A Comparative Study on the Integration of Text Augmentation and Pre-trained Language Models for Online Government Inquiry Message Classification (Postprint)

Authors: Shi Guoliang, Chen Yuqi

Date: 2023-04-01T16:02:52+00:00

Abstract

[Purpose/Significance] Government online petition platforms serve as a vital channel for government agencies to understand public opinion. To enhance the accuracy of petition message classification and tackle challenges including poor data quality and insufficient data volume, this study compares the classification effectiveness of various combinations of BERT-based improved models with text augmentation techniques, and investigates the underlying reasons for their performance differences. [Method/Process] We design an integrated comparative model for online petition message classification. For text augmentation, comparative experiments are conducted utilizing EDA and SimBERT text augmentation techniques. For text classification models, comparative experiments are performed employing multiple BERT-based improved pre-trained language models (e.g., ALBERT, RoBERTa). [Results/Conclusions] Experimental results demonstrate that the text classification model combining RoBERTa with SimBERT text augmentation achieves optimal performance, attaining an F1 score of 92.05% on the test set, which represents a 2.89% improvement over the BERT-base model without text augmentation. Additionally, SimBERT text augmentation yields an average F1 score improvement of 0.61% compared to non-augmented baselines. The experiments verify that the RoBERTa model with SimBERT text augmentation can effectively enhance multi-class text classification performance and offers strong reference value for addressing similar problems.

Full Text

Preamble

Volume 65, Issue 13, July 2021

A Comparative Study on the Integration of Text Enhancement and Pre-trained Language Models in the Classification of Online Political Inquiry Messages

Shi Guoliang, Chen Yuqi

Business School, Hohai University, Nanjing 211100

Abstract: [Purpose/Significance] Government online political inquiry platforms represent a crucial channel for authorities to understand public opinion. To improve the accuracy of message classification and address challenges such as poor data quality and limited quantity, this study compares the classification performance of various BERT-based improved models integrated with text enhancement techniques and explores the reasons for their differences. [Method/Process] We designed an integrated comparative model for online political inquiry message classification, employing EDA technology and SimBERT text enhancement techniques for comparative experiments in the text augmentation aspect, while utilizing multiple BERT-based improved pre-trained language models (such as ALBERT and RoBERTa) for comparative experiments in the text classification model aspect. [Result/Conclusion] Experimental results demonstrate that the text classification model based on RoBERTa and SimBERT text enhancement achieves the best performance, with an F1-score of 92.05% on the test set, which is 2.89% higher than the BERT-base model without text enhancement. Meanwhile, the F1-score after SimBERT enhancement is 0.61% higher on average compared to before enhancement. The experiments prove that the RoBERTa-SimBERT text enhancement model can effectively improve multi-category text classification performance and offers strong referential value for solving similar problems.

Keywords: Political inquiry platform; Text classification; Text enhancement; BERT model

Classification Number: G254

DOI: 10.13266/j.issn.0252-3116.2021.13.010

Online political inquiry has emerged as a novel form of democratic participation through which citizens engage in policy decision-making and safeguard their rights and interests [1]. Upholding the people-centered principle that "no issue concerning the masses is trivial," government agencies across China have successively launched various online political inquiry platforms such as "Mayor's Mailbox," "Please Share Your Thoughts," and "Public Opinion Message Board" in recent years. The rise of online political inquiry enables government departments to more conveniently, efficiently, and authentically understand public opinions and demands, significantly improving both administrative efficiency and citizen satisfaction. However, with the advent of the information age, the volume of online information has grown exponentially, and government online political inquiry messages are no exception. Manual classification methods can no longer keep pace with this data growth. Therefore, integrating Natural Lan-

guage Processing (NLP) technology into the "smart government" system is of great significance.

Traditionally, the classification of government online political inquiry messages has been viewed as a precursor to extracting valuable information from citizen feedback. Previous classification methods relied primarily on manual screening, which consumed substantial human and material resources. After introducing NLP technology, traditional text classification models often required retraining word vectors for different tasks to extract features, with model performance closely tied to the quality of the training corpus. Meanwhile, inquiry message data may contain considerable invalid information, and low data quality can affect classifier performance to some extent.

To address these issues, this paper conducts an integrated comparative study of multiple BERT (Bidirectional Encoder Representations from Transformers) series text classification models and different text enhancement algorithms. We propose a government online political inquiry message classification model based on RoBERTa and SimBERT text enhancement. We aim to contribute in the following aspects: Apply pre-trained language model technology to online political inquiry platform message classification tasks, leveraging multi-head attention mechanisms and bidirectional Transformer network structures to alleviate the "polysemy" problem that traditional classifiers cannot effectively resolve; Conduct integrated comparative studies using text enhancement models tailored to the specific characteristics of message texts to identify optimal combinations and partially solve data quality issues, thereby improving government work efficiency; Analyze the reasons for performance differences among experimental models from a model construction perspective, providing references and guidance for text processing practices in other domain text classification tasks and traditional NLP downstream tasks.

