Similarity analysis method of power unstructured text based on multi-dimensional fusion feature extraction

Li Yongle*[1], Chen Jiaqi[1], Liu Yang[1], Sheng Shuang[1], Zheng Ling[2], Chen Fei[2]

[1] Big Data Center of State Grid Corporation of China, Beijing, China
[2] North China Electric Power University, Beijing, China

* Corresponding author: lyl_d08@163.com

Abstract. Similarity analysis of power unstructured text is one of the most important tasks in power unstructured data management. This paper studies the feature extraction and similarity analysis of electric power unstructured text. A multi-dimensional fusion feature extraction-based similarity analysis method is proposed to capture the features of unstructured text with more keywords and strong professionalism in electric power. This method improves the MatchPyramid model. In the input layer, word vectors generated by BERT model are used to strengthen the relationship between semantics. In the matching layer, the matching matrix between texts is constructed according to word vectors. In the feature extraction layer, the multi-word feature vectors extracted by BERT model are extracted in a multidimensional text fusion feature vector by means of dense connection. The higher-order feature vector of unstructured text is obtained. The higher-order features of different unstructured texts were input into Pseudo Siamese Network for similarity analysis. This method improves the semantic feature extraction ability and similarity analysis accuracy of unstructured text. Experiments show that compared with the traditional MatchPyramid model, the proposed method improves the feature extraction accuracy of unstructured text by 2.66% and $F1$ value by 2.99%.

Keywords: Deep metric learning, Text feature extraction, MatchPyramid, Pseudo Siamese Network

1 Introduction

Text feature extraction is a process of extracting and presenting textual information, and is the basis for extensive text processing [1,2]. The aim is to transform textual information into numerical or vector form, which can further be processed by machine learning algorithms, thereby enhancing model performance and diminishing dimensionality. This is highly important to make algorithms more comprehensive and to extracted essential data from text for decision-making and problem-solving.

Researchers have conducted a series of explorations and researches on text feature extraction. Liu Qingquan's research invention uses training suppress text and gain text to effectively correct the weights calculated by TFIDF algorithm, thereby improving its

X. Ding et al. (eds.), *Proceedings of the 2023 4th International Conference on Big Data and Social Sciences (ICBDSS 2023)*, Atlantis Highlights in Social Sciences, Education and Humanities 12,
https://doi.org/10.2991/978-94-6463-276-7_11

processing ability for longer text. It can effectively detect and identify specific strings, so as to achieve better segmentation effect [3]. Ma Li et al. improved the commonly used feature extraction algorithm, that is, TFIDF algorithm, and proposed LDA to realize text modeling. By using Gaussian weighting, they solved the problem that when encountering a feature word with low frequency but representative significance, it could be regarded as an important feature word category. But there is still the problem of sparse data [4]. Researcher Dingli Liu and his team have developed a new text clustering algorithm, which uses the Word2vec model to identify and compare specific specific features more finely, greatly reducing the sparsity and specificity of text features, and more effectively identifying and comparing specific specific features [5].

With the development of technology, the written records of the electricity industry have become more complex and contain a lot of different information. Ai Sheng Liu et al. proposed a power single text classification model based on XLNet with multiple layers of enhanced attention, which solved the classification of automatic power customer service types and effectively extracted different categories of work orders. However, there are still some problems such as single category of power text recognition [6]. This paper proposes a multi-dimensional fusion feature extraction-based electric unstructured text similarity analysis method, which resolves the issue of distinguishing multiple categories and text similarity in electric text feature extraction, thereby enhancing the precision of electric text feature extraction.

2 Related theories and models

2.1 MatchPyramid model

MatchPyramid model is a text matching model in deep metric learning [7], which can be used to solve various text matching problems, such as question and answer matching and short text similarity calculation.

The MatchPyramid model mainly consists of two parts: feature extraction and matching calculation. It can effectively extract useful information from multiple inputs to generate two independent matrices. Then their similarity matrix is calculated, and then the similarity matrix is processed by convolutional neural network to get the final matching score.

