Computers, Materials & Continua CMC, vol.63, no.3, pp.1499-1514, 2020

MII: A Novel Text Classification Model Combining Deep Active Learning with BERT

Anman Zhang[1], Bohan Li[1, 2, 3, *], Wenhuan Wang[1], Shuo Wan[1] and Weitong Chen[4]

Abstract: Active learning has been widely utilized to reduce the labeling cost of supervised learning. By selecting specific instances to train the model, the performance of the model was improved within limited steps. However, rare work paid attention to the effectiveness of active learning on it. In this paper, we proposed a deep active learning model with bidirectional encoder representations from transformers (BERT) for text classification. BERT takes advantage of the self-attention mechanism to integrate contextual information, which is beneficial to accelerate the convergence of training. As for the process of active learning, we design an instance selection strategy based on posterior probabilities Margin, Intra-correlation and Inter-correlation (MII). Selected instances are characterized by small margin, low intra-cohesion and high inter-cohesion. We conduct extensive experiments and analytics with our methods. The effect of learner is compared while the effect of sampling strategy and text classification is assessed from three real datasets. The results show that our method outperforms the baselines in terms of accuracy.

Keywords: Active learning, instance selection, deep neural network, text classification.

1 Introduction

Supervised learning calls for much data to train. Collecting a large amount of data is costly and intractable. As an effective way to reduce labeling cost, active learning (AL) begins with a small training set, and then iteratively adds the most uncertain or informative instances into itself. Previous work about AL mainly involved traditional machine learning scene and sentiment analysis. Yue et al. [Yue, Chen, Li et al. (2018)] gave an extensive survey on classification of sentiment analysis about social media. Wan et al. [Wan, Li, Zhang et al. (2018)] focused on the analysis of sequential sentiment based

[1] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, 211106, China.

[2] Key Laboratory of Safety-Critical Software, Ministry of Industry and Information Technology, Nanjing, 211106, China.

[3] Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing, 210046, China.

[4] School of Information Technology and Electrical Engineering, University of Queensland, Queensland, Australia.

[*] Corresponding Author: Bohan Li. Email: bhli@nuaa.edu.cn.

Received: 31 January 2020; Accepted: 01 March 2020.

on a million-level Chinese micro-blog corpora to mine sequential sentiment features. Yan et al. [Yan, Rosales, Fung et al. (2011)] employed the logistic regression model to select samples with a predicted probability around 0.5. Rodrigues et al. [Rodrigues, Pereira and Ribeiro (2014)] incorporated Gaussian process into active learning and measured instance uncertainty. Zhong et al. [Zhong, Tang and Zhou (2015)] constructed a SVM classifier with RBF kernel to verify the effect of labeling confidence. Yue et al. [Yue, Zuo, Tao et al. (2015)] combined vector space model (VSM), and singular value decomposition (SVD) for the document classification. Hu et al. [Hu, Mac Namee and Delany (2016)] explored the reusability of training examples with naive Bayes, SVM and k-NN in text classification scenarios of AL. Wang et al. [Wang, Chen, Li et al. (2019)] applied Dynamic Bayesian Network (DBN) to Recurrent Neural Network (RNN) units to discover the similarity in time serial data analytics.

However, compared to hand-designed features, representations appeared better in deep neural networks (DNN). Zhang et al. [Zhang, Li, Wan et al. (2019)]. classified the cyberbullying text with bidirectional recurrent neural network and attention mechanism. Xu et al. [Xu, Zhang, Xin et al. (2019)] utilized convolutional neural networks to investigate the Chinese text sentiment. Zhang et al. [Zhang, Lease and Wallace (2017)] first applied AL in convolutional neural networks (CNNs) for text classification. They emphasized the significance of the update on the word embedding gradient when selecting instances. Shen et al. [Shen, Yun, Lipton et al. (2017)] used a lightweight architecture CNN-CNN-LSTM to speed up iterative retraining. Furthermore, Wang et al. [Wang, Zhang, Li et al. (2016)] incorporated deep convolutional neural networks into AL for image classification. Gal et al. [Gal, Islam and Ghahramani (2017)] combined active learning with Bayesian deep learning to process high dimensional data. Feng et al. [Feng, Liu, Kao et al. (2017)] detected civil infrastructure defects by applying a deep active learning system.

