exhibiting a marked increase (Amin et al., 2023). However, in this context, information security issues have become increasingly prominent, with frequent leaks of personal legal information. In particular, the irregular collection and processing of personal information by internet enterprises has become the focus of social attention (Nsugbe, 2023). As an important resource for internet enterprises, the collection and analysis of user information provides enterprises with the possibility of personalized services and precision marketing but also exposes users to potential threats of privacy disclosure (Hebbi & Mamatha, 2023; Wang et al., 2021a). The formulation and implementation of privacy policies are often controlled by enterprises, which can easily lead to the abuse of users' personal information. Although relevant laws and regulations such as the Personal Information Protection Law have been implemented, in practical operation, it is difficult for infringed users to obtain effective compensation, and the role of legal protection is limited (Wang et al., 2021b). With the increasing maturity of machine learning in recent years, applying it to legal privacy protection (LPP) has gradually become an effective choice (Chen et al., 2022). Given this, the study uses word2vec's word vector representation technology to finely capture the semantic relationships of vocabulary in legal texts through the skip-gram model. The introduction of hierarchical softmax (HS) optimizes the efficiency of traditional softmax, and combines word vector distance and TextRank algorithm to identify sensitive content in text. Finally, the long short-term memory network classifier (LSTM-B)

DOI: 10.4018/IJISP.365911

is introduced, and an LPP machine learning model based on the word2vec algorithm (LPPM-W2V) is designed. This study aims to explore the LPPM-W2V method and improve the personal information security protection system.

LITERATURE REVIEW

The progress of the internet has made more people begin to pay attention to the LPP problem on the internet. Wang proposed a new word2vec-based unsupervised anomaly detection method (LogUAD) to address the challenges faced by large-scale distributed system log analysis. Compared with log cluster, LogUAD's F1 score has increased by 67.25% (Wang et al., 2022). Ren et al. (2022) proposed a new k-mer context free alignment method (kmer2vec) to address the problem of high computational complexity and difficulty in effectively capturing sequence context structure in traditional multiple sequence alignment methods when dealing with large numbers of sequences. The kmer2vec performed well in constructing phylogenetic trees and species clustering, with a much faster running speed than the multiple sequence alignment method. Gao et al. (2022) proposed a method that combined the dynamic topic model and the word2vec model to address the issue of quantifying semantic distribution and its changing characteristics in topic evolution research. The themes in library and information science mostly corresponded to multiple semantic concepts, and there were three evolutionary patterns: convergence, diffusion, and stability, while the popularity of themes was independent of their evolutionary dynamics.

Lakshmanan and Anandha (2024) proposed a method that combines blockchain technology and a new privacy protection model to address the security vulnerabilities and privacy issues that medical data may face when transmitted through open communication channels. This method improved the privacy, reliability, and practicality of electronic medical models through two stages of data cleaning and recovery, and had significant advantages in comparison. Schulze et al. (2023) proposed a deep learning-based internal discrimination algorithm and an outer discrimination algorithm to handle the issue of privacy leakage in laparoscopic surgery videos. This algorithm was used to automatically identify and mask non-abdominal areas in videos to protect patient privacy. The outer discrimination algorithm performed well in classifying non-abdominal frames, with an average F1 score of 0.96 ± 0.01 (binary classification) and 0.97 ± 0.01 (fifth classification). Cao et al. (2022) proposed a blockchain privacy protection data mining algorithm based on decision tree classification to address the problems of low mining accuracy and high data noise in privacy protection data mining methods in blockchain. The mining accuracy of this algorithm was always above 90%, and the data noise was stable below 0.6 dB.

Although the above-mentioned studies have proposed innovative privacy protection or data processing methods in specific fields and achieved certain results, there are still some common or individual shortcomings. Firstly, some studies lack broad applicability validation for datasets of different types or sizes, and the results may not be universally applicable. Secondly, some methods may sacrifice computational efficiency or resource consumption while pursuing performance improvement, which may pose challenges in practical applications (Dhinakaran & Prathap, 2022). Therefore, based on word2vec's word vector representation, this study introduces HS to optimize the efficiency of traditional softmax, and ultimately constructs the LPPM-W2V model.

LPPM-W2V MODEL

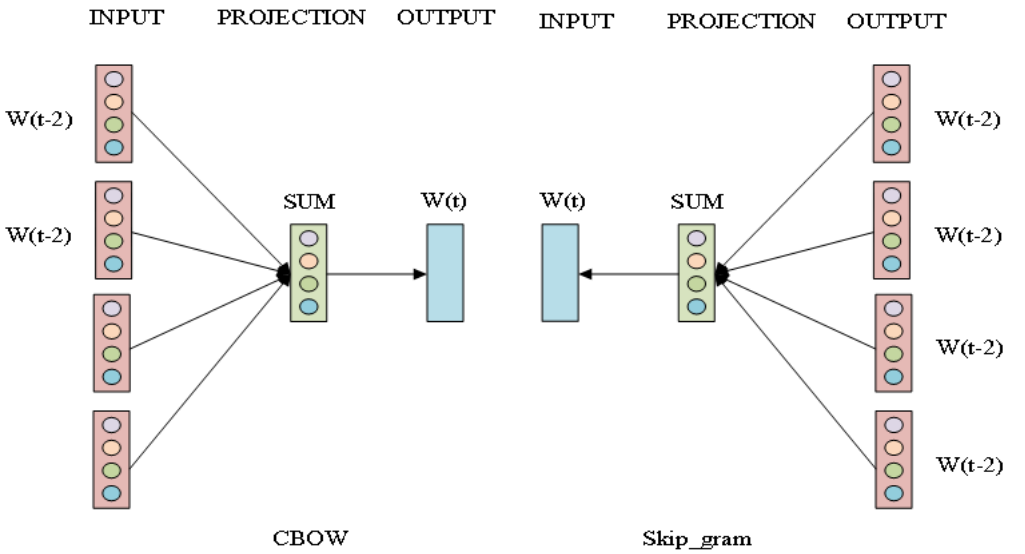
Improvement of Word2vec

In LPP, identifying and processing sensitive information is an important task for Dynamic Topic Model. Word2vec is utilized to generate word vectors. The model is a shallow and double-layer neural network, which is defined by words and requires guessing the input of adjacent positions (Balfagih et

al., 2022). Words are adopted to train to reconstruct linguistic word texts. The word2vec algorithm can be used in sensitive content recognition technology by converting text into vector representations and utilizing the spatial distance between vectors to determine whether sensitive words or content exist in the text (Ma et al., 2023). Sensitive texts for LPP are specifically identified according to the personal or sensitive data they contain, such as personal name, contact information, ID number, financial details, and any other information. Once disclosed, they may violate privacy laws or regulations. One may now refer to key legal frameworks, such as the Personal Privacy Protection Law, the General Data Protection Law, and relevant local laws to better understand the scope of these sensitive texts.

The schematic diagram of the word2vec algorithm's structure is displayed in Figure 1.

Figure 1. The schematic diagram of the word2vec algorithm's structure



Note. CBOW = continuous bag of words.

Word2vec is a word embedding method based on neural networks that generates word vectors by learning context information from a large number of corpora. Its computational requirements are usually related to the size of the training corpus, the complexity of the model (such as dimension size), and the optimization algorithm. The training time of word2vec is affected by the following factors: 1) corpus size: a larger corpus means more training iterations; 2) word vector dimension: higher word vector dimensions (such as 300 dimensions) usually require more computing resources; and 3) window size: the context window determines the context range of each word during training. A larger window will require the model to deal with more contexts, thus increasing the computational burden.