2.2 Pseudo Siamese Network Model

The Siamese network is a structure which divides input content into two distinct forms: the Siamese network and the pseudo Siamese network [8]. The pseudo Siamese network model is a text matching model based on the Siamese network, but it does not necessitate two identical networks; instead, by sharing the network layer, only one network is utilized to achieve text matching. The specific structure is shared network layer, left text network layer and right text network layer. By constantly adjusting network parameters, the pseudo Siamese network model gradually learns the semantic information of the text and can accurately match the text in the test.The speed of training, along

with the limited number of model parameters, make it an ideal choice for text matching tasks with a limited corpus size [9].

3 Similarity analysis of power unstructured text based on multi-dimensional fusion feature extraction

Electric power unstructured text has the characteristics of more keywords and strong professionalism. Existing text feature extraction methods have low feature extraction accuracy for electric unstructured text. This paper aims to solve this problem based on the improved MatchPyramid model and the Pseudo Siamese Network.

A thorough description of the power unstructured text similarity analysis model structure, which is based on multi-dimensional fusion feature extraction, is provided. The model can be divided into four sub-levels: input layer, matching layer, feature extraction layer and output layer. The structure can be seen in Figure 1.

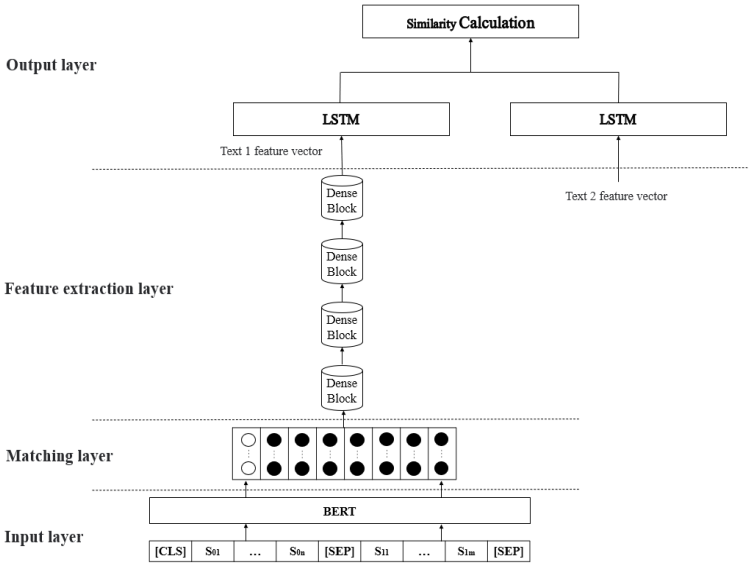

Fig. 1. Similarity analysis model of power unstructured power text based on multi-dimensional fusion feature extraction

3.1 Input layer

In the input layer, you first need to convert the text into a word vector. Our use of the Jieba word segmentation tool can effectively remove unnecessary words. Secondly, Glove can convert words into corresponding word vectors. Finally, with the BERT model, two paragraphs of text can be converted into an ordered vector and converted into a form that can be recognized by the MatchPyramid algorithm.

3.2 Matching layer

Word vectors processed by BERT model can be used to construct similarity matching matrix between words in two paragraphs of text. First, we need to multiply the word vectors in the two paragraphs of text to form a matrix and use the convolutional neural network to process it to get a matching score matrix. Where m_i and n_j represent the i and j words in sentences S1 and S2 respectively, and M_{ij} is used to represent this matrix, the specific formula is shown in (1).

$$M_{ij} = \frac{(m_i)^T \cdot n_j}{\|m_i\| \times \|n_j\|} \tag{1}$$

3.3 Feature extraction layer

The matching matrix formed by two text word vectors is a two-dimensional graph formed by dot product. Each element of the matrix represents the similarity of two word vectors, calculated by cosine similarity. This layer is composed of 4 Denseblocks, which are normalized network layer, convolutional layer, activation layer and pooling layer, which can effectively realize effective data sorting and processing.