Followed by them, we checked the effectiveness of AL for classifying text with a deep neural network model called bidirectional encoder representations from transformers (BERT). The core of BERT was the Transformer structure, which depended only on the multi-head self-attention to capture contextual information instead of learning the left-to-right and right-to-left sequence representation like RNN. Although CNN performed well in parallel computing, it failed to notice distant features as BERT. Considering the advantages of BERT on text representation, we took it as the learner of active learning, and explored the instance selection strategy based on it. To summarize, we made the following contributions:

(1) We propose a deep active learning model by introducing deep neural network BERT into AL framework. Since BERT has been pre-trained, we focused on the benefits of active learning in the fine-tuning phase for downstream subtasks.

(2) We design a sampling strategy MII, combining uncertainty with instance correlation to actively select instances. Minimizing the margin in two posterior classification probabilities is used to measure the uncertainty. The metric of instance correlation included intra-correlation of the instance and inter-correlation between instances.

Tab. 1 lists all of the notations and descriptions. The rest of this paper is organized as follows. Section 2 briefly overviews the advanced sampling strategies of active learning

for text classification. Section 3 introduces some basic knowledge of AL and BERT. Then we propose a deep active learning model, and design an instance selection method MII to find informative instances. Experiments are carried out in Section 4 to verify the effectiveness of our method. Finally, we conclude the paper in Section 5.

Table 1: Notations and descriptions

Notations	Descriptions
$\hat{C}$	The most probable class for the data x
L	The training dataset
$f_u(x)$	The uncertainty function of data x
$q_c(x)$	The correlation function of data x
P^L	The labeled data pool
M	The training model
p^U	The unlabeled data pool
I	The selected instances
X	The word embedding of each layer
Q	A query matrix derived from X
K	A key matrix derived from X
V	A value matrix derived from X
W^q	The weight of transforming X into Q
W^k	The weight of transforming X into K
W^v	The weight of transforming X into V
A_i	The i^{th} attention head
Z	The output that covers bidirectional words information
AM_i	The attention matrix of the head i
S_i	The sum of attentions for each word over the head i
X_{LL}	The text representation of X at the last encoding layer of the deep neural network
X_{LL}^c	The classification center

2 Related work

Active learning is meant to obtain the expected learning model at little cost. The core of it is to develop the sampling standard, which selects part of the data for labeling. The study about instance selection began with Sheng et al. [Sheng, Provost and Ipeirotis (2008)].

They assessed the impact of duplicate labels on data quality. Fu et al. [Fu, Zhu and Li (2013)] summarized the criteria for instance selection and divided it into two categories: the uncertainty and the correlation of independent identical distribution (IID) instances. Most of the work related to the measurement of uncertainty involves some common strategies such as label uncertainty, model uncertainty, and mixed uncertainty. Based on these uncertainty measurements, Zhang et al. [Zhang, Wu and Shengs (2014)] proposed three new instance selection strategies, including MLSI, CMPI and CFI. These strategies were combined with the thresholds of the dynamically adjusted labels to improve learning performance. Based on the bootstrap theory, Mozafari et al. [Mozafari, Sarkar, Franklin et al. (2014)] then took the current quality and uncertainty into account, selecting samples in minimum expected error and uncertainty.

Several popular strategies of AL for text classification were categorized from the perspective of individual instance and expected model.

Uncertainty sampling The basic principle of uncertainty sampling is to give priority to the sample data that cannot be accurately judged by the current classifier, which is usually located near the classification boundary. This strategy applied probabilistic models based on the current hypothesis to select uncertain instances. It included three measures: least confidence (LC), margin sampling and entropy. A general LC method followed by Lewis et al. [Lewis and Gale (1994)] was to realize

$$x_{LC}^{*} = \arg\max_{x} 1 - P_{\theta}\left(\hat{C} \mid x\right) \tag{1}$$

$\hat{C}$ is the most probable class for x. But it ignored the information from other labels. Scheffer et al. [Scheffer, Decomain and Wrobel (2001)] then introduced margin sampling by integrating multi-class uncertainty. They tried to minimize the margin between two most probable classes $\hat{C}_1$ and $\hat{C}_2$ to find ambiguous x.

$$x_{M}^{*} = \arg\min_{x} P_{\theta}\left(\hat{C}_1 \mid x\right) - P_{\theta}\left(\hat{C}_2 \mid x\right) \tag{2}$$

In addition, another common measure called entropy [Shannon (1948)] evaluated the information that we did not know. It was the expectation of the amount of information over entire uncertain classes C_i.

$$x_{E}^{*} = \arg\max_{x} -\sum_{i} P_{\theta}\left(C_i \mid x\right) \log P_{\theta}\left(C_i \mid x\right) \tag{3}$$

Zhu et al. [Zhu, Wang, Tsou et al. (2010)] took density into account based on entropy, realizing uncertainty sampling on the tasks of text classification and word sense disambiguation.