For large-scale corpus training, a GPU or a high-performance CPU cluster is used to speed up the training process. In training, a large-scale corpus may lead to a large demand for intermediate data storage. It is recommended to have enough RAM to store the data structure in the training process. Word2vec creates a word vector for each word, so the size of the vocabulary directly affects the memory requirements of the model. In this study, each word vector has 200 dimensions, and each dimension is stored by a 4-byte floating-point number, so the memory occupied by each word vector is about 800 bytes. For a model with millions of words, the total memory required for the model is 800 MB. During training, the model needs to store all kinds of temporary variables, which will also consume a lot of memory.

The word2vec model has two training models: continuous bag of words (CBOW) and skip-gram (Choudhary & Beniwal, 2021). CBOW focuses on predicting the central word W(t) through contextual vocabulary, such as W(t-2) to W(t+2). Its structure goes from the layers of input through the projection (the hidden layer [HL]) to the output, capturing the co-occurrence relationship between words through weight matrix mapping and summation operations. On the contrary, skip-gram operates in reverse, predicting its contextual vocabulary from the central word W(t), generating the probability distribution of contextual vocabulary through a dot product operation, and strengthening the semantic connection between words and context (Santos et al., 2021). Legal texts often contain a large number of professional terms and complex sentence structures. Therefore, in legal text analysis, using skip-gram is more advantageous than CBOW. This is because skip-gram predicts the context of the target word and can more finely depict subtle differences in meaning, which is crucial for understanding the precise meaning of legal provisions. Its HL can be expressed as in (1).

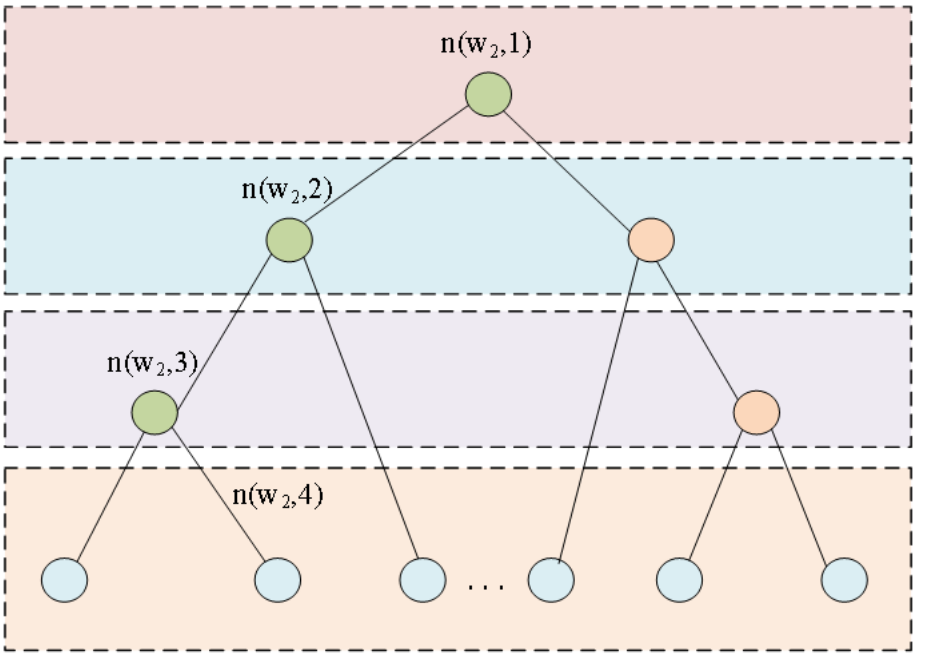
$$h = W_{(K,\cdot)}^{T} := v_{w_i}^{T} \tag{1}$$

where h is the output vector of the HL. $W_{(K,\cdot)}^{T}$: is the transpose of the k-th row of the weight matrix W. In the skip-gram, W is utilized to store vector representations of vocabulary, with each row corresponding to a vector of vocabulary. $v_{w_i}^{T}$ is the transpose of the vector representation of the word w_i. In skip-gram, each word has a corresponding vector representation used to capture semantic relationships between words. The probability value after softmax processing is represented in (2).

$$p\left(w_{c,j}|w_k\right) = y_{c,j} = \frac{exp\left(u_{c,j}\right)}{\sum_{j'=1}^{V} exp\left(u_{c,j'}\right)} \tag{2}$$

where $p\left(w_{c,j}|w_k\right)$ is the probability of the occurrence of contextual words $y_{c,j}$ and $w_{c,j}$ given the central word w_k. $u_{c,j}$ is the score or logarithmic probability of $w_{c,j}$ before performing softmax processing. It is obtained by a dot product operation of the vector representation of the central word and the contextual vocabulary. V is the vocabulary size, which is the gross of words considered in the model. $exp\left(u_{c,j}\right)$ is the exponential function value of $u_{c,j}$, used to convert scores into non-negative values for probability calculations. In legal text analysis, a large vocabulary and high precision are required. To make skip-gram more efficient and practical in processing such datasets, this study introduces HS to optimize the efficiency of traditional softmax on large-scale vocabulary (Yang et al., 2020). It reduces the computational complexity of traversing the entire vocabulary to the log level by constructing a binary tree and assigning a path from root to leaf for each word, as shown in Figure 2.

