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