Expected gradient length (EGL) This strategy was proposed by Settles et al. [Settles, Craven and Ray (2008)]. Expected model change was reflected in the impact on model parameters. The criterion of selecting instances is as follows:

$$x_{EGL}^{*} = \arg\max_{x} \sum_{i} P_{\theta}\left(C_i \mid x\right)\left\|\nabla f_{\theta}\left(L \cup \langle x, C_i \rangle\right)\right\| \tag{4}$$

f_{θ} is a learner based on gradient and L is the training dataset. $\left\|\nabla f_{\theta}\left(L \cup \langle x, C_i \rangle\right)\right\|$ represents the length of the model gradient in the Euclidean space after adding the new sample. Zhang et al. [Zhang, Lease and Wallace (2017)] applied variants of EGL in both sentence and document classification tasks. Especially, they combined EGL and entropy to form a beta model when classifying long text. Long et al. [Long, Bian, Chapelle et al. (2015)] utilized expected loss optimization (ELO) with discounted cumulative gain (DCG) for web search ranking.

Instance correlation This scheme uses feature-based, label-based and graph-based correlation among instances to determine the most informative samples. The basic idea is to define a utility function that combines uncertainty $f_u\left(x\right)$ and correlation $q_c\left(x\right)$.

$$x^{*} = \arg\max_{x} f_u\left(x\right) \times q_c\left(x\right) \tag{5}$$

Sun et al. [Sun and Hardoon (2010)] mapped the features into a coordinate system with canonical correlation analysis (CCA), and then measured the similarity between unlabeled data and original labeled data. Common similarity functions are cosine similarity, KL divergence similarity and Gaussian similarity. Chen et al. [Chen and Mani (2010)] presented a modified information density method to identify the correlation of instances and the mean point.

Compared with the three kinds of state-of-the-art approaches that were mentioned above, our method focused on the scenario of DNN, and gave priority to the output of the last encoder. Instead of relying on a single strategy, we simultaneously considered uncertainty and correlation for sampling.

3 Deep active learning model

3.1 Problem description

Due to the absence of sufficient label data, it is difficult to guarantee the effect of text classification. Given an unlabeled dataset, supervised learning asks for labeling the unlabeled data manually to form a labeled dataset, which participates in the training process of the learner. The more labeled data required for training, the higher the labor cost. Since active learning determines the data to be labeled by the learner, after manual labeling, it is sent to the learner for training to improve the performance of the learner. The difference between these two methods is shown in Fig. 1.

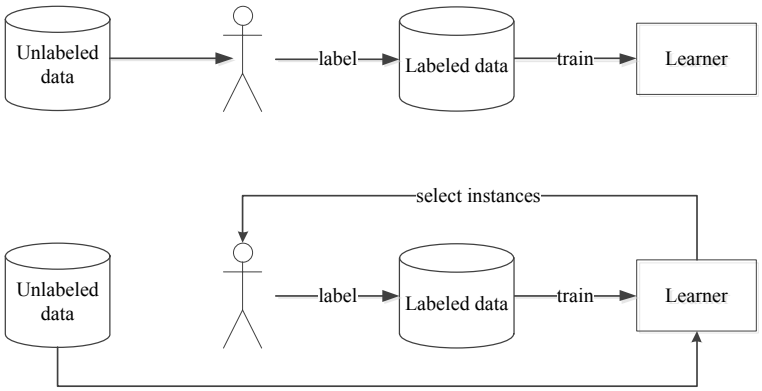

Figure 1: The general process of supervised learning and active learning

Active learning allows a model to actively query data that is most helpful to improve task performance in the case of limited labeled data. As shown in Algorithm 1, the process begins with the process of training a small amount of data in the labeled data pool P^L to initialize a model M. Then M is applied to select some instances I from the unlabeled data pool P^U. Labeled by specific annotators, I is combined with label L to form an instance-label pair $\langle I, L \rangle$. After that, new labeled data $\langle I, L \rangle$ is added to P^L. A new cycle starts here. Finally, the expected performance is achieved at a minimal labeling cost.