Figure 2. Structural schematic diagram of HS

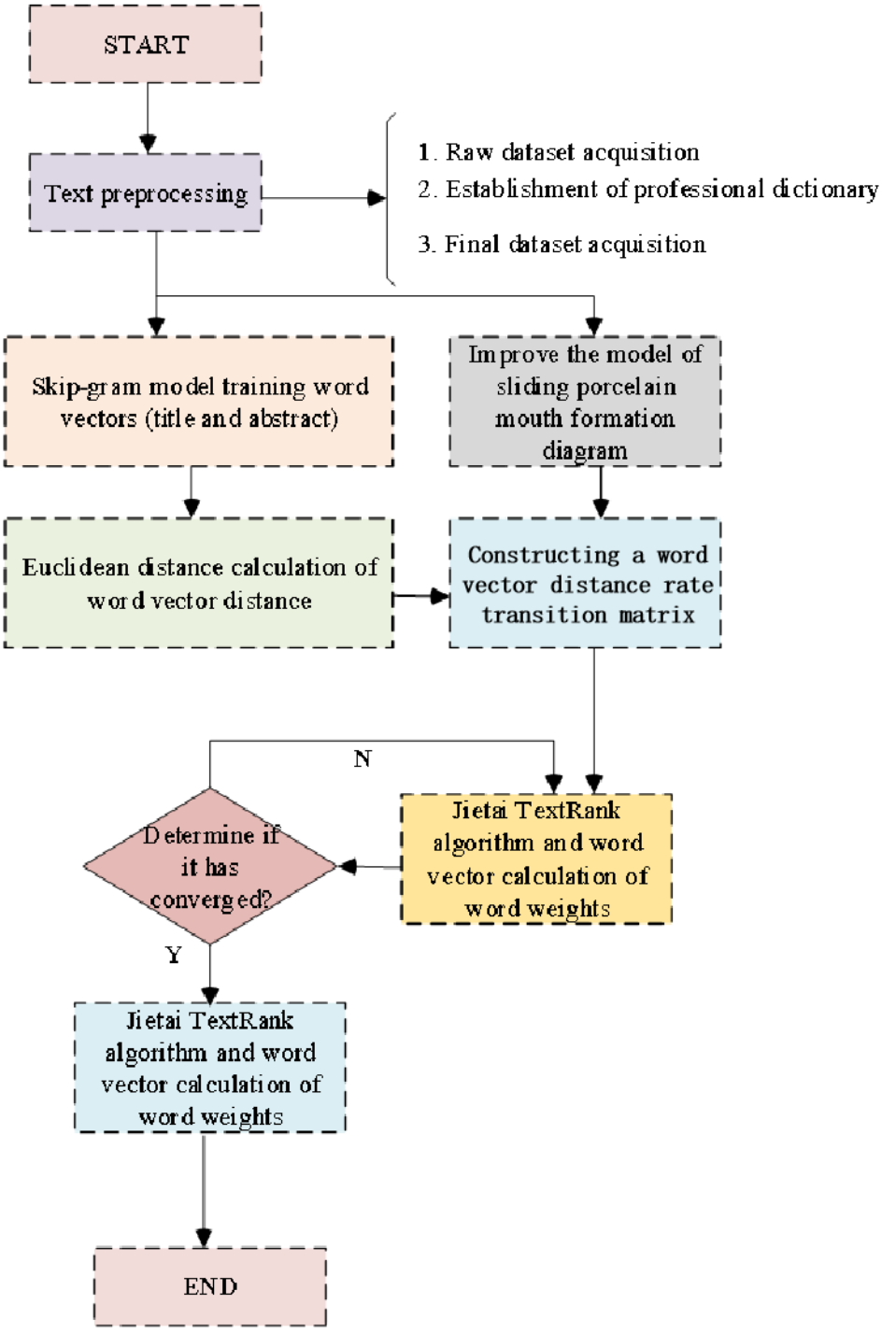


The HS model is an effective optimization method for the softmax function. In the HS model, when it is necessary to estimate the probability of a certain feature word,, the probability distribution of its category is evaluated to consider the attributes of the feature word itself, thereby achieving a more accurate estimation of the probability of the feature word (Zhu et al., 2020). The model will track the path from the root node of the tree to the leaf node where it is located. The model assigns a weight vector to each of the V-1 non-leaf nodes in the binary tree. For the sake of convenience, this study once again refers to these vectors as "output vectors" and uses symbols to represent the vectors of nodes. Furthermore, the probability of generating the feature word w is defined as a function of all non-leaf node vectors along the path from the root node to the leaf nodes representing the feature word. The *p(w)* referring to the probability of generating the feature word *w* is defined in (3).

$$p\left(w\right) = \prod_{j=1}^{L(w)-1} \sigma\left(\begin{matrix} switch(n(w,j+1) \\ = ch\left(n\left(w,j\right)\right)\right) \cdot v'_{n(w,j)}{}^{T} \cdot h \end{matrix}\right) \tag{3}$$

where *L(w)* is the path length from the root node to the leaf nodes representing *w*. *n(w,j)* is the j-th node on the path from the root node to the leaf node represented by *w*. *σ(·)* is the sigmoid function, and *switch* is an indicator function used to return different values based on conditions. *ch(n(w,j))* is the sub node type of *n(w,j)* to determine which direction of the vector should be used within the sigmoid function. $v'_{n(w,j)}$ is the weight vector of *n(w,j)*. This structure effectively reduces the number of parameters required for calculating the softmax function. Based on this, this study constructs a keyword extraction model (KEM), whose structure is exhibited in Figure 3.

Figure 3. Flow chart of the KEM



The logical process begins with the text preprocessing stage, which includes obtaining the original dataset, establishing a professional dictionary, and obtaining the final dataset. Next, the model uses

the skip-gram algorithm to train word vectors, especially for the title and abstract sections, to capture key information of the text. By calculating the Euclidean distance between word vectors, the model constructs a probability transition matrix for word vector distance. On this basis, the model combines the TextRank algorithm and word vectors to calculate the weight of words, and iteratively calculates until convergence to determine the final keywords.

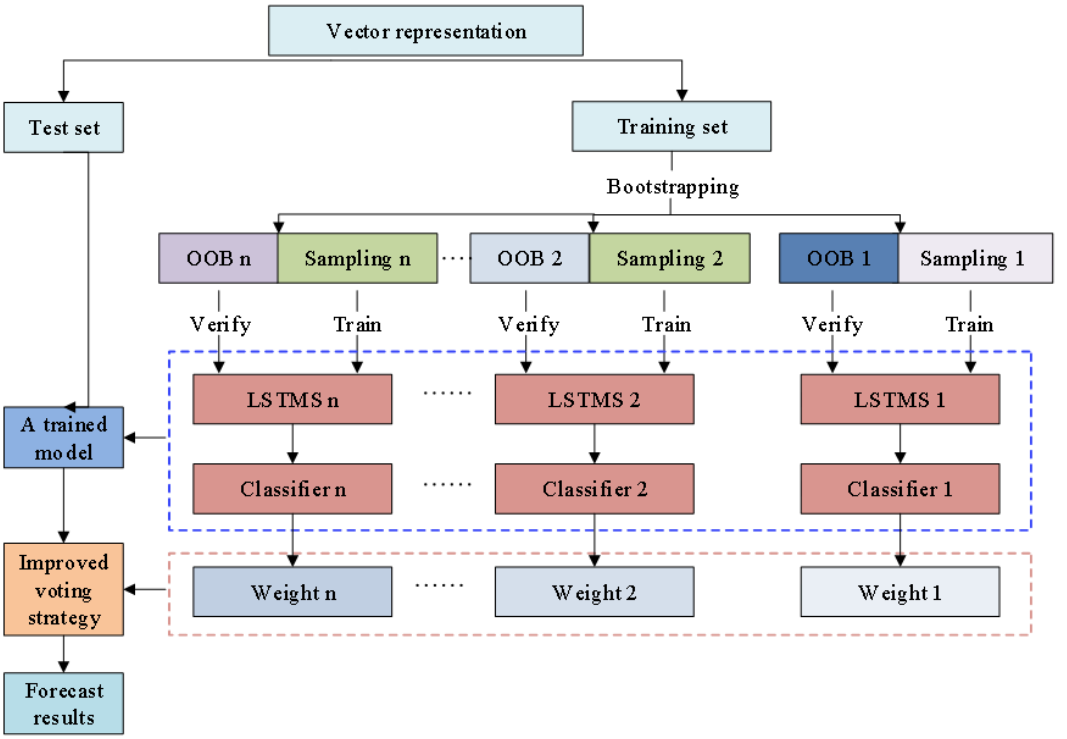
When implementing word2vec, one may encounter the following challenges:

- The vocabulary of a training corpus can be very large. Therefore, how to efficiently store and access vocabularies, especially when memory is limited, is a big challenge. It is common to use hash tables to store vocabularies, but it is necessary to ensure the speed of handling and searching hash conflicts.
- Although HS can improve efficiency in theory, it may need to deal with complex tree structure and efficient calculation methods in actual deployment.
- For the training of large-scale data sets, the implementation of word2vec usually supports multithreading, but it needs to optimize the synchronization between threads and access to shared data.
- For the growing corpus, word2vec may need to support incremental training. Incremental training requires that new data be continuously trained on the basis of the existing word vector model without retraining the whole model.

Painting Image Classification Based on Neural Networks and Support Vector Machines

After converting the text into a vector representation and extracting keywords, a classification algorithm is combined to classify the text and identify text containing sensitive information. Next, this study utilizes multiple long short-term memory (LSTM) models to form an ensemble, with each model capturing complex patterns in text data for different data distributions and constructing a new LSTM-B classifier. Its structure is shown in Figure 4. The classifier is integrated into a strategy to obtain strong classifier prediction results. Predictive category labels are very important for classifying legal texts according to their sensitivity. They predict categories to reflect the specific characteristics of the analyzed text, such as “highly sensitive,” “moderately sensitive,” and “low sensitive.”

