Research on Automatic Labelling of Imbalanced Texts of Customer Complaints Based on Text Enhancement and Layer-By-Layer Semantic Matching

XIAOBO TANG
Wuhan University
HAO MOU (mouhao2020@163.com)
Wuhan University
JIANGNAN LIU
Wuhan University
Xin Du
Wuhan University

Research Article

Keywords: auto-labeling, BERT, text enhancement, Word2Vec

Posted Date: February 16th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-198986/v1

Research on automatic labelling of imbalanced texts of customer complaints based on text enhancement and layer-by-layer semantic matching

XIAOBO TANG[1,2] HAO MOU[2,3] JIANGNAN LIU[2] XIN DU[2]

[1]Center for Information System Research, Wuhan University
[2]School of Information Management, Wuhan University
[3]Sichuan Xichang Electric Power CO.,LTD.

Corresponding author: Hao Mou (e-mail: mouhoo2020@163.com).

Abstract: [Purpose/meaning] Due to its potential impact on business efficiency, automated customer complaint labeling and classification are of great importance for management decisions and business-level applications. The majority of the current research on automated labeling uses large and well-balanced datasets. However, customer complaints' labels are hierarchical in structure, with many labels at the lowest hierarchy level. Relying on lower-level labels leads to small and imbalanced samples, thus rendering the current automatic labeling practices not applicable to customer complaints. [Methodology/process] This article proposes an automatic labeling model incorporating the BERT and Word2Vec methods. The model is validated on electric utility customer complaints data. Within the model, the BERT method serves to obtain shallow-level text tags. Further, text enhancement is used to mitigate the problem of uneven samples that emerges when the number of labels is large. Finally, the Word2Vec model is utilized for the deep-level text analysis. [Findings/conclusions] The experiments demonstrate the proposed model's efficiency in automating customer complaint labeling. Consequently, the proposed model supports the enterprises in improving their service quality while simultaneously reducing labor costs.

Keywords: auto-labeling, BERT, text enhancement, Word2Vec

1 Introduction

As the quality of life improves, customers have higher expectations of the purchased products and received services. Major enterprises strive to remain aligned with customer interests by creating customer complaint channels that resolve customer disputes and dissatisfaction. In addition, with a rise in business competition, customer churn reduction becomes increasingly important. The critical aspect of reducing customer churn is customer satisfaction improvement. Thus, customer complaints, which reflect customer satisfaction, form an essential bridge between customers and companies.

Although the number of complaints is often large, enterprises commonly rely on manual methods for processing complicated complaint content. As a result, customer complaint processing requires significant labor and time. Since text mining enables automatic analysis of customer complaints, it can reduce labor costs, promote customer satisfaction, and prevent customer churn. Thus, the research on text mining of customer complaints can play a crucial role in improving enterprises' efficiency.

Automatic labeling and classification refer to the use of computational procedures that generate tags to summarize, describe, and classify the textual content. The current research on automatic

labeling of both short and long text mainly relies on keyword extraction and topic modeling. However, the derived labels do not have a hierarchical structure. The customer complaint texts, on the other hand, are processed in accordance with the complaint classification, where the responsible customer service department is contacted, customer complaints are processed, and the processing results are returned. Based on the level of admissibility, the complaint texts are usually divided into different business categories that can be pictured as having a tree-like structure. For example, "complaint service" may be divided into "service behavior" and "service channel" on the first hierarchy level. The "service channel" may be further divided into "electronic channel" and "business hall channel" to form a second hierarchy level.

Therefore, the customer complaint labels are characterized by a hierarchical structure, but also the excessive use of deep-level labels, a relatively small sample size, and an imbalanced sample size. When a traditional auto-labeling algorithm assigns a text to the lowest-level admissible business category, the text can be understood to automatically belong to the relevant higher-level admissible business category, as well. However, such an approach does not make effective use of hierarchical information between categories. Further, customer complaints can have numerous deep-level labels, resulting in small and imbalanced samples. Such samples pose an obstacle in training an auto-labeling model.

This paper addresses the described issues by proposing a BERT and Word2Vec-based automated customer complaint labeling model. BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language representation model and is viewed as a landmark work in the field of natural language processing (NLP). BERT makes the full use of a large number of unsupervised texts for self-supervised learning and encodes the linguistic knowledge. The experiments demonstrate its superior performance on various NLP tasks [1]. Word2Vec is a technique used to calculate word vectors [2]. Based on a given corpus and a training model, Word2Vec quickly and efficiently represents a word in the form of a vector that enables the calculation of word-to-word similarity.

The model presented in this paper enables the hierarchical classification of customer complaints. Within this work, class labels at the second hierarchical level (i.e., Level 2 of admissible business classes) are called shallow-level labels. In contrast, class labels for Level 3 admissible business classes are denoted deep-level labels. The model's automatic labeling of customer complaints is divided into two stages. In the first stage, the BERT classification algorithm is used to identify shallow-level labels. In the second stage, deep-level labels are determined by calculating the similarity between the text and labels to be matched. The class label with the highest similarity to the text is selected as the complaint label. Using a BERT-based classification for text labels at the shallow level and a similarity calculation model based on Word2Vec for text labels at the deep level. The contribution of this paper:

(1) An automatic customer complaint text indexing model based on BERT and Word2Vec is

proposed.

(2) A text enhancement method is proposed to improve the problem of imbalanced text and improve the accuracy of automatic indexing model.

(3) In the process of matching the text with the indexing label, the method of layer-by-layer semantic matching is adopted. First, which shallow-level label belongs to the text is determined, and then the deep-level label under the shallow-level label is used to match the text, rather than directly match the text with the deep-level label, which significantly improves the accuracy of the automatic indexing model.

2 Literature review

2.1 Auto-labeling and auto-indexing

As noted in Section 1, auto-labeling or auto-indexing refers to the process of automatically assigning labels or tags to text to express its content[3]. More precisely, automatic labeling is the process of extracting words and phrases directly from the original text to describe the document's subject matter.