Algorithm 1: General Active Learning Process	
1	**repeat**
2	$M \leftarrow$ Train labeled data in P^L
3	$I \leftarrow$ Select unlabeled data from P^U
4	$\langle I, L \rangle \leftarrow$ Label I by annotators
5	$P^L \leftarrow$ Add $\langle I, L \rangle$ to P^L
6	**until** expected performance with minimal cost

3.2 BERT text representation

A latest model BERT proposed by Devlin et al. [Devlin, Chang, Lee et al. (2018)] has shown its advantages on many NLP tasks, such as semantic understanding and inference. This kind of generative performance benefits from pre-training based on a large scale of corpus and multi-headed attention mechanism [Vaswani, Shazeer, Parmar et al. (2017)] for deep contextual representations. By incorporating self-attention into word embedding, the model encodes words within the internal relationship.

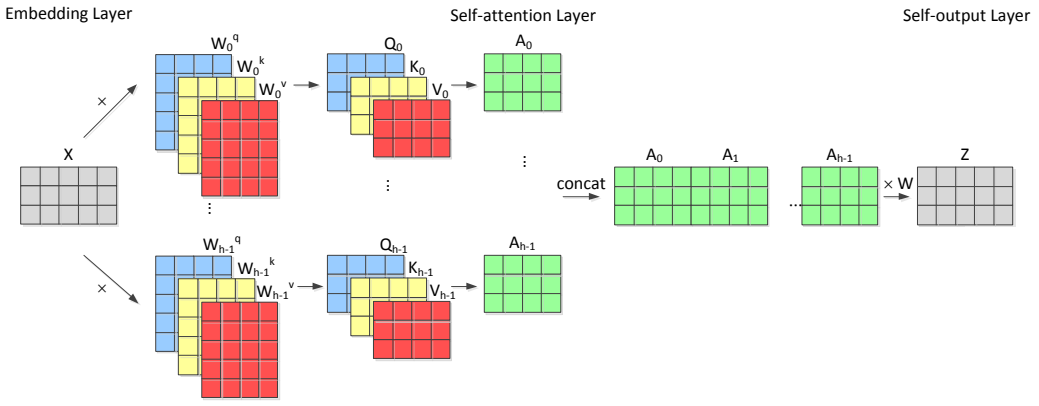

Figure 2: Self-attention mechanism

Fig. 2 presents the self-attention mechanism. According to the encoder principle of the Transformer network architecture, word embedding of each layer $X \in R^{d_{vocab} \times d_{hidden}}$ are transformed into three matrixes, query $Q \in R^{d_{vocab} \times d_q}$, key $K \in R^{d_{vocab} \times d_k}$ and value $V \in R^{d_{vocab} \times d_v}$, where $d_q = d_k = d_v$. They are obtained by multiplying X with corresponding weights W.

$$Q = X \times W^q, K = X \times W^k, V = X \times W^v \tag{6}$$

To measure the attention given to other words, a score is computed by dotting the query vector for current word and the key vector for the word to be scored. The scaled dot-product attention is then computed as following formula. Meanwhile, varying Q, K and V enables the model to find multiple words that have a significant impact on current word encoding. After integrating all the attention heads $A_0, A_1, \ldots A_{h-1}$, the output Z that covers bidirectional words information is obtained.

$$A_i = soft\max\left(\frac{Q_i \times K_i^T}{\sqrt{d_k}}\right) \cdot V_i, i = 0,1,\ldots,h-1 \tag{7}$$

$$Z = concat\left(A_0, A_1, \ldots A_{h-1}\right) \times W \tag{8}$$

3.3 Active learning with BERT

Fig. 3 shows the process of active learning in deep neural network BERT. This framework starts with collecting few labeled data as well as a large amount of unlabeled data. Then labeled data is used to initialize the BERT model. BERT adopts the structure of Transformer, which includes multiple encoded layers. We make use of the last encoded layer to mine the intra-correlation and inter-correlation among instances. Meanwhile, a margin of probabilities for two classes is combined to form a utility function. This function is then applied in selecting informative data from unlabeled data pool P^U to supplement the training data. The loop continues until the desired

classification performance is achieved.