For each batch of data, the normalized network layer is precisely calculated. First, the average value of each batch of data is calculated according to formula (2), where m represents the batch size:

$$\mu = \frac{1}{m}\sum_1^m x_i \tag{2}$$

By applying formula (3), we can calculate the variance of each set of data:

$$\delta^2 = \frac{1}{m}\sum_1^m (x_i - \mu)^2 \tag{3}$$

After normalization, we can use formula (4) to turn all the data into a normal distribution with 0 and 1.

$$\widehat{x_i} = \frac{x_i - \mu}{\sqrt{\delta^2 + \varepsilon}} \tag{4}$$

In order to capture and maintain effective features effectively, the normalization algorithm adopts an improved scheme. That is, by introducing transformations to improve the performance of the model, for example, by introducing (5), we can make the performance of the model better reflect the performance of the model.

$$y_i = \alpha \widehat{x_i} + \beta \tag{5}$$

The feature extraction section of this paper uses a number of different methods, and their inputs include many types of information. For example, l represents the sum of multiple pieces of information, while xl refers to the sum of individual pieces of information. For example, we use H_l, which can classify a variety of information and process it with appropriate methods.

$$x_l = H_l([x_1, x_2, \cdots, x_{l-1}]) \tag{6}$$

Through convolution and pooling operations, the MatchPyramid model can extract local and global features of different sizes from the matching matrix and combine them together to form a higher-level representation of the features.

3.4 Output layer

After obtaining the higher-order features of the output of the previous layer, they are input into two subnetworks, the long short-term memory network. They consist of several complete linking layers and one active function layer. The features are fused and analyzed, and the cosine approximation is used to measure their internal similarity, and then the similarity of the two texts is evaluated.

4 Experimental results and analysis

4.1 Collection of datasets

The datasets used in this paper can obtain public power text data sets through search engines or data warehouses. These data sets may be collected from real power systems, sensors, monitoring devices, etc. Synthesizing for particular objectives, like fault detection and load prediction, is a common practice.

According to the purpose of the text, the power text is divided into the following seven categories: power equipment and facilities (hereinafter referred to as "equipment"), power monitoring data text (hereinafter referred to as "monitoring"), power industry news and report text (hereinafter referred to as "news"), power policies and rules and regulations text (hereinafter referred to as "policies"). Power Industry Knowledge Base and encyclopedia text (hereinafter referred to as "Encyclopedia"), power industry social media and forum text (hereinafter referred to as "media"), power industry user feedback and complaint text (hereinafter referred to as "Complaint").

This dataset consists of several fields that not only cover the encoding of each text pair, but also provide the encoding used to distinguish their different meanings, with 1 stands for they have exactly the same meaning, and 0 stands for they have no meaning at all. A total of 500,000 texts were divided into three datasets based on a ratio of 0.85:0.10:0.05: training set, validation set and validation set. We cut out some irrelevant information and limited the minimum length of each paragraph to 40 pinyin characters. If the length of the paragraph is less than or equal to 40 pinyin characters, we will choose one of them.

4.2 Experimental environment

In this study, we used a 320-dimensional Glove vector for word embedding and Adam for optimization. In the hyperparameter adjustment, we set the learning rate to 0.001, batch_size to 64, and epoch to 20. This article uses pytorch2.0.0 and python3.8.1 as experimental environments.

4.3 Analysis of experimental results

To test the validity of this model, we conducted three experiments.

In the first experiment, the accuracy and F1 values of this model were compared with those of existing text models. By introducing BERT model, this model can capture the semantic information of text better, thus greatly enhancing the accuracy of text processing. Using high-density convolutional neural networks, we are better able to capture the details of the information and better able to deal with the complexity of the information. Through deep data mining of Pseudo Siamese Network, we have obtained a large amount of advanced information. The accuracy and F1 values of this model are better than those of other models. The specific results are shown in Table 1.

Table 1. Comparison of benchmark models

Model	Accuracy	*F*1
TF-IDF	79.11%	80.23%
Siamese Network	82.45%	83.73%
MatchPyramid	84.23%	82.91%
Improved MatchPyramid	86.89%	85.90%

The second experiment's primary aim was to investigate various algorithms for obtaining more precise data. To accomplish this, TF-IDF, Information gain (IG), CHI and mutual information algorithms were employed to draw out valuable information from the data set, and clustering algorithms were employed to delve into these facts more deeply. Then, we use MatchPyramid algorithm to extract valuable information from the original information set, and use clustering algorithm to dig these information deeply, so as to obtain more detailed information. Figure 2-4 shows the performance, recall results and F1 value of different feature extraction algorithms in the cluster respectively. By means of comparison, we discovered that employing deep neural network technology to recognize power industry data is highly beneficial [10].