Automatic indexing research began in 1957 when Luhn [4] introduced computer technology into bibliographic studies and devised a word frequency method based on Zipf's law. Earl [5] combined the syntactic analysis and word frequency statistical methods to extract keywords. The work of Salton et al. [6] proposes the application of the vector space model for auto-labeling, whereas Deerwester et al. [7] suggest the use of latent semantic analysis. Finally, Anjewierden and Kabel [8] proposed an ontology-based method for automatic labeling. Hugo [9]proposed an innovative method to deal with the complexity of events in medical event log. Based on automatic labeling, similar events are clustered in potential space to create accurate label. Su [10]proposed an automatic evaluation and labeling architecture of product perceptual attributes based on convolutional neural network.

2.2 Imbalanced Text

Imbalanced text is represented by a significantly lower number of elements in one category than in other categories. An important problem facing natural language processing and machine learning at this stage remains the efficient processing of imbalanced text in classification tasks. Major NLP tasks including sentiment analysis, propaganda detection and event extraction from social media are all examples of imbalanced classification problems.

S[11] proposes a hybrid imbalanced data learning framework (HIDLF) to deal with the imbalance of views in the movie review dataset, and then classifies the movie reviews by the proposed HIDLT-SVM algorithm. Harish[12] proposed the the BERT model to deal with the problem of data imbalance in text classification. Li[13] proposes a solution to the imbalanced text problem in a multi-classification task. The multi-class dataset is first decomposed into several binary

class datasets. It then uses spectral clustering to divide a minority of the subset of binary categories into subspaces and oversamples them according to the characteristics of the data. Spectral clustering-based sampling takes into account the distribution of the data and effectively avoids oversampling of outliers. After the data has approximately reached the equilibrium point, a multiclass classifier can be trained from this rebalanced data.

3 BERT andWord2Vec-based model for automatic labeling of customer complaint

Fig. 1 shows the proposed model for automatic customer complaint labeling based on BERT and Word2Vec. The model consists of four automatic text classification stages: data pre-processing, BERT-based text classification, Word2Vec-based semantic similarity matching, and label confirmation. BERT classification is used to validate shallow-level labeling, and Word2Vec semantic similarity matching serves for deep-level labeling. The model first confirms the shallow-level labels of the text to be indexed by using the BERT classification model; Then according to the hierarchical label set, find the deep level label under the shallow level label ; Finally, Word2Vec semantic similarity matching model is used to match the deep-level labels of the text to be indexed.

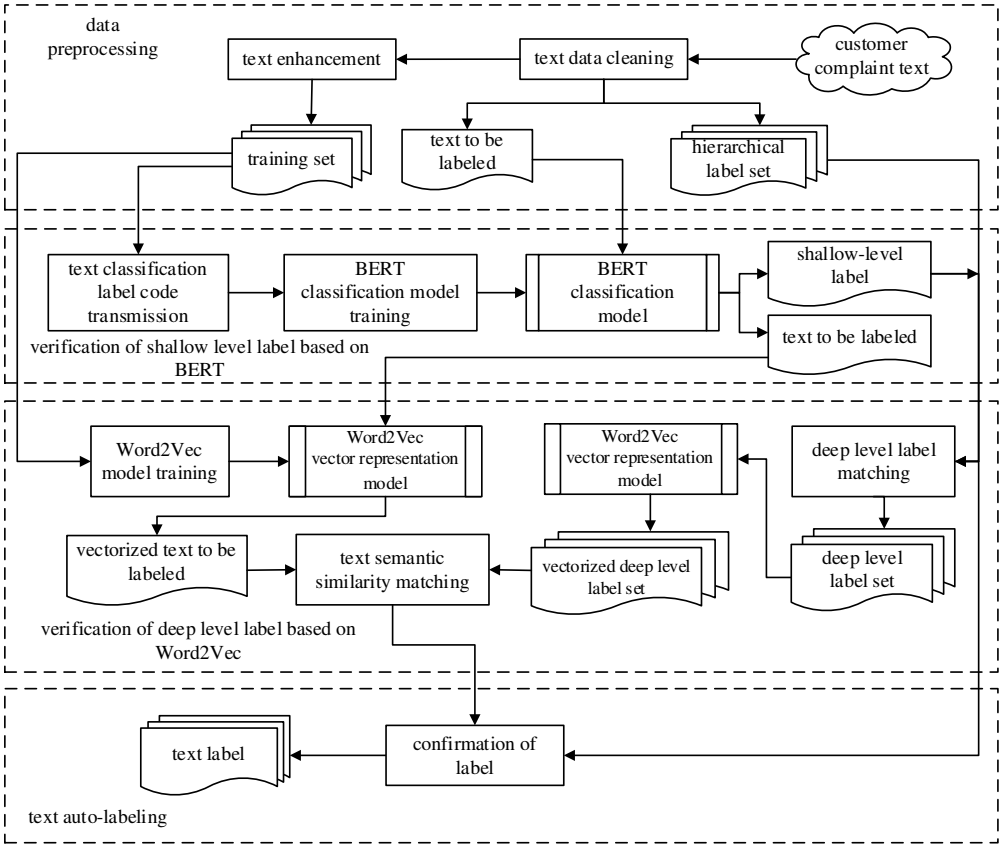

Figure 1 BERT and Word2Vec-based automatic text labeling model for customer complaints

3.1 Data pre-processing

Data pre-processing is a series of normative operations performed on the acquired irregular and chaotic texts. The goal of pre-processing is to reduce the irregularities, remove inaccuracies,

and reduce noise. This step serves as a basis for the subsequent text manipulation. Within this work, data pre-processing consists of the following steps:

(1) Data acquisition : Through consultation and cooperation with the power company, the customer complaint data by the power company background is obtained, which includes the customer complaint text and the corresponding tag.