2 Research Status

2.1 Research on Text Representation Methods

Before text classification, texts must be modeled to extract and represent their features. Selecting appropriate text data features can effectively improve classification model performance. Text feature representation methods mainly include vector space text feature representation, pre-trained word vector text feature representation, and pre-trained language model text feature representation. Vector space models employ the TF-IDF algorithm and assign weights based on word frequency. However, due to the sparsity of short text features, traditional vector space models become overly sparse, affecting classification results. Pre-trained word vector-based text feature representation can effectively solve the vector matrix sparsity problem. Ma Sidan et al. [2] divided text keywords into overlapping and non-overlapping parts during word vector training and used parameterized linear weighting to calculate similarity between the two parts, proposing a weighted Word2vec text classification method that signifi-

cantly outperformed traditional vector space algorithms. Cheng Jing et al. [3] noted that training word vectors often fails to effectively update low-frequency words due to insufficient samples, suggesting that low-frequency words could be trained and updated through guidance from similar high-frequency words to optimize word vectors. Using pre-trained word vectors based on Word2vec and GloVe for downstream tasks yielded significant improvements.

Recent research on pre-trained language models demonstrates that unsupervised pre-trained semantic representation models can effectively address the shortcomings of the aforementioned two text feature representation methods. Pre-training refers to training models on unlabeled text data with the objective of predicting the next word in a sentence, thereby learning contextual representations of different words. In 2018, M. E. Peters et al. [4] proposed the ELMo (Embeddings from Language Models) model at the NAACL conference, which constructed dynamic word vectors through bidirectional Long Short-Term Memory (LSTM) network structures, alleviating the polysemy problem and ushering in the era of pre-trained language models for NLP tasks. Building on this foundation, Google's J. Devlin et al. [5] proposed the BERT model in October of the same year. BERT employs the Transformer model [6] with stronger semantic representation capabilities and integrated self-attention mechanisms to replace the LSTM structure in ELMo, while utilizing massive public corpora for training, significantly enhancing the dynamic word vector representation capabilities of pre-trained language models. If the ELMo model opened the era of pre-trained language models, the BERT model represents the pinnacle of pushing pre-trained language models forward through attention mechanisms and massive training corpora.

2.2 Research on Text Classification Algorithms

Text feature extraction effectively transforms texts into feature vectors to support subsequent tasks, while classification algorithms distinguish short text features and assign them to correct categories based on this foundation. As research has deepened, an increasing number of scholars have applied machine learning algorithms to text classification tasks with promising results. Chen Yanfang et al. [7] constructed an online product credibility factor index system and integrated it into an SVM classifier, proposing a DDAG-SVM online product review credibility classification model. Yu Bengong et al. [8] employed a multi-channel modeling approach using SVM and random forest, proposing an nLD-SVM-RF short text classification algorithm that improved model generalization performance.

Meanwhile, with increasing data volumes and improved computer performance, the advantages of deep learning algorithms in text classification have gradually become apparent. Han Dong et al. [9] assigned higher weights to topic sentences and integrated them into character-level Convolutional Neural Networks (CNN) for text classification research. Yang Yunlong et al. [10] addressed the long-term memory limitations of single Recurrent Neural Networks (RNN) by

proposing a text sentiment analysis model G-Caps that fuses capsule features with Gated Recurrent Units (GRU), effectively improving Chinese sentiment analysis performance. In research using pre-trained language models as base classifiers, Zhao Kun et al. [11] utilized a Chinese medical pre-trained model (BERT-Re-Pretraining-Med-Chi) for literature classification research. Wu Jun et al. [12] employed the BERT model for text vectorization before connecting it to a BiLSTM-CRF model for Named Entity Recognition (NER) of Chinese professional terms, achieving significant improvements over traditional pre-trained word vectors. Liao Shenglan et al. [13] used the BERT model as a "teacher model" for model distillation to improve the classification performance of lightweight classifiers such as Text-CNN, optimizing classification algorithms from both effectiveness and efficiency perspectives.

2.3 Research on Text Enhancement Algorithms

Government online political inquiry messages belong to the category of short texts, which are generally brief, highly informal, and non-standardized. Particularly, message texts contain numerous internet slang terms, colloquial expressions, and abbreviations, resulting in high noise levels and limited quantities of qualified texts. Text data enhancement techniques can alleviate these issues to some extent. W. Jason et al. [14] summarized and proposed systematic text enhancement strategies (Easy Data Augmentation, EDA) in 2019, primarily generating new sentences through word-level modifications to achieve text enhancement. Yu Chang et al. [15] used RNN as a generative network and CNN as a discriminative network, proposing a generative adversarial network (seqGAN) based text generation model for power user intentions, and verified the effectiveness of generated texts using the BLEU algorithm.