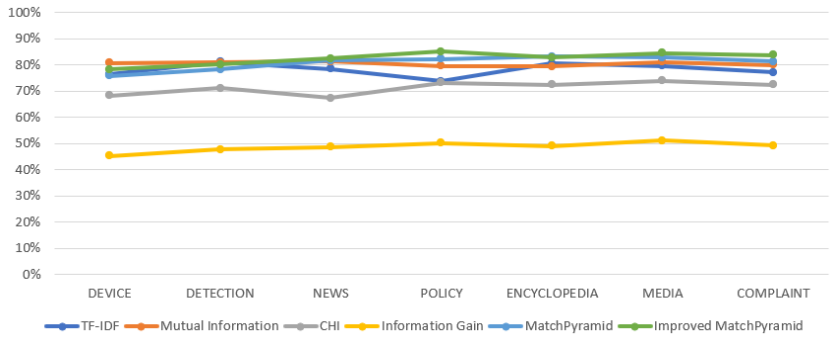

Fig. 2. Performance of different feature extraction algorithms in clustering

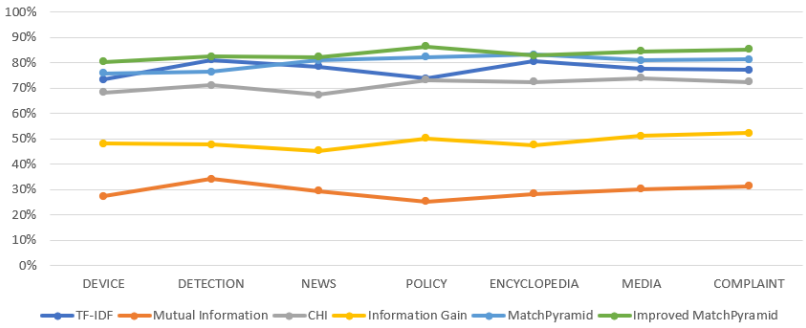

Fig. 3. The result of clustering recall by different feature extraction algorithms

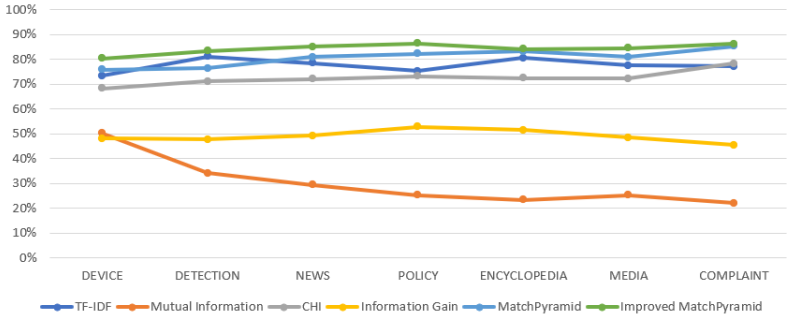

Fig. 4. The clustering F1 value of the feature extraction algorithm

In the third experiment, 40 different types of power text were randomly selected in the public power text dataset, and the categories of each group of power text were manually determined. The 40 power texts were divided into 20 text groups, that is, 20 experimental batches. The preprocessed text is input into the MatchPyramid model to learn the similarity features between the two texts. The extracted text features are matched and the similarity between them is calculated. According to the similarity calculation result, the output matching result determines whether it is the same category. Table 2 shows that multiple sets of non-identical power texts are input into the model to obtain the results of similarity and whether they are of the same category.