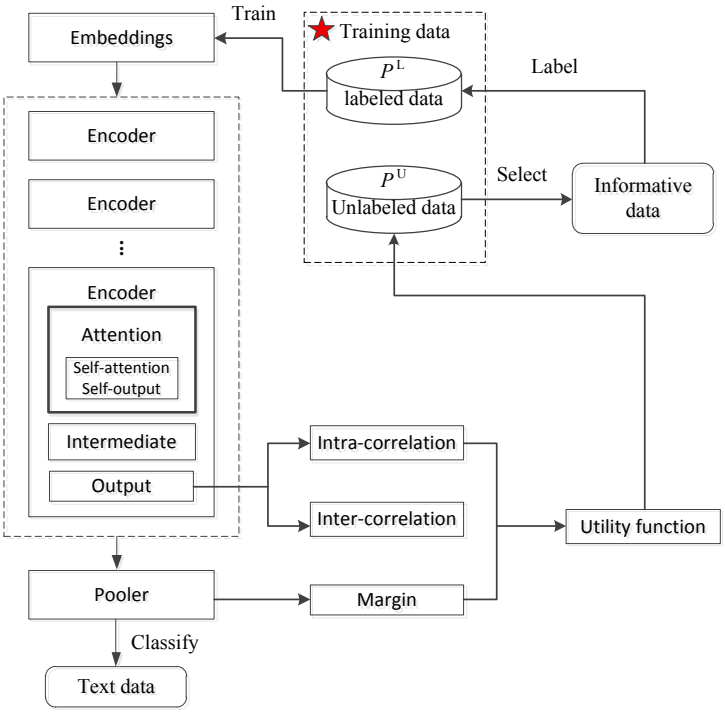

Figure 3: Active learning framework based on BERT

The classification part of the deep active learning framework is based on the deep text representation model BERT and the instance selection mechanism. BERT obtains a text representation that incorporates context semantics through the cascading encoder. The utility function combining intra-correlation, inter-correlation and posterior probability margin is used to select unlabeled data. After manual labeling, unlabeled data is converted into labeled data.

3.4 Instance selection strategy

With preliminary understanding of the above framework, we further design the sampling mechanism of deep active learning. Inspired by the method based on the instance correlation, we propose a sampling approach MII, combining posterior probabilities margin, the intra-correlation and the inter-correlation. We aim to minimize a utility function to select instances from both uncertainty and correlation views. As for uncertainty measure, we use the minimum margin between posterior probabilities of two class labels. Different from traditional feature correlation measure, we explore correlations within and between instances respectively in the last self-attention layer of BERT.

Definition 1 (Intra-correlation) Let AM_i be the attention matrix of the head i, which reflects the effect of all the words in the text on updating a word representation. The intra-correlation is to measure the interaction of words in each text by jointly considering

$AM_0, AM_1, \ldots AM_{h-1}$. Here n is equal to the dimension of the word in the text.

$$AM_i = \begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \ldots & a_{nn} \end{pmatrix}_{d_{vocab} \times d_{vocab}} \tag{9}$$

Firstly, we compute the sum of attentions for each word by summing over the rows of AM_i, and get S_i. Then the result $R_{intra}(X)$ is the mean of all the heads' variances.

$$S_i = (s_1, \ldots s_n)^T = \begin{bmatrix} a_{11} + \ldots + a_{1n} \\ \vdots \\ a_{n1} + \ldots + a_{nn} \end{bmatrix} \tag{10}$$

$$\bar{s} = \frac{1}{n} \sum_{j=1}^{n} s_j \tag{11}$$

$$R_{intra}(X) = \frac{1}{h} \sum_{i=0}^{h-1} \left(\frac{1}{n} \sum_{j=1}^{n} \left(s_j - \bar{s} \right)^2 \right) \tag{12}$$

Definition 2 (Inter-correlation) Given a word embedding X of unlabeled data in the pool P^U, X_{LL} is the text representation of X at the last encoding layer of the deep neural network. The inter-correlation is meant to compute the cosine similarity between X_{LL} and two classification centers $X_{LL}^{c_1}$, $X_{LL}^{c_2}$.

$$R_{inter}\left(X_{LL}, X_{LL}^{c}\right) = \max\left(\frac{X_{LL} \cdot X_{LL}^{c_1}}{\|X_{LL}\| \times \|X_{LL}^{c_1}\|}, \frac{X_{LL} \cdot X_{LL}^{c_2}}{\|X_{LL}\| \times \|X_{LL}^{c_2}\|} \right) \tag{13}$$

In combination with the above definitions, selected instances satisfy the objective function

$$X_{sel} = \arg\min_{X} \left\{ f_u(X) \times R_{intra}(X) - R_{inter}\left(X, X^c\right) \right\} \tag{14}$$

where $f_u(X) = \left| p(C_1 \mid X) - p(C_2 \mid X) \right|$. The intuition behind it includes two aspects. From the perspective of a single instance, small $f_u(X)$ and $R_{intra}(X)$ reflect large classification uncertainty and sparse semantic structure. In terms of instance feature correlation, samples similar to clustering centers are probably redundant information.