Figure 4. Structural diagram of the LSTM-B classifier



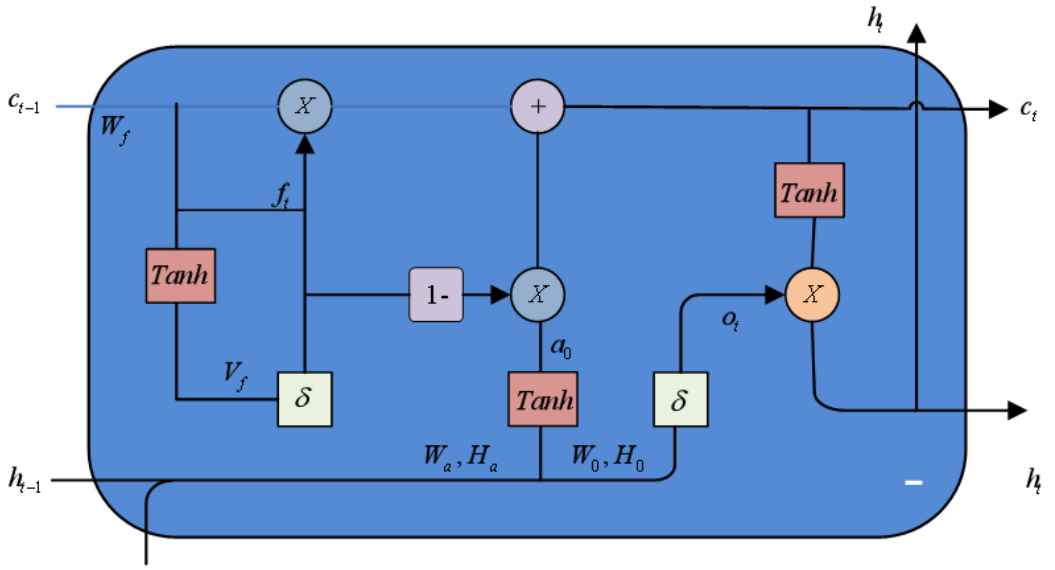
Note. OOB = out of bag.

The training set generates multiple sampling sets and corresponding validation sets (i.e. out of bag [OOB] data) through the bootstrap method. Each sampling set is used to train an LSTM model and obtain a classifier. The test set is taken to verify the accuracy of the trained model. The trained model is composed of multiple classifiers. Each classifier has its own weight, and the final prediction result is obtained through weighted voting. This study will improve the boosting voting strategy based on OOB data, calculate the reliability coefficient of the base classifier using OOB data, and select the base classifier for different categories of data. The reliability coefficient is represented in (4).

$$C_{in} = \frac{OOB\_F_{in}}{\sum_{n-1}^{N} OOB\_F_{in}} \tag{4}$$

where C_{in} is the reliability coefficient (reliability coefficient) of the i-th base classifier for category n. OOB_F_{in} is the number of correct classifications of category n by the i-th base classifier on OOB data. Multiple different sampling sets are obtained through multiple sampling, each of which is randomly selected from a certain number of samples in the original training set and allows for repeated selection. Each sampling set is used to train an LSTM model, resulting in a classifier. To further reduce the internal parameters of the neural network, this study introduces the attention model (AM) and LSTM to construct a new AM-LSTM neural structure, as shown in Figure 5.

Figure 5. Schematic representation of the AM-LSTM neuron structure



On the basis of classical LSTM, attention gates are added as regulatory mechanisms to adjust the interaction strength between input x and hidden state H_a through weights a_0 and a_1. AM not only affects the input processing of LSTM, but also directly acts on the output gate of LSTM through its output, finely regulating the final output h_t. The LSTM core in this structure—input/forget/output gates and unit state—works together to manage the flow and storage of information. The introduction of AM enables the model to focus more on key information in the input, reduce the interference of non-key information, and optimize the sequence modeling capability. Meanwhile, by processing the interaction between input and hidden states through functions f_1 and f_2, new unit states c_t and output h_t are generated, further enhancing the flexibility and expressiveness of the model. The constructed LPPM-W2V model is shown in Figure 6.

Figure 6. Flow diagram of the LPPM-W2V model

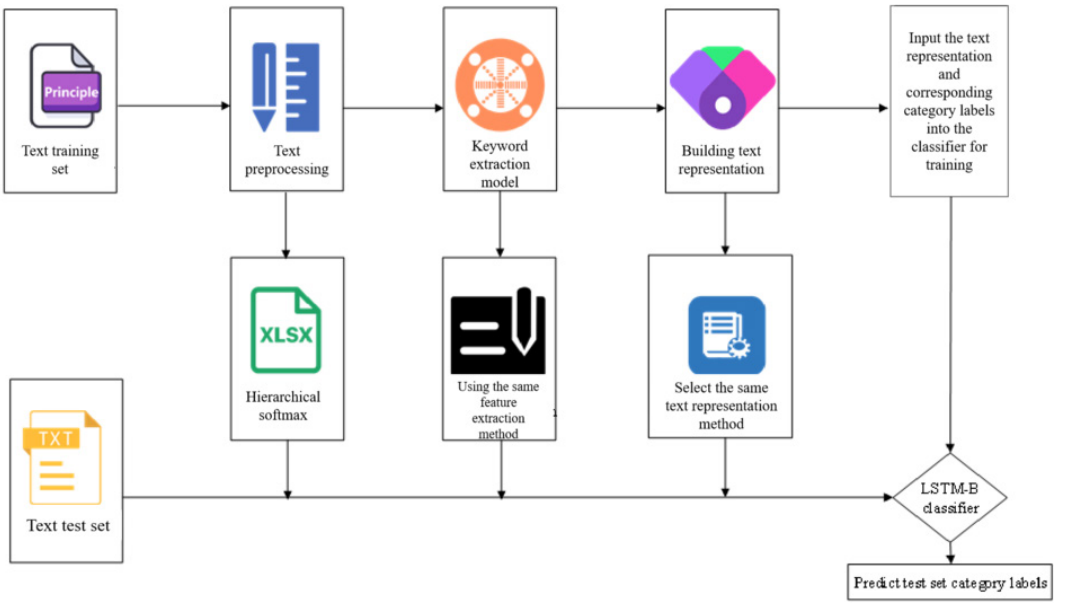


Figure 6 shows the comprehensive structure of the LPPM-W2V model. This model integrates multiple technical elements, including word2vec's word vector representation, HS optimization, KEM, and LSTM-B classifier (Liu & Wang, 2022; Wang & Poor, 2022). Firstly, word2vec converts legal text into vector representations using the skip-gram model, capturing semantic relationships between words. HS has optimized the efficiency issue of traditional softmax, making the model more efficient in handling large-scale vocabulary lists. Next, KEM uses the distance between word vectors and the TextRank algorithm to extract keywords from the text to identify sensitive content. Finally, the LSTM-B classifier integrates multiple LSTM models and obtains the final prediction result through weighted voting, which is used to classify text containing sensitive information (Bi et al., 2022). The collaborative work between various elements constitutes this efficient and practical LPP model.

APPLICATION EXPERIMENT AND PERFORMANCE TESTING OF THE LPPM-W2V MODEL

Performance Testing of the LPPM-W2V Model

To improve the accuracy of legal privacy sensitive word recognition and more efficiently perform LPP, this study designed a new LPPM-W2V model. To verify its superiority, the paper compares the white-box adversarial example generation algorithms-whale optimization algorithm (IWA-WOA) proposed in reference (Guo et al., 2024) with the LSTM-genetic algorithm (LSTM-GA) proposed in reference (Sun et al., 2022). The experiment implemented these algorithms on the Windows 10 platform using Python language and trained them on the Reuters-21578 dataset.