Among them, the label is divided into the first-level label, the second-level label and the third-level label. For example, a customer complaint text: “The customer asked a staff member to deal with the power failure at home, but the staff member did not deal with it and did not explain the reason to the customer.”, The text belongs to the "service attitude of other personnel" (third-level label) under "service behavior" (secondary label) under "service complaint" (first-level label).

(2) Data cleaning : data cleaning mainly includes text filtering and hierarchical label set extraction.

Text filtering is mainly to manually filter the customer complaint text. Due to the freedom of customer complaints, there are some meaningless noise data, so it is necessary to rely on manual screening of the obtained data.

The extraction of hierarchical label set is to organize the indexing labels of text data to form a hierarchical label set.

(3) Text enhancement: The (too) deep-level labels lead to a few samples in several categories. Thus, the classification process commonly deals with imbalanced samples that affect classification accuracy. Within this work, this problem is tackled by enhancing the texts falling in critical categories.

Oversampling is the process of generating a bias towards selecting the data from specific categories. In other words, it is a data enhancement technique that increases the likelihood of choosing a positive case, thus overcoming the issues stemming from the imbalance of positive and negative dataset sizes. However, the adjusted data set will often be much larger than the original data set, increasing the time overhead.

The main methods of data enhancement in NLP are adding noise and paraphrasing. Noise addition modifies a positive case by, for example, randomly deleting certain words or disrupting the word order to generate new data. Paraphrasing can be seen as a seq2seq task. For example, in the question and answer systems, question retelling deals with developing a better question format from the original one. Then, the new question is used instead of the old question in the question and answer system.

Images commonly add noise to several pixels, which often does not have a significant impact. However, in the study at hand, a small change in a text may greatly impact the task since the deleted words are likely important. Removing an unimportant word, on the other hand, is equivalent to

deleting a stop word. Disrupting many words' order is even more infeasible as sentence disruptions remove the contextual relationship among words. Thus, text generation is seen as a more effective way to enhance the texts within this work.

In this paper, data enhancement is divided into two steps. The first step extracts the texts from underrepresented categories and translates them. The texts are translated from Chinese to English and back. The newly obtained Chinese texts contain different expressions and can be saved as new texts under the category tag. This procedure doubles the number of complaints in the category. The second step – synonym substitution - has to be done to increase the number of texts even further.

The second step uses the partitioning tool in the NLP package pyltp and the Harbin Institute of Technology（HIT）synonym thesaurus. First, a synonym dictionary is traversed, and a synonym list is constructed for each word in the sentence. For example, had the word "pay" existed in a sentence, the list of synonyms might contain the words "compensate, give, refund, repay, pay off…". One restriction was introduced. Namely, the synonym replacement was not performed for very short words since they are commonly adverbs or prepositions (e.g., "of" or "in"), for which synonym replacement is not very meaningful. Next, the three top-ranking synonyms are selected as a final result for each word in the synonym list. If a word has less than three synonyms (e.g., "customer"), the exact word is copied three times (the synonym list becomes "customer, customer, customer"). In this manner, each word in the sentence is associated with three words, and a simultaneous replacement of multiple words in the sentence generates the new sentence. Since the synonym dictionary is not comprehensive, and short words are not replaced, the original sentence parts are retained, while the others are replaced with synonyms. For example, these three instances are obtained from a single sentence:

- The customer paid 400 RMB at the State Grid Pengshan Power Supply Office this morning.
- The customer refunded 400 RMB at the State Grid Pengshan Power Supply Office before noon.
- The customer repaid 400 RMB at the State Grid Pingshan Power Supply Office half a day before.

Following the described synonym substitution procedure, each original sentence could be expanded into three sentences. If additional data is required, more sentences can be generated by utilizing longer synonym lists. Substitution by synonyms introduces a level of randomness in the sentences, forcing the model to learn deeper levels of semantic information and, consequently, enhancing the model robustness and reducing the influencing data factors. Using the HIT synonym dictionary has – at least - two advantages. First, using the dictionary eliminates the need to calculate similarity and, thus, avoids the inaccuracies in calculating the Word2Vec similarity caused by scarce

customer complaints data. Second, HGMU synonym dictionary does not distinguish specific domains. Several words in the electricity field may not have synonyms such as, for example, "State Grid Pengshan Power Supply Business Hall". Using the dictionary-based synonyms' replacement can, to a certain extent, alleviate the emergence of fluency problems.

3.2 Verification of shallow level Label Based on BERT

Text belongs to unstructured data. To rely on the computer to classify the text, it must be transformed into computer-comprehensible structured data. Therefore, text representation is crucial for natural language processing tasks. Traditional word vector representation models include TF-IDF model,word2vec neural network language model and so on. TF-IDF performs vector representation of the text based on the product of term frequency TF and inverse document frequency IDF; Word2vec maps words to low dimension and high density vector space by training neural network model. However, these text representation models have such problems : they do not consider the meaning of words in the context. Take the word “book” for example, in the sentence “I am reading a good book on economics”, “book” means a written work or composition that has been published. And in the sentence “The agent booked tickets to the show for the whole family”, “book” means arrange for and reserve (something for someone else) in advance. When using the traditional word vector to represent the model, it cannot effectively combine context and distinguish semantics, which is obviously unreasonable. To solve this problem, this paper uses BERT model for text representation.

In recent years, several studies successfully mitigated the listed problems by pre-training deep neural network models and fine-tuning them to perform specific NLP tasks. BERT is one such pre-trained language model with output vectors, a word-level vector, and a sentence-level vector. Sentence-level vectors can capture the entire sentence's semantics and are often used in classification tasks.

In this paper, a BERT-based text classification model is proposed. Every BERT sequence starts with a special classification token denoted CLS. The model first takes a CLS vector representing the sentence, and then the sequence is passed to a fully connected layer. The model uses binary cross-entropy loss function and the softmax function, an activation function typically used in multi-classification tasks.

The specific structure of the model is shown in Fig. 2. In the figure, $Tok_1, Tok_2, \dots, Tok_N$ is the input vector at the word level, C is the input vector at the sentence level, $T_1, T_2, \dots, T_N$ is the BERT model output vector at the word level, and CLS is the BERT model output vector at the sentence level.