4 Experiment

4.1 Data preparation

We conduct experiments on three real-world datasets, and compare the presented methods

with state-of-the-art solutions. The first one is sentence-based, and the remaining two are document-based. All of them are comments with sentimental polarity. The purpose is to accurately divide each corpus into two categories with limited labeled data.

SST-2 This corpus is derived from the movie review dataset provided by Pang et al. [Pang and Lee (2005)], consisting of 11,855 single sentences for binary classification tasks. Each sentence is labeled with human sentiment.

Amazon This dataset covers the comments of people on products with a five-point scale, including 1,800,000 training data and 200,000 testing data. Comments of 1 and 2 are considered negative, and comments of 4 and 5 are considered positive. The review score of 3 is ignored.

Yelp Yelp is a platform for people to review various business service. Similar to the five-point principle of the Amazon dataset, 560,000 training data is divided into negative class 1 and positive class 2.

We randomly choose 1,000 samples from each of the above datasets as their initial training sets, and 800 samples as testing sets. The remaining are taken as their unlabeled sets. The average accuracy after 10-fold cross-validation is reported. We verify the experimental effect of active learning method with MII as the sampling strategy and BERT as the learner from two aspects. The first part compares the effect of different learners on active learning text classification under the premise of MII sampling strategy, while the second part analyzes the influence of different instance selection methods on the convergence speed of the algorithm when the learner (BERT) is fixed.

4.2 The effect of learner

We compare the performance of MII applied to different classifiers over three kinds of datasets. The basic version of BERT model is taken as the learner in the experiment. The network structure includes 12 layers with 768 hidden layer units and 12 self-attention heads. Fig. 4 compares the accuracy of four classifiers that introduce MII into model training. It can be seen that DNN-based approaches such as BERT and CNN outperform the traditional methods like logistic regression and SVM. To be specific, the accuracy of BERT and CNN ranges between 70% and 88%, while that of logistic regression and SVM falls into the range of 55%-75%. With the given classifier of MII, BERT performs slightly better than CNN. This is since that BERT combines the context of all layers to pretrain deep bidirectional text representation. The core is to focus multiple concerns at the same time with the help of Transformer structure to accurately learn the meaning of statements. Overall, deep neural network BERT assists sampling strategy MII in obtaining the informative instances. The combination of BERT and MII can provide a better text representation than other methods.