The Reuters-21578 data set, published by Reuters in 1987, is one of the earliest and most influential text classification data sets in the field of natural language processing. The Reuters-21578 data set is famous for its rich topic classification and high-quality text content. This data set contains 21,578 news articles covering a variety of topics and categories, providing researchers with a standardized benchmark to evaluate the performance of text classification algorithms. Its multi-label classification features enable researchers to explore more complex text analysis tasks. In addition,

the openness and wide use of data sets make it a standard benchmark for evaluating and comparing the performance of different algorithms. In practical application, the Reuters-21578 data set is used to build and optimize a news recommendation system. In addition, it is also used in financial risk assessment and market prediction to help financial institutions make more informed decisions. In the field of law and policy, Reuters-21578 is also used in text mining and public opinion analysis to monitor and predict the public's reaction and attitude to specific events.

Table 1 shows the experimental results.

Table 1. Classification effect of each model in the case of text vector dimension 200

Attribute domain	LSTM-GA		IWA-WOA		LPPM-W2V	
	Indicator 1	Indicator 2	Indicator 1	Indicator 2	Indicator 1	Indicator 2
D2V	0.8882	0.7770	0.7685	0.7685	0.8945	0.5391
LDA	0.7423	0.8393	0.8398	0.8393	0.8797	0.8398
RP	0.7685	0.7685	0.8391	0.8945	0.8755	0.8315
LSI	0.8652	0.9219	0.9219	0.7491	0.8755	0.9302
BOW	0.8941	0.8941	0.8357	0.8803	0.8941	0.8941

Note. Indicator 1 is the macro classification performance (Fl_macro), and indicator 2 is the micro classification performance (Fl_micro). D2V = doc2vec; LDA = latent Dirichlet allocation; RP = random projection; LSI = latent semantic indexing; BOW = bag of words.

Firstly, for the doc2vec (D2V) attribute domain, LPPM-W2V reaches a high value of 0.8945 on indicator 1, the macro classification performance (Fl_macro), which is significantly higher than IWA-WOA's 0.7685 and LSTM-GA's 0.8882. At the same time, it also shows a performance of 0.5391 on indicator 2, the micro classification performance (Fl_micro), which is slightly lower than LSTM-GA's 0.7770. In the latent Dirichlet allocation (LDA) attribute domain, LPPM-W2V also performs well, with an Fl_macro of 0.8797, slightly higher than IWA-WOA's 0.8398, but the Fl_micro is as high as 0.8398, the highest among all algorithms, indicating high accuracy in classifying individual categories. In the random projection (RP) attribute domain, the LPPM-W2V model achieves 0.8755 on the Fl_macro, which is the highest among all algorithms, but slightly lower than the 0.8945 on IWA-WOA on the Fl_micro. For the latent semantic indexing (LSI) attribute domain, LPPM-W2V reaches 0.9302 on the Fl_micro, which is the highest among all algorithms, but slightly lower than IWA-WOA's 0.9219 on the Fl_macro. In the bag of words (BOW) attribute field, the Fl_macro and Fl_micro values of LPPM-W2V are 0.8941, which is the same as LSTM-GA, but higher than the 0.8357 and 0.8803 values of IWA-WOA.

Overall, LPPM-W2V exhibits high classification performance in most attribute domains, especially in the Fl_micro metric, reaching its highest value multiple times. This indicates that LPPM-W2V has a significant advantage in individual classification accuracy. Although in some cases, Fl_macro is slightly lower than that of IWA-WOA, overall, LPPM-W2V performs well in the recognition of legal privacy sensitive words, with high accuracy and good application prospects. In addition, the experiment also records the changes in the mean squared error (MSE) and the root mean squared error (RMSE) when training these algorithms, as displayed in Figure 7.

Figure 7. Changes in MSE and RMSE during the training process of three algorithms

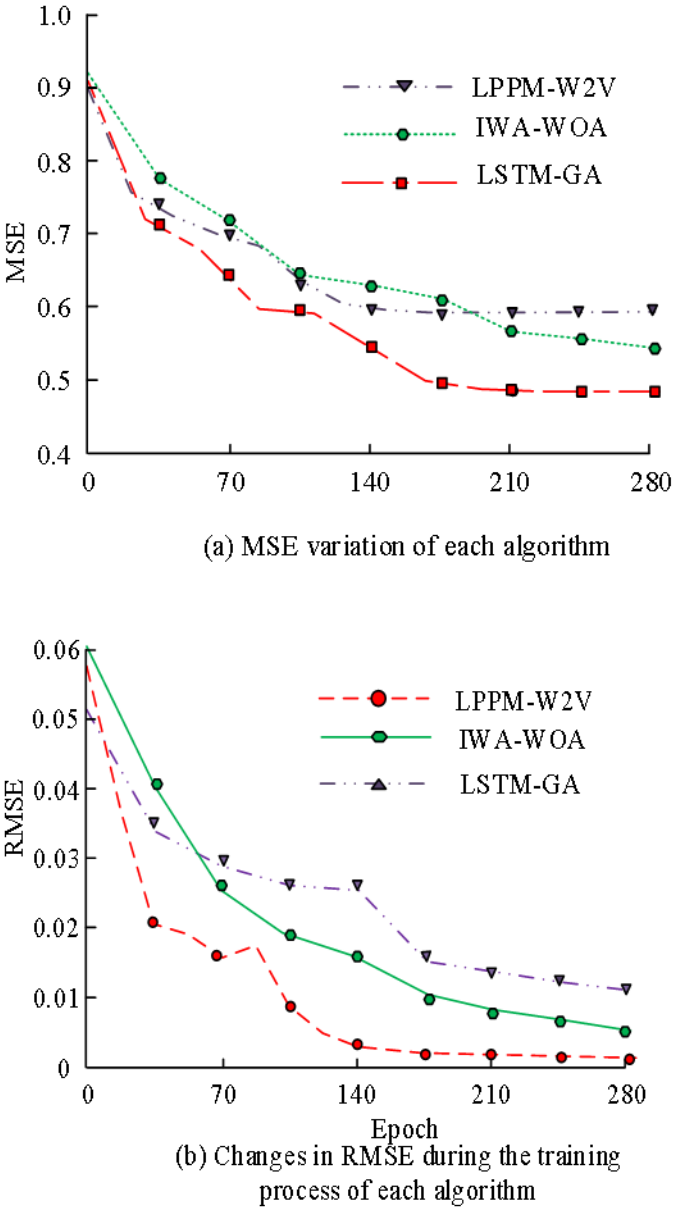
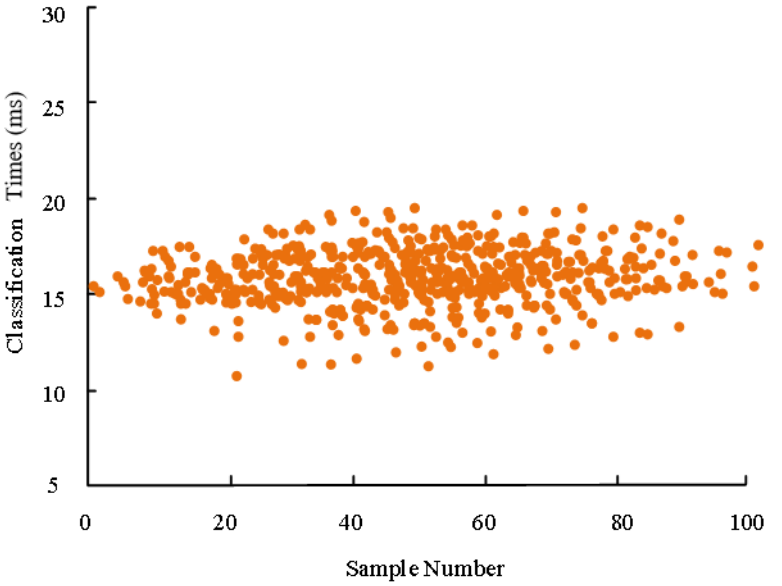
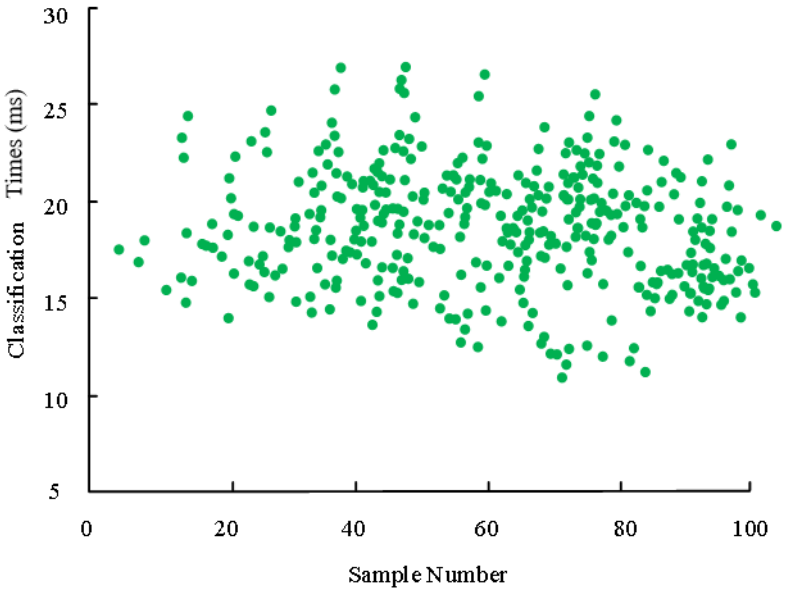


