Research on Key Point Acquisition and Intelligent Classification Technology

Niu Lu, Lu Kangning
Zhengzhou Logistic Support Center, Zhengzhou, Henan, China

Abstract: By classifying sensitive information, core technologies can be protected, ensuring information not accessed and used by unauthorized personnel, and preventing information from leakage, theft, and abuse. The effective classification work is able to establish an integral information protection mechanism, improve the level of information security, and protect the interests of all parties. In order to manage classified documents better, this paper proposes a multi-level processing method of dynamic key point acquisition and intelligent key point determination, and deep learning models are used to determine the level of document classification comprehensively. The main work of the paper is as follows: using Hidden Markov Model for part of speech tagging in stuttering segmentation algorithm, implementing word vectors by Word2Vec framework, and using CBOW model to predict center words based on contextual word information, and training the model through BERT to determine the classification level of the text automatically.

Keywords: Key Point Acquisition; Intelligent Encryption Point Determination; Skip-gram Algorithm; CBOW Model

1. Introduction

Confidentiality work can effectively prevent threats and risks faced by entities. Strictly control sensitive information through classification work. Its importance to entities lies in ensuring their security, maintaining their core interests, and maintaining their image and reputation. Establishing a sound classification system and mechanism, improving the level of confidentiality work, is of great significance for the development and stability of entities.

Traditional classification work requires the following processes: drafting classification documents, applying for permissions, reviewing data, confirming classification, implementing classification measures, publishing and distributing, using and transmitting, and conducting regular reviews.

However, the traditional process of classification has many problems, such as subjectivity and inconsistency, lack of evaluation and standards, risk of information leakage, and difficulty in management and protection.

2. Natural Language Processing Technology

2.1 Data Preprocessing

2.1.1 Word Segmentation Module Design

Jieba word segmentation is a widely used Chinese word segmentation algorithm, consisting of dictionary segmentation and prefix dictionary based segmentation methods. The following are the main characteristics and principles of the Jieba segmentation algorithm:

Jieba word segmentation is based on a large and accurate Chinese dictionary, which can be customized, expanded, and updated as needed. The stuttering segmentation can automatically recognize unrecorded words and use methods based on Chinese word formation ability and statistical information to segment new words. When a word cannot fully match the words in the dictionary, a prefix dictionary based segmentation strategy is adopted to divide the sentence into individual word segments, and then the final segmentation result is determined based on word frequency and rules.

2.1.2 Design of Part of Speech Annotation Module

In order to add the function of part of speech tagging in the Jieba word segmentation algorithm, a hidden Markov model (HMM) based approach is adopted for part of speech tagging. In text segmentation tasks, HMM can be used to infer the most likely segmentation result based on the observed character sequence.

Firstly, prepare a corpus with annotated parts of speech as training data, which should include the segmented text and corresponding part of speech labels. Use the words in the training data as observation sequences and the part of speech labels as corresponding state sequences. Calculate the frequency of transition from one part of speech to another in the training data, and normalize to obtain the probability of state transition. These probabilities represent the possibility of transitioning from one part of speech to another in a given context. Calculate the frequency of each word appearing under different parts of speech in the training data, and normalize to obtain the observation probability. These probabilities represent the likelihood that a given word belongs to different parts of speech. Smooth out the probability of state transition and observation to avoid zero probability when encountering unfamiliar words or parts of speech during the testing phase. Decode using the Viterbi algorithm to find the most likely part of speech sequence under a given observation sequence. The Viterbi algorithm considers both state transition probability and observation probability to maximize the joint probability of the entire sequence[1]. According to the Viterbi algorithm, label the most likely part of speech sequence for each word in the observed sequence.