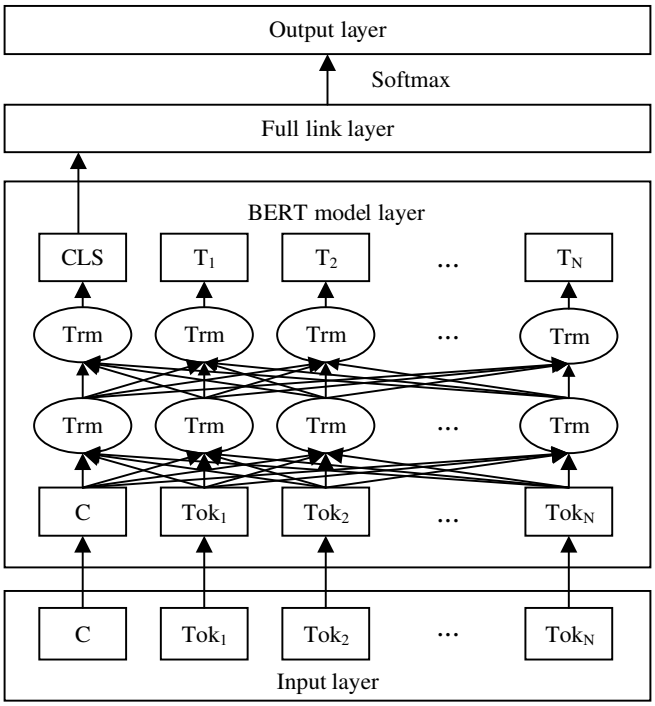

Figure 2 BERT-based text classification model

To enable text classification using the model, the class labels must first be encoded and transformed. The class labels of each sample are mapped to a list $[l_1, l_2, \ldots, l_N]$, where N is the total number of class labels and $l_i \in \{0,1\}, \forall i = 1, 2, \ldots N$. When $l_i = 1$, the sample obtains class label l_i.

Once the text is represented using BERT, the CLS vector is extracted and passed to the full link layer. The Softmax function normalizes the output nodes' values so that the sum of the output nodes equals 1. The Softmax function is defined as:

$$soft\max(Z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}.$$

After using text enhancement to solve the problem of category imbalance, this paper uses a BERT-based text classification model to classify customer complaint texts, and uses the classification label as the text indexing label of the text at the shallow level.

The next step is to find the deep-level label under the shallow-level label based on the shallow-level label obtained by the BERT-based text classification model and the level-level label set obtained by data cleaning, and use the text semantics based on Word2Vec Similarity matching is used to determine the deep-level tags.

3.3 Verification of Deep Level Label Based on Word2Vec

At present, there are many studies on question similarity matching, which can be roughly divided into question similarity research based on string, question similarity research based on vector space model and question similarity research based on deep learning. The focus of this paper is not to propose a question similarity matching model with high accuracy or innovation, but to

propose a set of suitable for this problem, and effective solutions, so this paper uses the commonly used question similarity calculation method: based on Word2Vec text vector representation, the cosine similarity of the calculation vector as the text semantic similarity.

Word2Vec uses two models, the Continuous Bag-of-Words model (CBOW) and the Continuous Skip-gram model. The goal of CBOW is to predict the probability of a word based on the current context, while Skip-Gram does the opposite, determines the probability of a context based on the presented word (Fig. 3). Both models use artificial neural networks as their classification algorithms. Initially, each word is a random *n*-dimensional vector, but - upon training - the models obtain an optimal vector for each word.

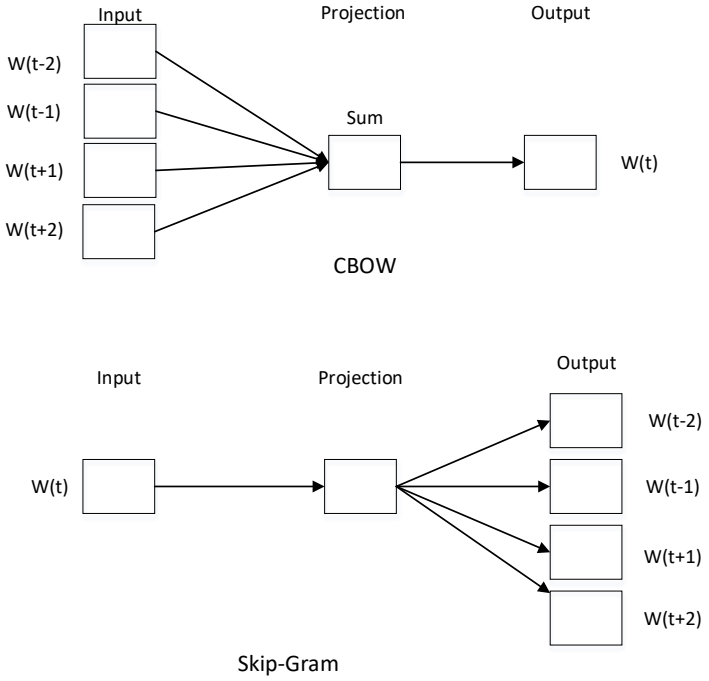

Figure 3 Two models of Word2Vec

Word2Vec trains both CBOW and Skip-Gram models on a given corpus, and the output yields vector representations of all the words that appear in the corpus. The obtained word vectors enable the calculation of word-word relationships, such as word similarity and semantic relatedness.

This paper uses training set text data to train Word2Vec text vector representation model; Through the shallow level label obtained by the previous text classification model based on BERT and the hierarchical label set obtained by data cleaning, the deep level label under the shallow level label is found. Word2Vec vector representation of indexing text and deep level label respectively; By calculating the cosine similarity between vectors, the deep-level label to be matched with the highest similarity is taken as the deep-level label of the text.