Figure 7 shows the MSE and RMSE changes of LSTM-GA, LPPM-W2V, and IWA-WOA during the training process. In the initial stage, the MSE and RMSE of IWA-WOA are the highest, at 0.900 and 0.060, and then gradually decrease to about 0.55 and 0.005. LSTM-GA steadily decreases from an MSE of 0.89 and an RMSE of 0.059 to approximately 0.61 and 0.013. LPPM-W2V performs particularly well, with its MSE decreasing from 0.89 to about 0.48 and its RMSE decreasing from 0.051 to about 0.001. LPPM-W2V performs the best in both MSE and RMSE, demonstrating the best fitting effect on training data. Finally, this study compares the classification time of LPPM-W2V and IWA-WOA, which performed well, as shown in Figure 8.

Figure 8. Comparison of classification time between LPPM-W2V and IWA-WOA



(a) LPPM-W2V model classification time consumption



(b) IWA-WOA model classification time consumption

In Figure 8, for the LPPM-W2V model, the classification time for most samples is concentrated between 5ms and 15ms, while IWA-WOA is mainly distributed between 10ms and 20ms, with some outliers exceeding 20ms. This may indicate that IWA-WOA may encounter a higher computational burden in certain situations. In addition, the maximum classification time of LPPM-W2V does

not exceed 20ms, while IWA-WOA even takes about 30ms. This further proves the advantage of LPPM-W2V in processing speed.

Application Experiment of the LPPM-W2V Model

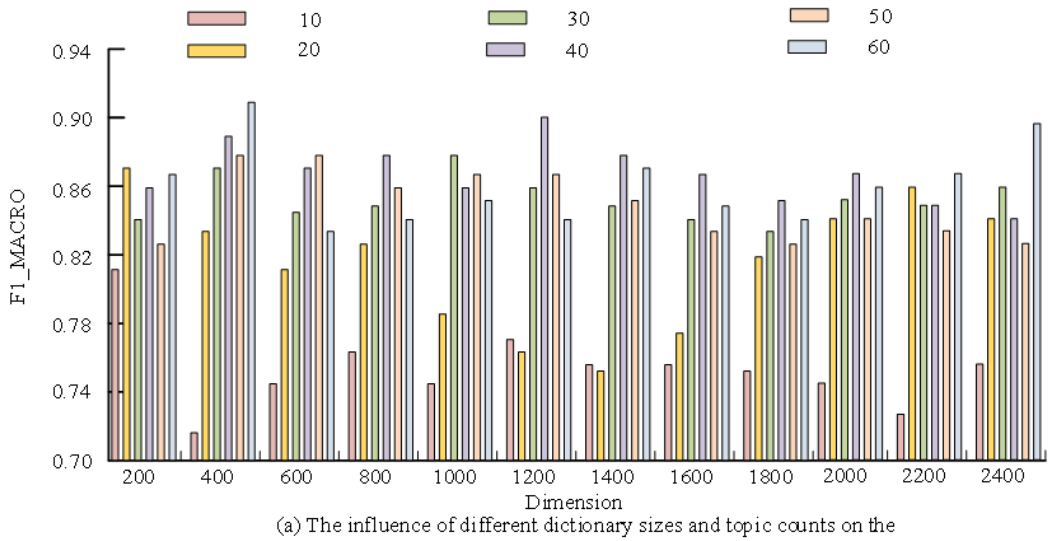
This study fully validates the superior performance of the LPPM-W2V model. To further validate the model's equally good performance in the actual LPP field, this study applies it to the actual legal sensitive word recognition process and conducts model ablation experiments. Accuracy, recall, F1 score, and processing speed are evaluation indicators, as shown in Table 2.

Table 2. Results of the ablation experiments of the LPPM-W2V model

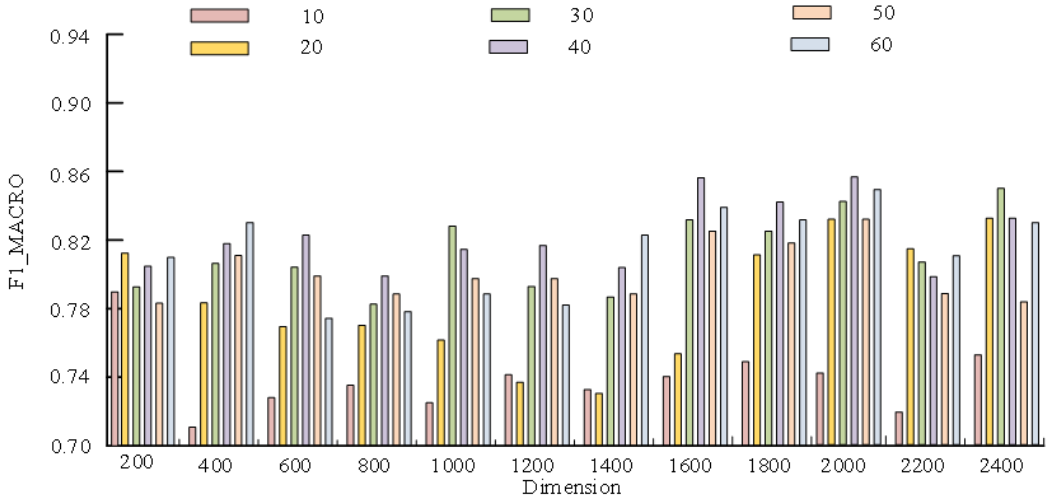
Method / parameter configuration	Accuracy(%)	Recall(%)	F1 score(%)	Processing speed (ms)
Baseline model, with no special configuration	85.22	80.31	82.13	250.03
The skip-gram and the HS are used to optimize	87.56	83.78	85.62	229.87
Word weights are calculated by combining the TextRank algorithm and the word vector price	89.65	85.98	87.76	298.56
Integrated LSTM model, using the bootstrapping square law	91.87	87.65	89.71	348.76
LPPM-W2V	92.56	88.78	90.62	319.87

In Table 2, compared with the baseline, the model optimized using skip-gram and HS improved accuracy by 2.34%, recall by 3.47%, F1 score by 3.49%, and processing speed by 20.16ms. Secondly, the model that combined the TextRank algorithm and word vector to calculate word weights further improved performance, with an accuracy increase of 4.43% compared to the baseline model, a recall increase of 5.67%, and an F1 score increase of 5.63%. However, the processing speed slowed down slightly, increasing by 48.53ms. Finally, the model that integrated the LSTM model and used the bootstrap method, as well as the complete LPPM-W2V model, showed better performance. Compared with the baseline, LSTM improved accuracy by 6.65%, recall rate by 7.34%, and F1 score by 7.58%, but its processing speed was relatively slow, increasing by 98.73ms. LPPM-W2V improved accuracy by 7.34%, recall by 8.47%, F1 score by 8.49%, and processing speed by 69.84ms. To further test its superiority, this study introduced the Operating Systems Design and Implementation proposed in reference (Du et al., 2023) for comparison, as shown in Figure 9.