2.4 Research Gaps and Summary

Regarding text representation methods and classification models, although pre-trained word vector text representation methods effectively alleviate the shortcomings of traditional vector space representation by mapping different tokens into single vectors through word embedding, they still face several issues: Word segmentation is required before training word vectors, and inaccuracies in segmentation lexicons lead to ineffective recognition of out-of-vocabulary words, affecting vector representation accuracy. Applying the same pre-trained word vectors to text representation tasks in different contexts (medical or legal backgrounds, etc.) cannot achieve optimal results, while training word vectors based on specific tasks requires substantial training corpora and equipment support, resulting in low practical feasibility. Pre-trained word vectors can solve the "one meaning, multiple words" problem but cannot resolve the "one word, multiple meanings" problem. Therefore, this study selects pre-trained language models for text feature representation and classification. Meanwhile, compared to the bulky BERT model, this research employs improved versions based on BERT, using RoBERTa [16] and ALBERT [17] pre-trained language models for

online political inquiry message classification, and attempts to connect neural networks as classifiers to improve model performance.

In terms of text data enhancement, EDA text enhancement techniques are mostly rule-based, and the vector feature representations of generated texts may not differ significantly from the original texts, leading to repetitive and ineffective training samples. Meanwhile, texts generated through generative adversarial networks are random and domain-specific without rules, and their category labels must be predicted by the model's discriminative network, inevitably introducing errors that affect model training. To address these issues, this study adopts the SimBERT [18] text enhancement technique. SimBERT is primarily trained using supervised similar text pairs based on the BERT model, capable of generating similar sentences for specific sentences while using original data labels, simultaneously solving both the polysemy problem in generated texts and the label inaccuracy issue.

3 Model Design and Overall Framework

To improve the accuracy of government online political inquiry message classification and advance the construction of a "smart government" service system, this paper designs an integrated comparative model for online political inquiry message classification. The model employs the popular BERT pre-trained language model and its improved versions from recent NLP research as text representation models, combined with EDA and SimBERT text enhancement algorithm models to complete message text classification. The overall design framework of the experimental model is shown in Figure 1.

3.1 Text Classification Model Selection and Design

3.1.1 BERT Model We select BERT and its improved pre-trained language models to generate text character embedding vectors for solving text classification tasks based on the following considerations: Pre-trained language models can complete text vectorization representation and fine-tuning in an end-to-end manner without relying on traditional manual feature extraction; The BERT model based on bidirectional Transformer structure and attention mechanisms can effectively solve problems such as polysemy during text feature extraction; Chinese texts have two different segmentation granularities (character and word). Traditional pre-trained word vectors are mostly trained after word segmentation, which inevitably introduces errors, while BERT can perform character-level embedding, generally achieving better performance in Chinese tasks.

BERT, fully known as Bidirectional Encoder Representations from Transformers [5], is a pre-trained language model proposed by Google at the end of 2018 based on the Transformer [6] structure integrated with self-attention mechanisms. BERT's advantage lies in its powerful word vector generalization capability. Unlike traditional one-hot encoding and static Word2vec pre-trained

word vectors, BERT dynamically adjusts word vectors through bidirectional Transformer structures, fully incorporating contextual information of words, which can effectively solve polysemy problems. The structure diagram of the BERT model is shown in Figure 2.

In Figure 2, $E_1, E_2, \ldots, E_N$ represent the word-level Embedding layer of input text, which then passes through the bidirectional Transformer encoder to obtain outputs $T_1, T_2, \ldots, T_N$ that incorporate contextual information through the Self-attention mechanism. In the original BERT model, feeding these outputs into a softmax layer yields the final classification results.

The basic internal structure of the BERT model is the Transformer model [6], which is a Seq2seq model based on the Self-attention mechanism and represents a typical Encoder-Decoder structure. It primarily encodes input sequences into fixed-length vectors through the Encoder layer, then decodes these vectors into output sequences of required length through the Decoder layer. The input to the Transformer Encoder module is the word embedding representation of text, incorporating position encoding information (Position Encoding). The model's core lies in the Self-attention layer that replaces traditional RNN and CNN structures. The Self-attention layer functions as a sequence encoding layer, primarily incorporating the relationship between a given word and other parts of the sentence into the word's vector representation, thereby solving polysemy problems.

The main principles and calculation steps are as follows:

(1) Input sentence text is embedded into word vectors at the character/word level.
(2) Word vectors are multiplied with weight matrices W_Q, W_K, and W_V to obtain corresponding Queries (q), Keys (k), and Values (v) vectors, where the dimensions of W_Q, W_K, and W_V are $N \times d_k$, $N \times d_k$, and $N \times d_v$ respectively. The dimensions of Queries and Keys vectors are d_k, while Values have dimension d_v.
(3) Calculate the score for each vector: $score = q \cdot k$, representing the attention a word pays to other parts of the sentence when encoding that word.
(4) Apply the softmax activation function to convert scores for different words into weights between 0-1 that sum to 1.
(5) Multiply the softmax weights with v to obtain weighted score vectors, then sum them to get the final output z, where $z_i = \sum w_i \cdot v_i$.