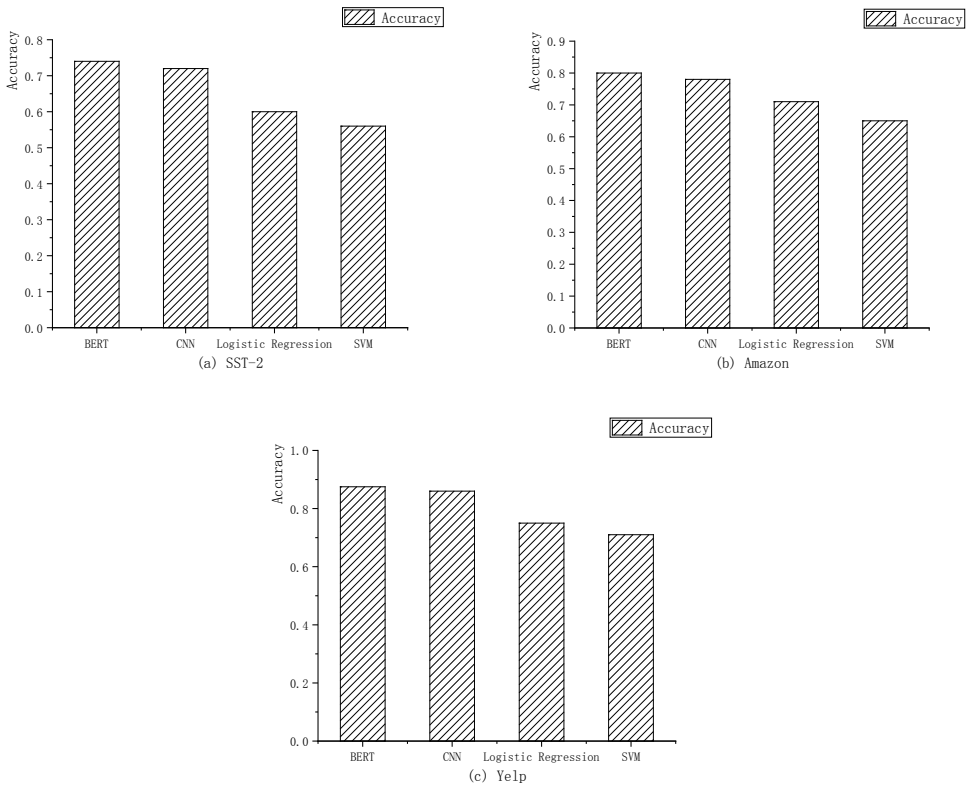

Figure 4: The accuracy of four classifiers under instance selection strategy MII

4.3 The effect of sampling strategy

The sampling strategy determines whether the selected samples are representative or not and whether contain positive information for correct text classification. By manually labeling these samples, the expected accuracy can be obtained at a small labeling cost. This part verifies the influence of different sampling strategies on deep active learning text classification, and gives the AUC (area under ROC curve) results of different experimental methods. The range of AUC is [0.5, 1]. The higher the value of AUC is, the better the classifier under the given sampling mode performs. Tab. 2 shows AUC scores of the four methods on the three datasets. Compared with baselines such as random, entropy and EGL, our proposed method MII scores a little higher. Random selection is the least effective way, while the performance of EGL is just second to our method, with AUC value approximately between 0.7 and 0.86. All methods perform better on Yelp dataset than on SST-2 and Amazon datasets.

Table 2: AUC comparison of classification model with BERT and four sampling strategies

	MII	Random	Entropy	EGL
SST-2	0.717	0.683	0.692	0.708
Amazon	0.752	0.721	0.734	0.745
Yelp	0.863	0.837	0.841	0.858

To verify the impact of active learning, each learner selects 25 samples from the unlabeled pool P^U at one time, and this process iterates 20 times. Fig. 5 demonstrates increasing tendency of accuracy as the growing number of selected instances. The sampling strategy helps classifier achieve higher accuracy with fewer samples compared with random process. Especially, as for SST-2, only 100 samples are needed to reach the accuracy of 0.675. On the whole, experiments on the document level outperformed that on sentence level. It makes sense that short sentence contains less information than long text. Poor embedding integration leads to difficulty in correlation measurement. To be specific, MII converges more rapidly than benchmark methods on all the datasets. It is consistent with the characteristics of active learning. The final accuracy of it falls into the range of 74%-87%. Although the curves of accuracy are not smooth on some datasets, the overall trend is acceptable.