2.1.3 Design of Word Vector Module

Using the Word2Vec model for keyword extraction on electronic documents can effectively capture semantic associations between words in the document. The steps to implement the word vector module using the Skip gram algorithm in the Word2Vec framework are as follows:

Collect a dataset containing a corpus, preprocess the corpus, construct a vocabulary based on the preprocessed corpus, use the vocabulary to generate training samples, use the Skip gram algorithm to select a central word as input, and then use its contextual words as output to form training samples. The Skip gram model consists of an embedding layer and an output layer[2]。The embedding layer converts the one hot encoding of words into word vector form, and then predicts the probability distribution of contextual words through the output layer. Train the Skip gram model using training samples, update the model parameters through optimization algorithms, and make the predicted results of the model as close as possible to the true probability distribution of contextual words. After training, extract word vectors from the Skip gram model as word representations[3]. These vectors can be used to calculate the similarity between words, thereby preparing for the next stage of natural language processing.

2.2 Keyword Extraction

2.2.1 TF-IDF Algorithm

TF-IDF is a commonly used text feature extraction method used to evaluate the importance of a word in the document it is in. TF (Word Frequency) represents the frequency of a word appearing in an article[4] 。 TF indicates that the higher the frequency of the word appearing in the article, the greater its importance to the document. IDF (inverse document frequency) is used to measure the importance of a word to the entire document set. The function of IDF is to penalize common words, reduce their weight, and weight rare words to increase their weight. A higher TF-IDF value means that the word is of high importance to a particular document and is unique and uncommon throughout the entire document collection[5]。

The steps of the TF-IDF algorithm are as follows:

1. Calculate the TF and IDF values for each word in the document..
2. Multiply the TF value of each word by its corresponding IDF value to obtain the TF-IDF value of that word.
3. Sort the words in the document based on the TF-IDF value to obtain the most important words.

TF-IDF can help extract keywords from documents, remove common but irrelevant words, and increase feature weights that are relevant to the document.

2.2.2 CBOW Model

The CBOW model consists of three layers: input layer, mapping layer, and output layer[2]。

CBOW is a model used to train word embedding vectors, which can predict the central word based on contextual word information.

Using the CBOW model for keyword extraction on electronic documents can effectively capture the semantic associations between words in the document. Firstly, collect a dataset containing

confidential electronic documents, preprocess them, and based on the preprocessed documents, establish a vocabulary table, assigning a unique index to each different vocabulary. This vocabulary will be used to convert vocabulary into corresponding numerical representations. The training sample of the CBOW model consists of contextual words and central words. For each central word, select a fixed size contextual window from the document, with words within this window as contextual words and central words as targets. Encode the training samples into numerical representations so that the model can process them, converting context words and central words into corresponding indexes. Constructing a CBOW model using neural networks, where the input of the model is the index of contextual words and the output is the index of central words[6].The hidden layer in the middle can be a fully connected layer or an embedded layer. Train the CBOW model using the encoded training samples. Optimization algorithms such as gradient descent can be used to update the weights of the model based on the difference between the predicted and actual values of the central word. After training, extract word embedding vectors from the CBOW model. These vectors will use pre trained weights to apply word embedding vectors to keyword determination tasks, and use cosine similarity algorithms to find the words closest to the keywords. Based on similarity or distance scores, a threshold can be set to select words with scores higher than the threshold as keywords.

3. Deep learning techniques

3.1 Transformer Framework

Transformer is a neural network architecture based on attention mechanism, with its core being self-attention mechanism, which requires the model to simultaneously consider relevant information from other positions in the input sequence.

Transformer based dense point annotation can be used to classify or annotate each vocabulary or token in the text, first collecting a text dataset with corresponding labels. Each text sample should contain vocabulary or sub words and corresponding labels. Preprocess the text to ensure that its format and quality are suitable for model training. Convert the preprocessed text dataset into the input and output representations of the Transformer model. The input is usually a text sequence, and the output is the label corresponding to each token. Using the architecture of Transformer as the model for the dense point annotation system, training the constructed Transformer model with prepared training data, calculating the difference between the model output and labels using cross entropy loss, and updating parameters through backpropagation. Infer on test data and annotate each token using a trained Transformer model. The model will output corresponding labels for each token. Perform post-processing on the predicted labels, such as mapping to specific label categories, decoding and interpreting the output of the model, to obtain the final dense point annotation result.