Table 2. Power text similarity discriminations

Group	Electricity Text Type	Similarity	Determine if it is the same category
1	Equipment + Detection	74.32%	No
2	Complaints + News	78.21%	No
3	Encyclopedia + Encyclopedia	83.73%	Yes
4	Equipment + Encyclopedia	57.89%	No
5	Policy + News	81.23%	Yes

6	Media + Media	85.89%	Yes
7	Encyclopedia + Complaints	53.45%	No
8	Detection + Detection	84.79%	Yes
9	Complaints + Media	82.12%	Yes
10	News + Media	85.19%	Yes
11	Encyclopedia + News	79.86%	No
12	Equipment + Equipment	76.89%	No
13	Detection + Policy	45.39%	No
14	Media + News	83.27%	Yes
15	Policy + Policoes	84.67%	Yes
16	News + Detection	76.49%	No
17	Media + Detection	69.32%	Yes
18	Complaints + Complaints	89.09%	Yes
19	Complaints + News	84.54%	Yes
20	Equipment + Policy	68.34%	No

According to the experimental results, batches 3, 6, 8, 10, 15 and 18 are identified as the same category and the similarity is 80%, indicating that the model has a good performance. Batches 5, 9, 14, 17 and 19 are identified as the same category, but in fact they are not the same category, so there is an overfitting situation when the model extracts features. Therefore, it can be concluded that when the similarity exceeds 80%, the category of the power text combination is the same, and even though the similarity of the group is less than 80%, it is still the same category of text. In conclusion, the improved MatchPyramid text matching accuracy is higher.

5 Conclusion

A multi-dimensional fusion feature extraction-based MatchPyramid model is presented in this paper for power unstructudred text similarity analysis. By using BERT model to preprocess text, word vector is generated, and MatchPyramid model is used to generate matching matrix and feature extraction. Intensive connections were implemented in the feature extraction layer to avert the loss of effective features at the base layer. The extracted higher-order features were input to Pseudo Siamese Network to obtain text similarity scores. Experiments proved that our model performed well in processing power text similarity problems. The accuracy of forecasting the likeness of the authoritative text can be determined. This method has good plasticity and extensibility, and can provide strong support and guidance for the task of power text similarity calculation.

Funding project

This work was financially supported by Science and Technology Project from Big Data Center of State Grid Corporation of China (Contract No: SGSJ0000YYJS2310054)

References

1. Zeng M, Yuan M, et al. Research progress in text feature extraction [J] . Journal of Nanjing University of Information Science and Technology Science, 2019,11(06):706-715. DOI: 10.13878/j.cnki.jnuist. 2019.06.008.
2. Research and application of deep learning in text feature extraction [D] . Xi'an Shiyou University, 2021. DOI: 10.27400/d.cnki.gxasc. 2021.000802.
3. Liu Q. Application of improved TFIDF algorithm in text analysis [D] . Nanchang University,2019.DOI:10.27232/d.cnki.gnchu.2019.000283.
4. Ma L, Liu H. An improved text feature extraction algorithm [J] . Journal of Xi'an University of Posts and telecommunications, 2015,20(06) : 79-81 + 120. DOI: 10.13682/j.issn. 2095-6533.2015.06.017.
5. Liu D. Research on clustering algorithm of text feature extraction based on deep learning [D]. Guilin University of Electronic Technology, 2021. DOI: 10.27049/d. cnki.ggldc.2021.000866.
6. Liu A, Ou W, et al. Power Text Classification Based on XLNet with Multi-Layer Enhanced Attention[J]. Automation and Instrumentation, 2023, 38(05): 1-4+27. DOI:10.19557/j.cnki.1001-9944.2023.05.001.
7. Hou Y. Short text classification based on attention mechanism [J] . Computer Literacy and technology, 2020,16(28) : 185-186 + 201. DOI: 10.14004//j.cnki.ckt. 2020.3221.
8. Li Yilin, Zhou Yanping. Text similarity matching based on Siamese network and combined word vector[J]. Computer Systems & Applications, 2022, 31(10): 295-302. DOI: 10.15888/j.cnki.csa.008756.
9. Xie T. Research on Text Semantic Matching Based on Dual Siamese Network [D]. Southwest University of Finance and Economics, 2022. DOI: 10.27412/d.cnki.gxncu.2022.001371.
10. Lei L. Research on Chinese text classification algorithm based on depth feature extraction [D]. Shanghai Jiao Tong University, 2020. DOI: 10.27307/d.cnki.gsjtu. 2020.001081.