The formulas are:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QW_Q \cdot (KW_K)^T}{\sqrt{d_k}}\right) VW_V \quad (1)$$

$$\text{Head}_i = \text{Attention}(QW_Q^i, KW_K^i, VW_V^i) \quad (2)$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_n)W^O \quad (3)$$

Formula (1) is the Self-attention mechanism calculation formula, while formulas (2) and (3) are the Multi-Head Attention calculation formulas. W_Q^i represents the weight matrix in the i-th "head" of the self-attention mechanism, and W^O is the matrix of the fully connected layer, which constructs the final output result matrix of specified dimensions after concatenating the outputs of different attention heads and passing through the fully connected layer.

Meanwhile, one of BERT's major innovative contributions is its unique pre-training approach. BERT employs Masked Language Model (MLM) and Next Sentence Prediction (NSP) as pre-training tasks, effectively enhancing the model's deep bidirectional prediction and reasoning capabilities. BERT's input vector primarily consists of the weighted sum of word vectors, segment vectors, and position vectors. The beginning and end of sentences are marked with [CLS] and [SEP] tokens respectively, with [SEP] also used to separate sentences. The specific structure is shown in Figure 3.

The MLM model primarily masks one or several words in a sentence with a 15% probability, training the model to predict the masked words using remaining words, similar to a cloze test. This approach prevents performance degradation from excessive use of [MASK] tokens without affecting model comprehension. Unlike bidirectional LSTM models that can only train the model to understand left-to-right and right-to-left context separately, the MLM model enables deep bidirectional training of BERT's inter-sentence information understanding capabilities.

The NSP model aims to train the model's sentence-level contextual relationships by inputting numerous sentence pairs AB from the corpus, where sentence B has a 50% probability of being the next sentence of A and a 50% probability of being a randomly selected sentence. The two sentences are separated by [SEP] tokens, and the model performs binary classification prediction training on these sentence pairs to improve its understanding of sentence-level relationships.

Additionally, this paper improves upon BERT by connecting the output vectors from BERT's input layer to other text classifiers, specifically selecting Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), and Recurrent Convolutional Neural Networks (RCNN) as subsequent text classifiers. The Text-RNN model, proposed by P. F. Liu et al. [19], is a text classifier based on RNN and its variants (LSTM, GRU) that can incorporate contextual semantics through bidirectional RNNs, achieving good classification results but with slower training speed. Y. Kim [20] first applied CNN, originally used for image processing, to text classification tasks, proposing the Text-CNN model that treats sentences or words as word vector matrices input to the model, extracting important sentence features through convolutional and pooling layers for classification. S. W. Lai et al. [21] combined Text-RNN and Text-CNN models to propose the Text-RCNN model, which captures contextual information through RNN's bidirectional recurrent structure and captures key information through CNN's max pooling layer, addressing the bias problem of Text-RNN

and the fixed window feature extraction limitation of Text-CNN.

3.1.2 RoBERTa Model To further improve model performance, XLNet [22] and RoBERTa [16] models optimized the original BERT pre-training approach while increasing training data volume and duration. However, this approach leads to excessively large model parameters. Although model performance improves as parameters increase, when model complexity becomes too high and parameters too numerous, performance 反而 decreases, a phenomenon known as "Model Degradation."

To address this issue, researchers have used Knowledge Distillation (KD) methods to reduce model parameters, represented by DistilBERT [23] and TinyBERT [24] models proposed by Huawei Noah's Ark Lab, both achieving the goal of reducing model size and parameters. Although knowledge distillation can reduce model scale and improve computational speed, it sacrifices model performance. For example, TinyBERT is only 13.3% the size of BERT with 28% of its parameters, but its performance on the GLUE benchmark drops by 3 percentage points compared to BERT.

To solve these problems, Google's Zhenzhong Lan et al. [17] proposed the ALBERT model, which successfully reduced model parameters by 18 times while 反而 exceeding the performance of BERT, XLNet, and other large-scale pre-trained language models. ALBERT's main improvements over BERT include three aspects:

(1) **Embedding layer factorization**. In BERT, both the initial word embedding and the final output embedding after encoding have dimension 768. The ALBERT research team argued that the original word embedding contains much less information than the hidden layer output embedding, thus the word-level embedding dimension can be reduced. This method maps one-hot vectors to a low-dimensional space to reduce word-level embedding dimension (E), then maps to a high-dimensional space to maintain the hidden layer encoder output dimension (H) unchanged, ultimately reducing model parameters from O(V×H) to O(V×E + E×H).

(2) **Cross-layer parameter sharing**. The Transformer model proposed parameter sharing methods but only shared parameters of fully connected layers or Attention layers. ALBERT combines both approaches by sharing parameters across all encoder layers, significantly reducing model parameters while improving training speed. Although model performance decreases slightly, it enables training with larger-scale data to improve effectiveness.

(3) **Sentence Order Prediction (SOP) task**. While RoBERTa replaced the NSP pre-training task with FULL SENTENCES, ALBERT improved upon NSP by proposing the SOP pre-training task. Compared to NSP, SOP better trains the model's ability to infer inter-sentence relationships. Its essence remains training a binary classifier with the same positive sam-

ples as NSP, but negative samples are changed to predicting whether two adjacent sentences are in reversed order.