3.2 Bert Model

BERT is a pre trained language model based on the Transformer model[7].It has been pre trained on large-scale unlabeled text data and can then be used for fine-tuning various downstream natural language processing tasks. It learns bidirectional contextual representations in the text through mask language modeling in the pre training stage and next sentence prediction tasks.

In the next sentence prediction task, BERT receives a pair of sentences as input and determines whether they are adjacent sentences in the original text. For downstream tasks, the output of the BERT model can be used as input and adjusted according to specific tasks.

The advantage of Bert lies in its ability to capture long-distance contextual information, allowing the model to understand the semantics and context of language better through its bidirectional nature. It can be used to assist in classification systems, automatically classifying and marking the confidentiality level of files based on text content. Firstly, collect file datasets with confidentiality level labels. Each file should contain textual content and corresponding confidentiality level labels. Using a pre trained BERT model, use the BERT model as the basic model for auxiliary classification systems. Use prepared training data to fine tune the BERT model. Input the text into the BERT model to obtain the representation vector of the text. Then input the representation vector into the subsequent fully connected layers for classification prediction.

Train on the training data using the fine-tuned BERT model. Usually, the cross entropy loss function is used to calculate the difference between model prediction results and labels, and parameter updates are performed through backpropagation[7]。Infer on the test data and use the trained BERT model for confidentiality level prediction. For each input text, the model will output a corresponding confidentiality level label. Post process the predicted confidentiality level labels, such as mapping them to specific confidentiality level labels, transforming and interpreting the output of the model, to obtain the final classification result.

4. Systems Design

Based on the actual work situation of classified personnel, combined with information technology and artificial intelligence technology, the following system design scheme is proposed in figure 1.

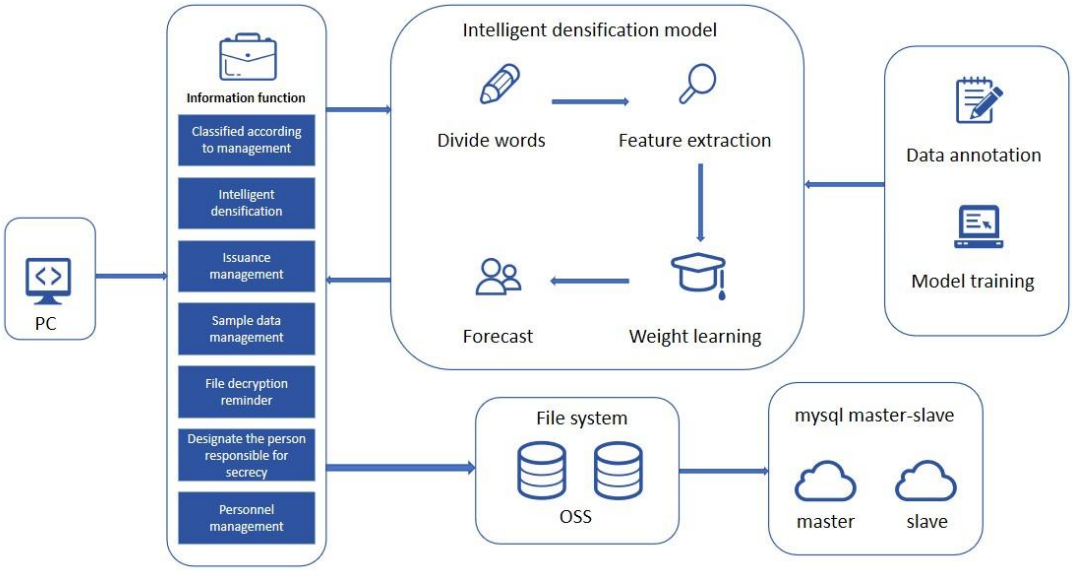

Figure 1. System Design

(1)Informatization of classified business: Standardize the process of classified business and control the process of electronic document classification, change, and release. The specific functions are as follows:

Classification basis management: Import standardized classification basis into the system, and add, modify, delete classification basis operations, as well as import and export corresponding classification basis.