Figure 9. Effect of dictionary size and topic counts on the classification accuracy of the two models



(a) The influence of different dictionary sizes and topic counts on the classification performance of legal texts in the LPPM-W2V model

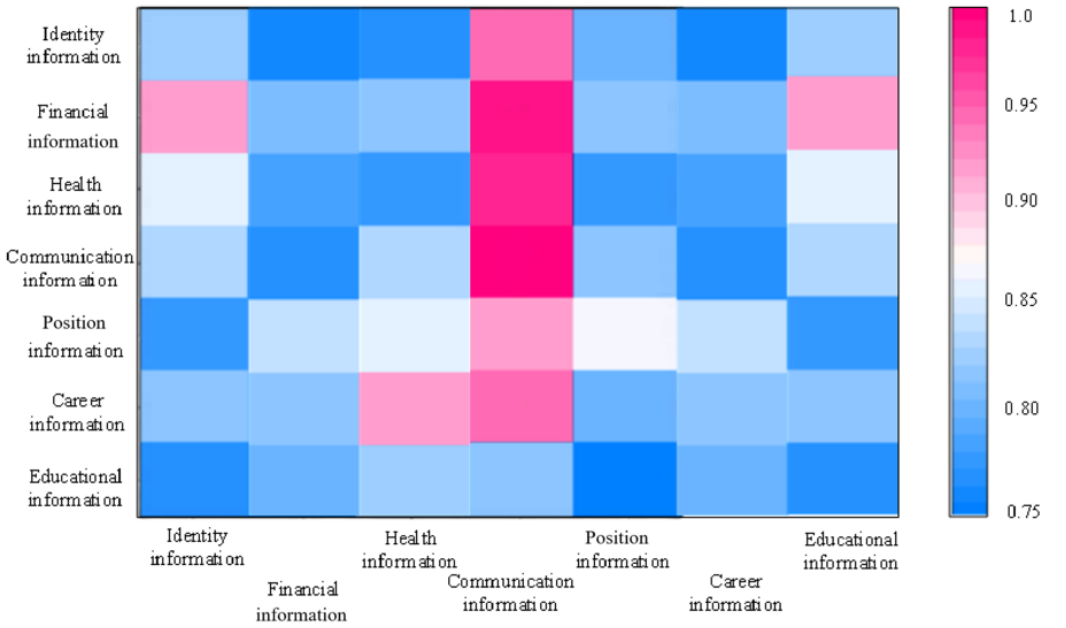


(b) The influence of different dictionary sizes and topic counts on the classification performance of legal texts in the OSDI model

In Figure 9(a), as the number of topics grows, the F1_macro value generally shows an upward trend followed by a downward trend. Specifically, when the number of topics is 10, the F1_macro value is relatively low, at 0.79. When the number of topics increases to 20/50, the F1_macro value increases to 0.86 and decreases to 0.78. This may mean that the model has encountered difficulties in capturing more detailed thematic information. The F1_macro value of the research model generally shows a gradually increasing trend with the increase of dictionary size. The average F1_macro value increases from 0.83 when the dimension is 200 to 0.88 when the dimension is 2400. In Figure 9(b), the Operating Systems Design and Implementation model performs poorly, with an average F1_macro

value of only around 0.81 during the training process. The confusion matrix results of the LPPM-W2V model in the task of legal sensitive word recognition are shown in Figure 10.

Figure 10. Confusion matrix of the LPPM-W2V model in the legally sensitive word recognition task



In Figure 10, the model performs the best in identifying the two types of legal privacy, "identity information" and "health information," with accuracy rates of 0.95 and 0.90, respectively. This means that 95% and 90% of the samples identified as "identity information" or "health information" are correctly classified. This indicates that the model has strong recognition ability for these two types of information. In the recognition of "communication information" and "location information," the accuracy is 0.85 and 0.80 respectively, indicating that the model's performance in these two aspects is slightly lower than the first two types of information. For "occupational information" and "educational information," the accuracy further decreases to 0.80 and 0.75, which means that the model encounters more challenges when processing these two types of information. The accuracy of "occupation" and "education" information is low, which may be due to the small proportion of these categories in the training data, resulting in the model failing to fully learn the characteristics of these categories. One can carry out domain-specific pre-training, especially for the text data related to "occupation" and "education" on a larger scale. For example, one could collect more text data from different occupational backgrounds, covering sensitive information related to occupations, such as work experience, job description, company name, etc. For "education" information, one can pay special attention to information such as degree, school name, major, and grades. In the process of training, one can weight different kinds of information. For example, giving higher weight to the information categories of "occupation" and "education" makes the model pay more attention to the identification of these categories. After that, through continuous evaluation and feedback, the performance of the model on different types of information is further optimized.

CONCLUSION

This study proposed an LPPM-W2V model to handle the issue of legal privacy information disclosure in the internet era. For the D2V attribute domain, LPPM-W2V reached a high value of 0.8945 on the Fl_macro, significantly higher than IWA-WOA's 0.7685 and LSTM-GA's 0.8882. Meanwhile, it also showed a performance of 0.5391 on the Fl_micro, which was slightly lower than LSTM-GA's 0.7770. In addition, LPPM-W2V also performed the best on the MSE and the RMSE, proving its best fitting effect on training data. In terms of classification time, LPPM-W2V took no more than 20ms, significantly faster than the maximum time of IWA-WOA (around 30ms). The ablation experiment further validated the performance improvement of the LPPM-W2V model, with accuracy, recall, and F1 score reaching 92.56%, 88.78%, and 90.62%, respectively, which were 7.34%, 8.47%, and 8.49% higher than the baseline model. Especially when identifying "identity information" and "health information," the accuracy was as high as 0.95 and 0.90, demonstrating the powerful ability of the model in processing such information.

To sum up, the LPPM-W2V model significantly improved the accuracy and efficiency of legal privacy sensitive word recognition and provided a powerful tool for LPP in the internet era. It performed particularly well in handling identity information and health information, with accuracy rates of 0.95 and 0.90. Although the LPPM-W2V model performed well in most tests, its accuracy in handling occupational and educational information was only 0.80 and 0.75. This suggests that the model may have limitations when facing such specific information. This study focuses on specific data sets and attributes, and tests different legal data sets from different regions and legal systems, which will be the next research direction of this study. And in order to improve the cross-language ability of the model, machine translation technology and cross-language pre-training model will be introduced. Testing legal documents in different languages will help to evaluate the performance of the model in cross-language situations and further improve the applicability of the model. With the change of the legal environment, the legal language itself is constantly developing, so the model can be optimized and updated according to the new legal text data through continuous learning. This is very important to maintain the validity and applicability of the model in the changeable legal environment.