Additionally, ALBERT made improvements such as removing dropout layers, truly achieving simultaneous model size reduction and performance improvement, laying a solid foundation for the industrial deployment of BERT series models.

3.2 Text Enhancement Model Selection and Design

Text classification tasks face varying data quality and quantity across different scenarios. The government online political inquiry message data studied in this paper also suffers from high noise levels and limited qualified data. To mitigate classification model performance degradation caused by training data issues, this paper employs EDA [14] and SimBERT [18] text enhancement techniques for data augmentation and conducts comparative experiments to explore the effects and advantages/disadvantages of different text enhancement techniques.

3.2.1 EDA Text Enhancement As a representative of systematic text enhancement strategies, the EDA text enhancement algorithm primarily adopts traditional rule-based text enhancement methods, making modifications at the word and syntax levels. The main data operations and examples of EDA text enhancement are shown in Table 1 .

3.2.2 SimBERT Text Enhancement Experiments show that EDA text enhancement technology can improve model performance to some extent, but rule-based data augmentation methods still have limitations. To further enhance text augmentation effectiveness, this paper attempts to adopt the SimBERT text enhancement model [18]. SimBERT is a generative language model that integrates retrieval and generation, built upon the BERT model and incorporating Microsoft's UniLM model [25] training philosophy.

UniLM is essentially a unified pre-trained language model that jointly pre-trains multiple different language models with common objectives through special Attention MASK methods, achieving parameter sharing across different language models through a single Transformer model. This parameter sharing approach enables the model to simultaneously learn and fuse different text feature representations, achieving joint optimization.

SimBERT primarily borrows the Seq2Seq training approach from UniLM, belonging to supervised training where the training corpus consists of large amounts of collected similar text pairs. The main training objective is to construct the Seq2Seq component capable of predicting similar sentences for given sentences. During SimBERT training, different sentences in similar sentence pairs are separated by [SEP] tokens, and a special Attention MASK method is applied: tokens before [SEP] perform bidirectional Attention, while tokens after [SEP] perform unidirectional Attention, enabling the model to

recursively predict the latter half and possess NLG capabilities. Additionally, SimBERT incorporates random [MASK] tokens during input, allowing the model to perform MLM tasks during training, which cultivates its NLU capabilities.

During training, SimBERT concatenates all [CLS] sentence vectors in each batch to form a sentence vector matrix $D \in \mathbb{R}^{b \times d}$ (where b is batch_{size} and d is hidden_{size}), applies L2 regularization on the hidden_{size} dimension to obtain the regularized matrix $\tilde{D}$, and performs inner product operations to obtain the final similarity matrix $\tilde{D}\tilde{D}^T \in \mathbb{R}^{b \times b}$, with diagonal elements masked out. SimBERT uses this similarity matrix for classification tasks, where negative samples are dissimilar texts, and employs softmax operations to increase similarity for positive samples while decreasing it for negative samples. Ultimately, SimBERT's loss function is the joint loss combining Seq2Seq loss and softmax layer loss from the similar sentence classifier. The training approach is illustrated in Figure 4 [Figure 4: see original paper].

4 Experimental Process and Results Analysis

4.1 Data Source and Preprocessing

All experimental data in this paper originates from real message data from a provincial online political inquiry platform in China from 2014-2020, retrieved in April 2020, totaling 9,281 entries. The data includes seven categories: urban and rural construction, labor and social security, education and culture, transportation, etc. For data preprocessing, we removed duplicates based on message text length and repetition, desensitized location-sensitive vocabulary including cities, districts, counties, and towns in the message texts, and finally divided the data into training and test sets at an 8:2 ratio, with the training set further divided to create a validation set at the same ratio. Partial data samples are shown in Table 2 , and the message category distribution is shown in Figure 5 [Figure 5: see original paper].

The category distribution reveals that messages in urban and rural construction, labor and social security, and education and culture categories are more numerous. These three categories permeate the basic aspects of daily life. Most issues in urban and rural construction reflect property management problems and surrounding environmental issues. Having good living conditions and environments is fundamental to people's livelihoods and the most basic prerequisite for conducting other social activities. Labor and social security issues are closely related to people's personal interests, and resolving these issues is also an important way to promote social fairness. Education and culture issues concern personal and children's learning and development, representing people's pursuit of better development after meeting basic living needs. Therefore, in online political inquiry messages, these three categories receive the highest public attention and are the most urgent issues people hope to resolve. Effectively and correctly identifying the categories of concerns from large volumes of complex

messages is fundamental to improving government administrative efficiency and an important means of safeguarding people's basic interests.

4.2 Experimental Setup

4.2.1 Experimental Environment Configuration Experiments were primarily conducted on PyCharm using Python 3.7.3. Detailed experimental environment and hardware/software configurations are shown in Table 3 .