Intelligent classification: The classification personnel upload the files that need to be classified to the system and set relevant attributes such as but not limited to document value, importance, disclosure scope, leakage impact, etc. The system reads the files and parses them to call the intelligent classification model to predict the classification points and levels.

Document issuance management: For documents that have completed intelligent classification, the issuance function can label the file classification level and encryption point.

File Decryption Management: Establish a historical database of classification management, which can quickly query and access classification files, as well as view information such as classification responsible persons and classification levels. It also provides decryption reminder function

Classification sample data management: Based on the classification criteria, correctly label the classification of various documents, and ensure that the sample classification distribution conforms to the overall data classification distribution. Randomly select a large number of documents as model training samples for classification annotation. After the model is trained, it predicts the confidentiality level of other documents through the model. At the same time, if there is any deviation from the predicted data, this module is used to correct it and finally form new sample data, completing the closed-loop process.

Management of classification responsible persons: setting and adjusting the classification

responsible persons of government agencies and units

Personnel management: The system administrator is mainly responsible for the daily operation and maintenance of the system, the security and confidentiality administrator is mainly responsible for the daily security and confidentiality management of the system, and the security auditor is mainly responsible for querying and analyzing the operation behavior logs of the system administrator and security and confidentiality personnel.

(2)Intelligent classification business: Utilizing deep learning technology to intelligently analyze file content, automatically identify and label classification points, and match corresponding classification levels and classification criteria

Collect file datasets with confidentiality level labels.

Each file should contain text content that requires auxiliary classification and corresponding confidentiality level labels.

Preprocess the file to ensure that its format and quality are suitable for processing. Use Jieba segmentation technology to segment text and convert it into a sequence of words. Use HMM to model the segmented word sequence and estimate the implicit state corresponding to each word. These implicit states can indicate the level of confidentiality of words.

Train a word vector model using the Word2Vec framework and CBOW model to convert the segmented word sequence into a vector representation. Calculate the importance of each word in the text set using the TF-IDF algorithm [8].

Input the segmented text sequence into Transformer and BERT pre trained language models to obtain the representation vector of the text. Using a preprocessed file dataset, use representation vectors as input features and confidentiality level labels as output for model training and fine-tuning.

Use the trained model to predict and infer on the test dataset, and obtain the predicted confidentiality level of the text. Post process, map, and interpret the predicted confidentiality level labels, and convert the model's output into the actual confidentiality level labels.

References

[1] Lu Xiao. Research and Implementation of Chinese Word Segmentation Technology Based on Conditional Random Fields[D]. Huazhong University of Science and Technology, 2011.

[2] Sheng Wuping. Research on Automatic Text Classification Based on Machine Learning[D]. East China Jiaotong University, 2021.

[3] Guo Hongqi, Li Guojia. A Sentence Similarity Calculation Method Based on Word Multi Prototype Vector Representation[J] Intelligent Computers and Applications, 2018, 8 (02): 38-42.

[4] Jin Yilin, Hu Feng. Research on Chinese Text Classification Algorithm Based On Three Branch Decision-making[J]. Journal of Nanjing University (Natural Science), 2018, 54 (04): 794-803.

[5] Luo Ling, Li Shuokai, He Qing, Yang Chengqi, Wang Yuyang Heng, Chen Tianyu. A Winter Olympics Knowledge Q&A System Based on Knowledge Graph, TF-IDF, and BERT Models[J]. Journal of Intelligent Systems, 2021, 16 (04): 819-826.

[6] Pan Wei. Research on the Discovery and Vectorization of New Words in Professional Fields[D]. Shandong University, 2021.

[7] Tan Junjie. Research on Short Text Classification Algorithm Based on Graph Model[D]. University of Electronic Science and Technology, 2021.

[8] Wang Boru, Fan Jing, Zhang Wangce, Li Chenguang, Ni Min. News Title Classification Based on Deep Neural Decision Forest[J]. Journal of Yunnan University for Nationalities (Natural Science Edition), 2020, 29 (05): 472-479.