3.4 Text auto-labeling

The last phase of the auto-labeling process deals with determining the final label from shallow-level and deep-level text labels. At the shallow-level, the label is validated using the text-enhanced

BERT classification model. Similarly, at the deep-level, validation relies on Word2Vec to calculate the semantic similarity between the text and the potential labels. The final text's label is determined by combining the shallow-level and the deep-level citation labels.

4 Experiments and analysis of results

4.1 Data validation

The power company provided experimental data in an anonymized form to protect the privacy of its customers. The experimental data consist of the electricity complaint text and the primary, secondary, and tertiary admissibility business categories to which the text belongs. The experiments reported in this paper were performed on the first- and second-level categories. The data was cleaned to ensure the accuracy of the experiment. After the initial manual screening and the removal of noisy data, a total of 16 shallow-level labels corresponding to 3438 complaints were selected as the experimental dataset. The distribution of the experimental dataset is shown in Table 1.

Table 1 Distribution of the samples by the second-level categories

Serial number	Category Tags	Number of texts containing the tag
1	Electricity construction	225
2	Electricity supply facilities	52
3	Voltage quality	557
4	Power supply reliability	1434
5	Power supply frequency	11
6	Acts of service	403
7	Service Channels	29
8	Repair service	143
9	Power outage issues	58
10	Power Outage Information Bulletin	51
11	meter reading and reminder	86
12	Electricity tariffs	22
13	power metering	138
14	Business expansion report	138
15	Operating charges	34
16	Change of electricity consumption	57

4.2 Experimental procedures

As described in Section 3.1, data pre-processing includes the text enhancement step aimed at supporting the subsequent BERT classification. The categories with a small number of samples are

detected and the texts within these categories are enhanced. This step resulted in the dataset containing 3698 complaints, and the data distribution is shown in Table 2.

As described, the BERT-based text classification algorithm is first used to determine the shallow-level label for every complaint. The data were divided into training and test set in a 9:1 ratio, yielding a test set of 369 complaints. Within the BERT classification model, the initial learning rate was set to 0.00001. Further, Adam optimizer was utilized, and the Epoch was set to 2. For deep-level labels, the Word2Vec model was trained on an enhanced dataset with dimensionality set to 100, and the resulting Word2Vec model was used for semantic similarity matching.

Table 2 Distribution of the complaints after the text enhancement

Serial number	Category Tags	Number of texts containing the tag
1	Electricity construction	225
2	Electricity supply facilities	87
3	Voltage quality	557
4	Power supply reliability	1434
5	Power supply frequency	20
6	Acts of service	403
7	Service Channels	47
8	Repair service	143
9	Power outage issues	101
10	Power Outage Information Bulletin	86
11	meter reading and reminder	152
12	Electricity tariffs	34
13	power metering	138
14	Business expansion report	138
15	Operating charges	63
16	Change of electricity consumption	70

The model's performance was evaluated using accuracy, precision, recall, and F1-score measures. The measures are defined below, and TP denotes "true positives" (i.e., the number of texts correctly recognized as belonging to a class), TN stands for "true negatives" (i.e., the number of texts correctly identified as not belonging to a category). FP and FN stand for "false positives" and "false negatives," i.e., the number of complaints falsely classified as either belonging or not

belonging to a class.

(1) Accuracy denotes the probability that the classifier correctly classifies a sample. In other words, it is the ratio of the number of correctly classified samples and the total number of samples:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

(2) The precision rate is the ratio of true positives and the number of samples the classifier assigned to the class:

$$P = \frac{TP}{TP+FP} \quad (3)$$

(3) The recall rate is the ratio of true positives and the number of truly positive samples. In other words,

$$R = \frac{TP}{TP+FN} \quad (4)$$

(4) The F1 value is the harmonic mean, which combines P and R and is calculated as follows:

$$F1 = \frac{2\times P\times R}{P+R} \quad (5)$$

4.3 Analysis of results

4.3.1Analysis of shallow-level auto-labeling results

The results of the shallow-level labeling are shown in Table 3. The table reports the overall accuracy (i.e., accuracy of all samples and categories), average accuracy (i.e., average categorical accuracy), average recall, and average F1 score. The results for each category are shown in Table 4.

Table 3 Overall evaluation of shallow-level text citation tags

Model	Overall accuracy	Average accuracy	Recall rate	F1 value
BERT text classification after text enhancement	0.9175	0.9279	0.9175	0.9168

Table 4 The results of the BERT model's shallow-level labeling by the categories

Serial number	Accuracy	Recall rate	F1 value
1	0.5000	0.6000	0.5455
2	0.9545	0.9545	0.9545
3	1.0000	0.9963	0.9982
4	0.6154	0.8889	0.7273
5	0.5000	0.5000	0.5000
6	0.8750	0.9333	0.9032
7	0.9286	0.8125	0/8667
8	0.8889	0.8000	0.8421

9	0.9130	0.6562	0.7636
10	1.0000	0.1667	0.2857
11	0.6962	0.8333	0.7586
12	0.4167	0.7143	0.5263
13	0.5000	0.7500	0.6000
14	0.9773	0.8776	0.9247
15	0.9664	1.0000	0.9829
16	1.0000	0.8462	0.9167

We also choose to use the LR classification models of BERT without text enhancement and TF-IDF with text enhancement to conduct experiments respectively. The performance results are shown in Table 5.

Table 5 Model comparison results

Model	Overall accuracy	Average accuracy	Recall rate	F1 value
Text-enhanced BERT model	0.9175	0.9279	0.9175	0.9168
Unenhanced BERT model	0.9073	0.8672	0.8642	0.8548
TF-IDF-LR model	0.8960	0.9011	0.8960	0.8899

The table shows that the text-enhanced BERT model outperforms both the BERT model (without enhancement) and the TF-IDF-LR model with respect to every measure. The TF-IDF-LR model had lower overall accuracy, recall rate, and F1 score than the BERT model applied to the non-enhanced dataset. The results by each category are shown in Table 6 (for BERT model without text-enhancement) and Table 7 (for TF-IDF-LR model).