4.2.2 Model Parameter Settings The integrated comparative model for online political inquiry message classification in this paper consists of text classification and text enhancement modules. To achieve optimal performance on the validation set, specific parameter settings are as follows:

For text classification models, we selected three different Chinese pre-trained language models: the open-source bert_{{{base}}*{{chinese}}} model, albert{{{base}}*{{chinese}}} model, and robert{{{base}}*{{chinese}}} model. The BERT model has 12 layers, employs 12-head attention mode, has a hidden layer dimension of 768, and contains 110M parameters. The ALBERT model has 12 layers, a hidden layer dimension of 128, and contains 12M parameters. The RoBERTa model has 6 layers, a hidden layer dimension of 384, and contains 200M parameters. In the BERT baseline model, the pad{size} for text truncation/padding is set to 128, batch training size batch_{size} is 32, initial learning rate learning_{rate} is 2e-5, and the optimizer is BertAdam. For ALBERT and RoBERTa models, the optimizer is AdmLR with learning_{rate} set to 1e-4, while other parameters remain the same as the BERT baseline model.

For text enhancement models, in EDA parameters, the probability alpha for each word in a sentence to be replaced is set to 0.3, and the number of augmentations num_{aug} is set to 1. In SimBERT parameters, n is set to 25 and k is set to 1, where n represents the number of similar sentences generated by seq2seq, and k represents the number of most similar sentences returned after similarity calculation by the encoder.

4.3 Experimental Results Analysis and Discussion

4.3.1 Model Effectiveness Evaluation Metrics Online political inquiry message classification is a text classification problem. To evaluate and compare model effectiveness, this paper adopts Precision (P), Recall (R), and F1-score (F-score) as model performance metrics. The calculation formulas are as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (5)$$

$$F1 = \frac{2 \cdot P \cdot R}{P + R} \quad (6)$$

Where Precision (P) represents the ratio of true positive samples (TP) predicted as positive to all samples predicted as positive (TP + FP). Recall (R) represents the ratio of true positive samples (TP) predicted as positive to all actual positive samples (TP + FN). The F1-score is the harmonic mean of precision and recall, accurately reflecting overall model performance.

4.3.2 Integrated Comparative Model Effectiveness Analysis To compare the integrated effects of different text enhancement techniques and pre-trained language models on online political inquiry message classification, this paper designed an integrated comparative model. Models were trained on the training set, optimized using the validation set, and finally evaluated on the test set using the aforementioned metrics. Specific experimental results are shown in Table 4 and Figure 6 [Figure 6: see original paper].

(1) Analysis of Online Political Inquiry Message Text Classification Effectiveness. The comparative results show that without considering text enhancement, the RoBERTa model achieves the best performance on the political inquiry message classification task, with an F1-score of 91.28%. The ALBERT model also achieves a high F1-score of 90.89%. The average F1-score of BERT series models is 89.42%, significantly lower than RoBERTa and ALBERT.

Among the first four models, BERT-base, BERT+RNN, and BERT+CNN show little performance difference, while BERT+RCNN demonstrates significantly better classification performance, even exceeding the non-augmented ALBERT model after text enhancement. The reason is that the RCNN network structure combines the advantages of RNN and CNN networks: it can capture sentence contextual information through bidirectional RNN structure while capturing key information through CNN's max pooling layer, enabling more accurate expression of sentence semantic structure.

Furthermore, ALBERT and RoBERTa models perform better primarily because they are both improved pre-trained language models based on BERT, employing larger training data volumes and longer training times compared to BERT. RoBERTa mainly improves BERT's pre-training tasks by converting static Masking to dynamic Masking and proposing the FULL-SENTENCES pre-training task to train the model's inter-sentence understanding capability. ALBERT, while changing the NSP pre-training task to SOP, proposes factorizing the embedding layer and cross-layer parameter sharing to drastically reduce model parameters, enabling training with more data under equivalent time and space complexity. In reference [17], ALBERT outperformed RoBERTa on most tasks, but in this paper, RoBERTa's F1-score is 0.39% higher than ALBERT's. Apart from different data background influences, this is also because political inquiry message classification is a labeling task where RoBERTa performs slightly better than ALBERT, consistent with our experimental results.

(2) Analysis of Text Enhancement Effects for Online Political Inquiry Messages. This paper primarily employs rule-based EDA text enhancement algorithm and similarity sentence generation-based SimBERT text enhancement model. After data preprocessing and dataset splitting, text data enhancement techniques expanded the training set to twice its original size, and models were retrained for comparison with non-augmented models. The comparison results between non-augmented and augmented data under both models are shown in Table 5 .

Table 5 demonstrates the effects of EDA and SimBERT text enhancement on identical text content. EDA text enhancement, true to its principle, only makes rule-based modifications to individual words, while SimBERT text enhancement is not simple reordering but tends to rewrite original statements in interrogative form.

The model effectiveness comparison results in Table 4 and Figure 6 show that due to data volume and quality issues, model classification performance differs significantly before and after text enhancement. Comparative experiments reveal that models after EDA and SimBERT text enhancement achieve average F1-score improvements of 0.59% and 0.61% respectively compared to before enhancement, proving that constructing training data through text enhancement models can indeed improve model performance to some extent when data is limited or of low quality.