COMPETING INTERESTS

The author of this publication declares there are no competing interests.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. Funding for this research was covered by the author of the article.

PROCESS DATES

Received: September 18, 2024, Revision: November 21, 2024, Accepted: November 29, 2024

CORRESPONDING AUTHOR

Correspondence should be addressed to Rongrong Wang; fe244nhcd8351@163.com

REFERENCES

Amin, S. N., Shivakumara, P., Jun, T. X., Chong, K. Y., Zan, D. L. L., & Rahavendra, R. (2023). An augmented reality-based approach for designing interactive food menu of restaurant using Android. *Artificial Intelligence and Applications (Commerce, Calif.)*, *1*(1), 26–34. DOI: 10.47852/bonviewAIA2202354

Balfagih, A., Keselj, V., & Taylor, S. (2022). N-gram and word2vec feature engineering approaches for spam recognition on some influential Twitter topics in Saudi Arabia. In *Proceedings of the 6th international conference on information system and data mining* (pp. 101-107). Association for Computing Machinery, New York, NY, United States. DOI: 10.1145/3546157.3546173

Bi, R., Zhao, M., Ying, Z., Tian, Y., & Xiong, J. (2022). Achieving dynamic privacy measurement and protection based on reinforcement learning for mobile edge crowdsensing of IoT. *Digital Communications and Networks*, *10*(2), 380–388. DOI: 10.1016/j.dcan.2022.07.013

Cao, Y., Wei, W., & Zhou, J. (2022). Privacy protection data mining algorithm in blockchain based on decision tree classification. *Web Intelligence*, *20*(2), 103–112. DOI: 10.3233/WEB-210485

Chen, W., Wu, H., Chen, X., & Chen, J. (2022). A review of research on privacy protection of Internet of Vehicles based on blockchain. *Journal of Sensor and Actuator Networks*, *11*(4), 86. DOI: 10.3390/jsan11040086

Choudhary, K., & Beniwal, R. (2021). Xplore word embedding using CBOW model and skip-gram model. In *2021 7th international conference on signal processing and communication (ICSC)* (pp. 267-270). IEEE. DOI: 10.1109/ICSC53193.2021.9673321

Dhinakaran, D., & Prathap, P. J. (2022). Protection of data privacy from vulnerability using two-fish technique with Apriori algorithm in data mining. *The Journal of Supercomputing*, *78*(16), 17559–17593. DOI: 10.1007/s11227-022-04517-0

Du, H., Wen, Q., Zhang, S., & Gao, M. (2023). An improved conditional privacy protection scheme based on ring signcryption for VANETS. *IEEE Internet of Things Journal*, *10*(20), 17881–17892. DOI: 10.1109/JIOT.2023.3279896

Gao, Q., Huang, X., Dong, K., Liang, Z., & Wu, J. (2022). Semantic-enhanced topic evolution analysis: A combination of the dynamic topic model and word2vec. *Scientometrics*, *127*(3), 1543–1563. DOI: 10.1007/s11192-022-04275-z

Guo, C., Weijun, P., Jing, W., Youxuan, F., Keke, Y., & Yanshuang, X. (2024). A blockchain-based proxy re-encryption scheme with conditional privacy protection and auditability. *China Communications*, *21*(7), 267–277. DOI: 10.23919/JCC.fa.2022-0863.202407

Hebbi, C., & Mamatha, H. R. (2023). Comprehensive dataset building and recognition of isolated handwritten kannada characters using machine learning models. *Artificial Intelligence and Applications (Commerce, Calif.)*, *1*(3), 179–190. DOI: 10.47852/bonviewAIA3202624

Lakshmanan, M., & Anandha Mala, G. S. (2024). Merkle tree-blockchain-assisted privacy preservation of electronic medical records on offering medical data protection through hybrid heuristic algorithm. *Knowledge and Information Systems*, *66*(1), 481–509. DOI: 10.1007/s10115-023-01937-z

Liu, J., & Wang, S. (2022). All-dummy k-anonymous privacy protection algorithm based on location offset. *Computing*, *104*(8), 1739–1751. DOI: 10.1007/s00607-022-01067-4

Ma, J., Wang, L., Zhang, Y. R., Yuan, W., & Guo, W. (2023). An integrated latent Dirichlet allocation and word2vec method for generating the topic evolution of mental models from global to local. *Expert Systems with Applications*, *212*, 118695. DOI: 10.1016/j.eswa.2022.118695

Nsugbe, E. (2023). Toward a self-supervised architecture for semen quality prediction using environmental and lifestyle factors. *Artificial Intelligence and Applications (Commerce, Calif.)*, *1*(1), 35–42. DOI: 10.47852/bonviewAIA2202303

Ren, R., Yin, C., & S.-T. Yau, S. (2022). kmer2vec: A novel method for comparing DNA sequences by word2vec embedding. *Journal of Computational Biology, 29*(9), 1001-1021. DOI: 10.1089/cmb.2021.0536

Santos, F. A. O., Bispo, T. D., Macedo, H. T., & Zanchettin, C. (2021). Morphological skip-gram: Replacing FastText characters n-gram with morphological knowledge. *Inteligencia Artificial*, *24*(67), 1–17. DOI: 10.4114/intartif.vol24iss67pp1-17

Schulze, A., Tran, D., Daum, M. T. J., Kisilenko, A., Maier-Hein, L., Speidel, S., Distler, M., Weitz, J., Müller-Stich, B. P., Bodenstedt, S., & Wagner, M. (2023). Ensuring privacy protection in the era of big laparoscopic video data: Development and validation of an inside outside discrimination algorithm (IODA). *Surgical Endoscopy*, *37*(8), 6153–6162. DOI: 10.1007/s00464-023-10078-x PMID: 37145173

Sun, H., Tan, Y. A., Li, C., Lei, L., Zhang, Q., & Hu, J. (2022). An edge-cloud collaborative cross-domain identity-based authentication protocol with privacy protection. *Chinese Journal of Electronics*, *31*(4), 721–731. DOI: 10.1049/cje.2021.00.269

Wang, H., Han, G., Hou, Y., Guizani, M., & Peng, Y. (2021a). A multi-channel interference based source location privacy protection scheme in underwater acoustic sensor networks. *IEEE Transactions on Vehicular Technology*, *71*(2), 2058–2069. DOI: 10.1109/TVT.2021.3135438

Wang, H., Han, G., Zhang, Y., & Xie, L. (2021b). A push-based probabilistic method for source location privacy protection in underwater acoustic sensor networks. *IEEE Internet of Things Journal*, *9*(1), 770–782. DOI: 10.1109/JIOT.2021.3085586

Wang, J., Zhao, C., He, S., Gu, Y., Alfarraj, O., & Abugabah, A. (2022). LogUAD: Log unsupervised anomaly detection based on word2vec. *Computer Systems Science and Engineering*, *41*(3), 1207–1222. DOI: 10.32604/csse.2022.022365

Wang, Y., & Poor, H. V. (2022). Decentralized stochastic optimization with inherent privacy protection. *IEEE Transactions on Automatic Control*, *68*(4), 2293–2308. DOI: 10.1109/TAC.2022.3174187

Yang, S., Wei, R., Guo, J., & Tan, H. (2020). Chinese semantic document classification based on strategies of semantic similarity computation and correlation analysis. *Journal of Web Semantics*, *63*, 100578. DOI: 10.1016/j.websem.2020.100578

Zhu, Q., He, Z., Zhang, T., & Cui, W. (2020). Improving classification performance of softmax loss function based on scalable batch-normalization. *Applied Sciences (Basel, Switzerland)*, *10*(8), 2950. DOI: 10.3390/app10082950