Table 6 The results of the BERT model's (without text enhancement) shallow-level labeling by category

Serial number	Accuracy	Recall rate	F1 value
1	1.0000	0.2500	0.4000
2	0.8750	0.9545	0.9130
3	1.0000	0.9961	0.9980
4	0.5333	0.8889	0.6667
5	0.0000	0.0000	0.0000
6	0.7647	0.8667	0.8125
7	0.8824	0.9375	0.9091
8	1.0000	0.6842	0.8125
9	0.8889	0.5000	0.6400
10	0.5000	0.5000	0.5000
11	0.6667	0.8065	0.7299
12	0.5000	0.8571	0.6316
13	1.0000	0.5000	0.6667
14	1.0000	0.8776	0.9348
15	0.9800	0.9899	0.9849
16	0.8571	0.9231	0.8889

Table 7 The results of the TF-IDF-LR model's shallow-level labeling by category

Serial number	Accuracy	Recall rate	F1 value
1	1.0000	0.5455	0.7059
2	0.8421	0.8000	0.8205
3	0.9890	1.0000	0.9945
4	1.0000	0.7692	0.8696
5	0.0000	0.0000	0.0000
6	0.8889	0.6667	0.7619
7	0.7143	0.7143	0.7143
8	0.8571	0.7059	0.7742
9	0.9167	0.6875	0.7857
10	0.5000	0.3333	0.4000
11	0.6047	0.9286	0.7324
12	1.0000	0.5333	0.6957
13	0.0000	0.0000	0.0000
14	0.9535	0.8367	0.8913
15	0.9821	0.9910	0.9865
16	0.9286	0.9286	0.9286

Table 6 shows that BERT achieves 0.0 accuracy for the fifth category. The training set for this category is very small, containing only two complaints. Similarly, TF-IDF-LR has 0.0 accuracy for the fifth category, but also for category 13, whose training set includes nine texts. However, the text-enhanced BERT model does not show the result of 0.0000, which improves the poor performance of the BERT model when there are too many index tags and too little text, which can prove the necessity of text enhancement in this paper.

4.3.2 Analysis of deep-level auto-labeling results

The results of the deep-level labeling are shown in Table 8. The table shows that - when the text is correctly classified on the shallow level, and then Word2Vec semantic similarity matching performed on the deep-level label - the model performs well in terms of accuracy, precision, recall, and F1 values. The results are substantially improved compared to the direct Word2vec semantic similarity matching on the texts. These results demonstrate the proposed model's validity regarding performing classification before the similarity calculation.

Table 8 Results' comparison for semantic similarity calculation methods

Methodology	Overall accuracy	Average accuracy	Recall rate	F1 value
Use layer-by-layer semantic matching first and then calculate similarity with Word2Vec	0.7357	0.7350	0.7357	0.7173
Calculate similarity directly with Word2Vec	0.3018	0.4035	0.3018	0.3313

Using the method of semantic matching layer by layer and similarity calculation to automatically labeling the customer complaint text, the results are shown in Table 9.

Table 9 Auto-labeling results (partial)

Serial number	Type of business level received	Secondary type of operations received	Receiving business Type III	Textual content	Deep-Level Labels
1	Business complaints	Electricity tariffs	tariff	After the meter was changed here, the staff indicated that the meter was in arrears and charged the...	tariff
2	Business complaints	Electricity tariffs	tariff	Unreasonable charge for electricity penalty (customer says monthly number. Even if you have paid the bill but do not renew the invoice...)	tariff
3	Business complaints	power metering	Rotation, household meter conversion	Without the customer's knowledge ... many times to the customer meter into ...	Meter wiring error
…	…	…	…	…	…
664	Service Complaints	Acts of service	Service attitude of other personnel	After reflecting the farm network charges. Subsequently, the staff of the Zhongqiang power supply house telephone contact with the user...	Service attitude of electrical inspectors
665	Service Complaints	Service Channels	E-Channel Services	I'm not going to be able to buy electricity, but I still can't buy electricity.	E-Channel Services
666	Service Complaints	Acts of service	Service Standard for Business Office Personnel	In the process of transferring the name change, the customer asks if he can print out the previously unchanged VAT invoice...	Service attitude of electrical inspectors

5 Conclusion

This work tackles the problems emerging from the hierarchical structuredness of customer complaints, where too many deep-level labels result in small category size and imbalanced samples. The paper proposes a new model based on BERT and Word2Vec that enables automatic labeling of customer complaints. This model uses text enhancement to mitigate the problem of small category sizes without changing the semantics. In accordance, the model improves the sample sizes' balance. The developed model relies on the BERT model to determine shallow-level labels and uses the Word2Vec model to derive deep-level labels, thus taking full advantage of the hierarchical characteristics of customer complaint labels.

The innovations of this paper are: (1) an automatic indexing model of customer complaint text based on Bert and word2vec is proposed. (2) Before determining the shallow level text indexing label based on Bert, we first process the text with too small sample size under some categories. By text enhancement, we can get more samples and keep the semantic of the samples unchanged, so as to improve the problem of insufficient model training caused by the small sample size and imbalance. (3) Firstly, the shallow level label is determined, and then the deep level label under the shallow level label is used for similarity matching, which makes full use of the hierarchical structure between labels and greatly improves the accuracy.

The experiments demonstrated the model's feasibility and validity. Nevertheless, there are several limitations to this work. The first is the issue of data availability since, due to the unique nature of the customer complaint data, the data can be obtained only through cooperation with a power company. Further, since the data is not public, data quality cannot be determined. Finally, the quality of the enhanced text can be improved by advancing the developed algorithm. The future work will be directed at optimizing the details and improving the model to advance auto-labeling.

Acknowledgements

The work is partially supported by the Chinese National Natural Science Foundation under Project 71673209.

Author contributions

XB.T and H.M designed the study. H.M, JN.L and X.D performed the experiments and was supported by XB.T. The manuscript was written by H.M, JN.L, and X.D. All authors approved the manuscript. This manuscript is our original work, and it is submitted for frst publication.

bibliography

[1] Atliha V, Sesok D. Text Augmentation Using BERT for Image Captioning[J]. Applied Sciences-Basel, 2020, 10(17).