The F1-score after SimBERT enhancement is 0.02% higher on average than after EDA enhancement. The reason is that EDA data enhancement is rule-based, and although it improves performance, it still has drawbacks: Synonym replacement may produce words with little vector difference from original words, resulting in limited enhancement effects; Random word deletion may remove core keywords, causing deviation from the original label; Insertion and replacement operations may alter sentence structural semantics, potentially counterproductive in tasks with structural requirements. In contrast, SimBERT employs joint training of Seq2Seq structure and similar sentence classification tasks, demonstrating better performance in natural language generation tasks and improving data quality to some extent. However, due to the randomness of generation, EDA enhancement occasionally yields better results on individual models. Therefore, text enhancement technology can improve classification model performance, but the overall improvement remains limited.

(3) Analysis of RoBERTa-SimBERT Model Results. The integrated comparative model for online political inquiry message classification shows that the RoBERTa model with SimBERT text enhancement achieves the best classification performance. Compared with the non-augmented BERT-base and BERT+RNN models, the F1-score improves by 2.89% and 3.73% respectively. Detailed classification results of the RoBERTa-SimBERT model are shown in Table 6 .

Table 6 shows that the RoBERTa-SimBERT model achieves good overall classi-

fication performance on political inquiry messages, with the best F1-scores for education and culture (94.69%), labor and social security (93.15%), and urban and rural construction (92.21%)—precisely the categories with the most message data and highest public concern. Relatively, health and family planning and commerce and tourism show poorer classification performance with F1-scores of only 91.29% and 88.68% respectively.

Table 7 presents typical misclassification instances of the RoBERTa-SimBERT model on the test set. By analyzing these errors, we attempt to explore reasons for performance differences between different message categories.

Misclassification analysis reveals that ambiguous categories affecting classification results concentrate mainly between commerce/tourism and health/family planning. As shown in Table 6, commerce/tourism has an F1-score of only 88.68%, while health/family planning, although exceeding 91%, has a recall rate of only 89.18%. Most misclassification instances incorrectly classify commerce/tourism messages into other categories, resulting in poor overall classification performance for this category. Among misclassified categories, many messages are incorrectly assigned to health/family planning, leading to its low recall rate.

Qualitative analysis of misclassification reasons through examining specific message content shows that: (1) Messages that should be classified as commerce/tourism but were misclassified tend to emphasize "commerce" rather than "tourism." While tourism-oriented messages are easier to classify correctly, commerce-oriented messages easily confuse with other categories, especially urban/rural construction. (2) Many messages misclassified as health/family planning involve hygiene issues of manufacturers and shops, which easily confuse with commerce themes in commerce/tourism, affecting classification results.

Many similar misclassification examples exist, and classifier errors are closely related to message text expression. Some messages are inherently ambiguous or multi-categorical, affecting model performance. While text enhancement improves data quality and quantity, bringing some improvement, it can only maintain relatively good classification levels. Further performance improvement requires optimizing the original data source. Governments can implement finer-grained input restrictions on their online political inquiry platforms to make message content more detailed and standardized, thereby improving classification effectiveness and efficiency.

Conclusion

The rise of online political inquiry platforms provides channels for citizens to express their opinions. Effective classification of messages can help government departments better understand public sentiment and improve themselves. To enhance classification accuracy and model deployment efficiency, this paper combines pre-trained language models with text enhancement technology, proposing

through comparative experiments a government online political inquiry message classification model based on RoBERTa and SimBERT text enhancement.

Traditional pre-trained word vector-based text feature extraction models cannot effectively handle text polysemy. BERT series pre-trained language models successfully solve this problem through bidirectional Transformer network structures and multi-head attention mechanisms. Meanwhile, using pre-trained language models enables effective end-to-end model deployment, allowing fine-tuning on data for application to multiple downstream tasks. Character-level text vectorization is also more suitable for Chinese text representation.

To effectively address data quality issues in the political inquiry message domain, this paper utilized EDA and SimBERT technologies for text enhancement to alleviate problems of insufficient training data and poor data quality. Research results demonstrate that texts generated by SimBERT show improved quality over those generated by EDA, while also solving label prediction problems in text classification data enhancement, providing references for related text enhancement problems.

Furthermore, in comparative experiments, we applied popular BERT-based improved pre-trained language models such as BERT+RCNN, ALBERT, and RoBERTa to political inquiry message classification, explaining the advantages, disadvantages, and applicable scopes of different models in accessible terms. We attempted to explain performance differences from the perspective of model structure and pre-training methodology principles. BERT series models are trained on high-quality corpora covering all industries, providing highly referential and reusable solutions for text classification problems in other domains.

This study has certain limitations. The dataset coverage is relatively small, focusing mainly on multi-classification problems of political inquiry messages. Future research can employ datasets from different domains for comparison to explore more general solutions and models. Meanwhile, domain-specific corpora can be used to fine-tune pre-trained language models to enhance their text representation and discrimination capabilities in specific fields. Regarding model structure, subsequent research can attempt to connect other network structures after excellent pre-trained language models such as ALBERT and RoBERTa to achieve better results.

References

[1] Xu Xiaowen, Cao Shouxin. The influence of online political inquiry on public policy formulation—Based on SWOT analysis [J]. Shandong Social Sciences, 2015(6): 179-183.

[2] Ma Sidan, Liu Dongsu. Research on text classification method based on weighted Word2vec [J]. Information Science, 2019, 37(11): 38-42.

[3] Cheng Jing, Liu Nana, Min Kerui, et al. A low-frequency word vector optimization method and its application in short text classification [J]. Computer

Science, 2020(8): 255-260.

[4] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations [C]//Proceedings of the conference of the north american chapter of the Association for Computational Linguistics: human language technologies. Stroudsburg: Association for Computational Linguistics, 2018: 2227-2237.

[5] Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the conference of the North American chapter of the Association for Computational Linguistics: human language technologies. Stroudsburg: Association for Computational Linguistics, 2019: 4171-4186.

[6] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Proceedings of advances in neural information processing systems. California: MIT Press, 2017: 6000-6010.

[7] Chen Yanfang. Online product review credibility classification model based on DDAG-SVM [J]. Information Studies: Theory & Application, 2017, 40(7): 132-137.

[8] Yu Bengong, Cao Yumeng, Chen Yangnan, et al. Short text classification research based on nLD-SVM-RF [J]. Data Analysis and Knowledge Discovery, 2020, 4(1): 111-120.

[9] Han Dong, Wang Chunhua, Xiao Min. Short text classification method based on sentence-level learning improved CNN [J]. Computer Engineering and Design, 2019, 40(1): 256-260.

[10] Yang Yunlong, Sun Jianqiang, Song Guochao. Text sentiment analysis based on GRU and capsule feature fusion [J/OL]. [2021-02-10]. http://kns.cnki.net/kcms/detail/51.1307.TP.20200429.1704.010.html.

[11] Zhao Kun, Zhang Zhixiong, Liu Huan, et al. Chinese medical literature classification research based on BERT model [J]. Data Analysis and Knowledge Discovery, 2020(8): 41-49.

[12] Wu Jun, Cheng Kui, Hao Han, et al. Chinese professional term extraction research based on BERT embedding BiLSTM-CRF model [J]. Journal of the China Society for Scientific and Technical Information, 2020, 39(4): 409-418.

[13] Liao Shenglan, Ji Jianmin, Yu Chang, et al. Intent classification method based on BERT model and knowledge distillation [J/OL]. [2021-02-10]. https://doi.org/10.19678/j.issn.1000-3428.0057416.

[14] Jason W, Kai Z. Eda: easy data augmentation techniques for boosting performance on text classification tasks [C]//Proceeding of the 2019 conference on empirical methods in natural language processing. Hong Kong: ACL, 2019.

[15] Yu Chang, Ouyang Yu, Zhang Bo, et al. Power user intention text generation based on adversarial generative networks [J]. Information Technology and Network Security, 2019, 38(11): 67-72.

[16] Liu Y, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach [J/OL]. [2021-02-10]. https://arxiv.org/abs/1907.11692.

[17] Lan Z Z, Chen M D, Goodman S, et al. ALBERT: a lite BERT for self-supervised learning of language representations [C]//Proceedings of the international conference on learning representations. Ethiopia: ICLR, 2020.

[18] Su Jianlin. SimBERT model integrating retrieval and generation [EB/OL]. [2020-05-18]. https://kexue.fm/archives/7427.

[19] Liu P F, Qiu X P, Huang X J. Recurrent neural network for text classification with multi-task learning [C]//Proceeding of the international joint conference on artificial intelligence. New York: IJCAI, 2016.

[20] Kim Y. Convolutional neural networks for sentence classification [C]//Proceeding of the 2014 conference on empirical methods in natural language processing. Doha: ACM, 2014.

[21] Lai S W, Xu L H, Liu K, et al. Recurrent convolutional neural networks for text classification [C]//Proceeding of the 29th national conference on artificial intelligence. Austin: AAAI, 2015.

[22] Yang Z L, Dai Z H, Yang Y Y, et al. XLNet: generalized autoregressive pretraining for language understanding [C]//Proceedings of the 33rd conference on neural information processing systems. Vancouver: MIT Press, 2019.

[23] Sanh V, Debut L, Chaumond J, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter [J/OL]. [2021-02-10]. https://arxiv.org/abs/1910.01108v4.

[24] Jiao X Q, Yin Y C, Shang L F, et al. TinyBERT: distilling BERT for natural language understanding [J/OL]. [2021-02-10]. https://arxiv.org/abs/1909.10351v3.

[25] Bao H B, Dong L, Wei F R, et al. UniLMv2: pseudo-masked language models for unified language model pre-training [J/OL]. [2021-02-10]. https://arxiv.org/abs/2002.12804.

Author Contributions

Shi Guoliang: Proposed research ideas, revised research plans and paper, finalized manuscript.

Chen Yuqi: Designed research plan, processed data, built models, and wrote paper.

Note: Figure translations are in progress. See original paper for figures.