[2] Kim S, Park H, Lee J. Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis[J]. Expert Systems with Applications, 2020, 152: 12.

[3] Bharti S K, Babu K S. Automatic keyword extraction for text summarization: A survey[J]. arXiv preprint arXiv:1704.03242, 2017.

[4] Luhn H P.A Statistical Approach to Mechanized Encoding and Searching of Literary Information[J].IBM Journal of Research and Development,1957,1(4):309-317

[5] LOIS, L. E. Experiments in automatic indexing and extracting. Information Storage and Retrieval, 1970, 6.4: 313-330.

[6] Salton G, Wong A,Yang C S. A Vector Space Model for Automatic Indexing[J]. Communications of ACM,1975,18(11):613-620.

[7] DEERWESTER, Scott, et al. Indexing by latent semantic analysis. Journal of the American society for information science, 1990, 41.6: 391-407.

[8]Anjewierden A, Kabel S.Automatic Indexing of Documents with Ontologies[A].In:Proceedings of the 13th Belgian/Dutch Conference on Artificial Intelligence(BNAIC-01)[C].Amsterdam,Neteherlands,2001:23-30.

[9] De Oliveira H, Augusto V, Jouaneton B, et al. Automatic and Explainable Labeling of Medical Event Logs With Autoencoding[J]. IEEE Journal of Biomedical and Health Informatics, 2020, 24(11): 3076-3084.

[10] Su Z, Yu S, Chu J, et al. A novel architecture: Using convolutional neural networks for Kansei attributes automatic evaluation and labeling[J]. Advanced Engineering Informatics, 2020, 44: 101055.

[11]Adinarayana S, Ilavarasan E. A Hybrid Imbalanced Data Learning Framework to Tackle Opinion Imbalance in Movie Reviews[M]//Communication Software and Networks. Springer, Singapore, 2021: 453-462.

[12] Madabushi H T, Kochkina E, Castelle M. Cost-sensitive BERT for generalisable sentence classification with imbalanced data[J]. arXiv preprint arXiv:2003.11563, 2020.

[13] Li Q, Song Y, Zhang J, et al. Multiclass imbalanced learning with one-versus-one decomposition and spectral clustering[J]. Expert Systems with Applications, 2020, 147: 113152.

Figures

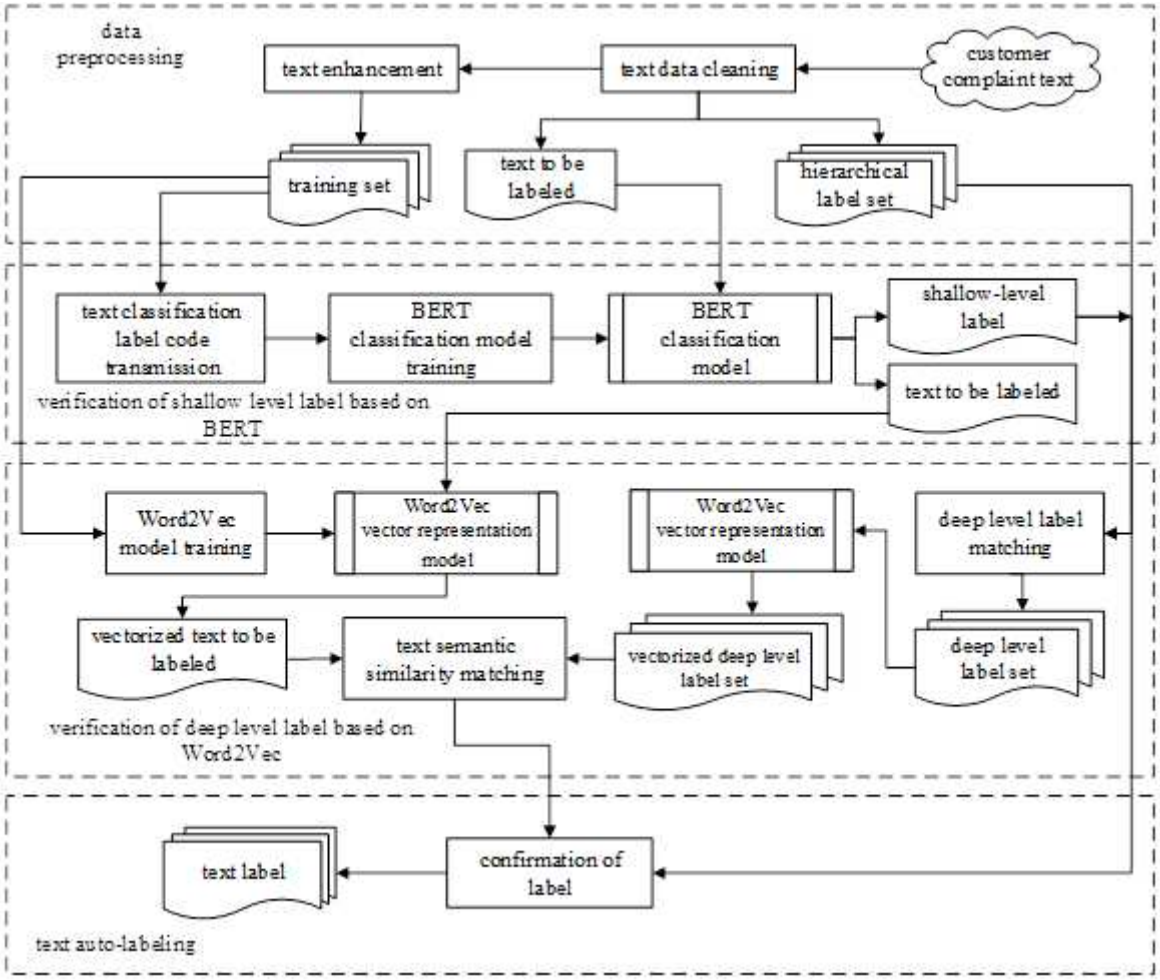

Figure 1

BERT and Word2Vec-based automatic text labeling model for customer complaints

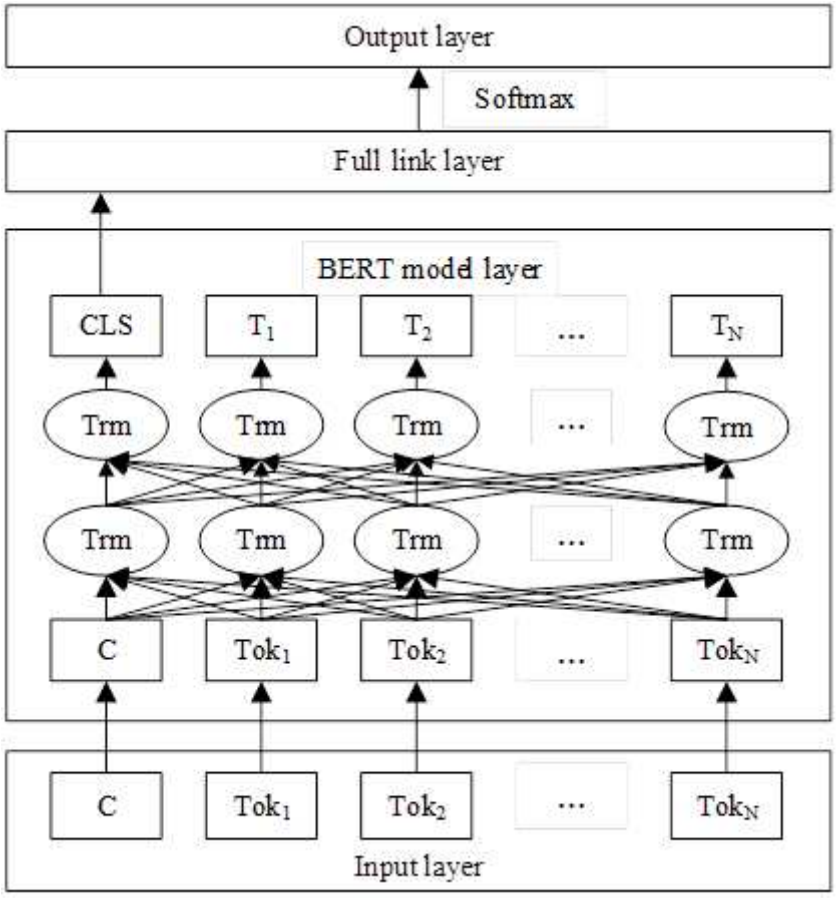

Figure 2

BERT-based text classification model

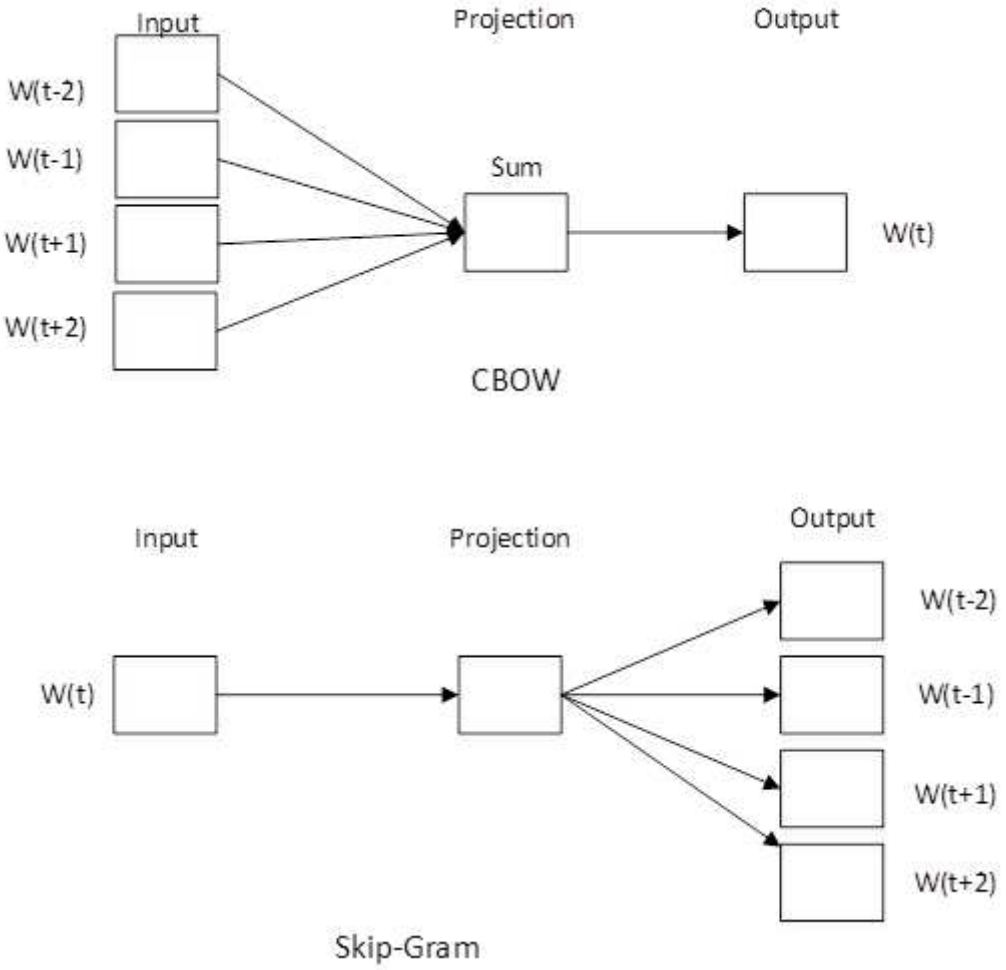

Figure 3

Two models of Word2Vec