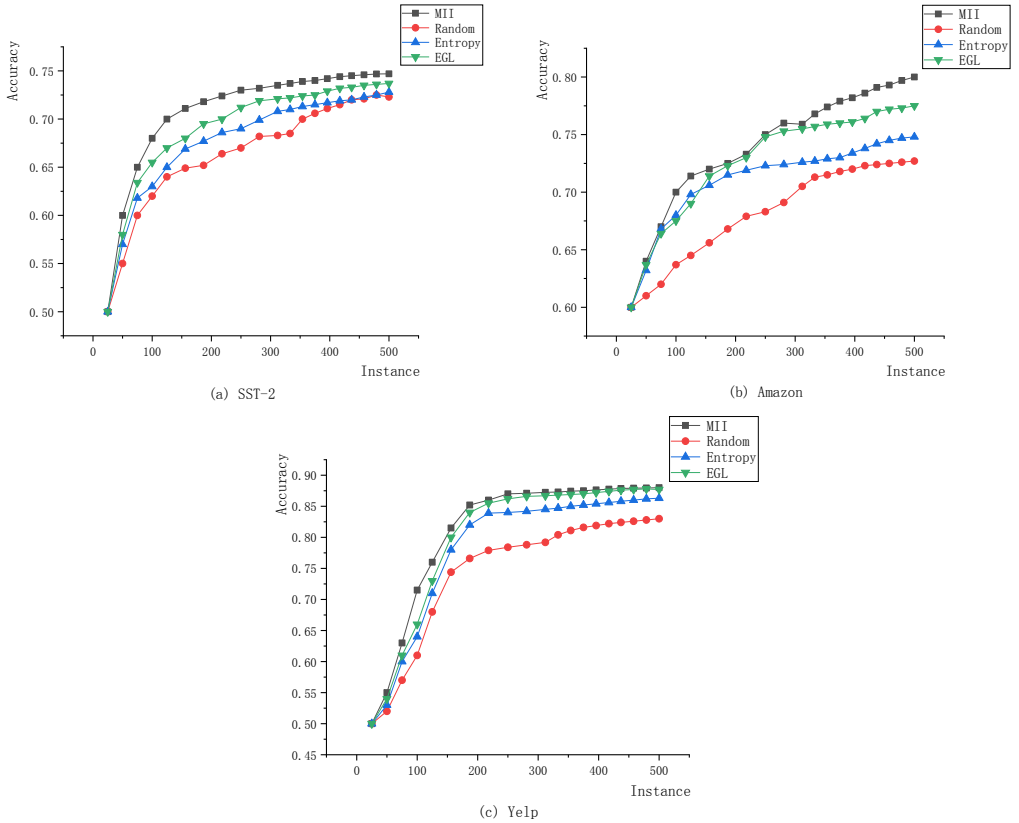

Figure 5: The accuracy convergence of instance selection in BERT

The following experiment studies the influence of MII sampling strategy and its components on the final classification accuracy in deep active learning framework, thus analyzing the advantages of MII. According to Tab. 3, vertically, MII is more favorable to improve the accuracy of classification model. The effect of using posterior probability margin is second only to the proposed method. It indicates that as one of components of MII, the posterior probability margin has great contribution to combination method, while the separated intra-correlation and inter-correlation are not effective when they are used alone. Horizontally, the experiment performs better on Yelp than on the other two datasets. Experimental results show that exploiting the uncertainty and correlation of samples is an effective way for instance selection in deep active learning.

Table 3: The effect of MII and its components on classification accuracy

	MII	Margin	Intra-correlation	Inter-correlation
SST-2	0.720	0.715	0.654	0.688
Amazon	0.756	0.742	0.671	0.692
Yelp	0.868	0.846	0.693	0.710

4.4 Text classification results

The above two experimental results show that BERT has advantages in classification performance, and the combination of uncertainty and correlation sampling is an effective method for instance selection in deep active learning. This part compares the proposed deep active learning text classification method (BERT+MII) with mainstream text classification methods. Tabs. 4, 5 and 6 compare text classification methods in terms of accuracy, precision and recall in three datasets respectively. It can be seen that deep learning (DL) method is generally better than traditional machine learning (ML) method. The proposed text classification method BERT+MII performs best with classification accuracy between 82% and 86% respectively. Also the experiment performs best on Yelp because of its moderate text length, and deep model can integrate context to represent text.

Table 4: Evaluation of mainstream text classification model (SST-2 dataset)

Type	Method	Accuracy	Precision	Recall
ML	KNN(k=1)	0.830	0.824	0.830
	Random Forest	0.821	0.815	0.822
	Naïve Bayes	0.814	0.808	0.814
	SVM	0.716	0.711	0.711
	Logistic Regression	0.710	0.704	0.704
DL	CNN	0.824	0.820	0.816
	LSTM	0.807	0.802	0.793
	BERT+MII	0.836	0.829	0.831

Table 5: Evaluation of mainstream text classification model (Amazon dataset)

Type	Method	Accuracy	Precision	Recall
ML	KNN(k=1)	0.820	0.814	0.820
	Random Forest	0.735	0.716	0.733
	Naïve Bayes	0.748	0.737	0.746
	SVM	0.814	0.812	0.810
	Logistic Regression	0.812	0.810	0.810
DL	CNN	0.816	0.813	0.806
	LSTM	0.809	0.814	0.804
	BERT+MII	0.825	0.820	0.822

Table 6: Evaluation of mainstream text classification model (Yelp dataset)

Type	Method	Accuracy	Precision	Recall
ML	KNN(k=1)	0.841	0.836	0.840
	Random Forest	0.829	0.823	0.827
	Naïve Bayes	0.818	0.812	0.815
	SVM	0.797	0.792	0.794
	Logistic Regression	0.783	0.776	0.781
DL	CNN	0.837	0.831	0.833
	LSTM	0.812	0.810	0.811
	BERT+MII	0.853	0.849	0.851

5 Conclusion

In this paper, we explore the effectiveness of active learning in deep neural network BERT. We combine margin, intra-correlation and inter-correlation to design a novel instance selection method MII. This approach is meant to select instances with high uncertainty and low correlation. By adding these informative samples to the training dataset, the accuracy of model converges within limited steps. However, there is still a lack of a unified method for selecting instances. In the future, we plan to rely on the structure of neural network to train a general mechanism for AL instead of designing specific approaches.

Funding Statement: This work is supported by National Natural Science Foundation of China (61402225, 61728204), Innovation Funding (NJ20160028, NT2018028, NS2018057), Aeronautical Science Foundation of China (2016551500), State Key Laboratory for smart grid protection and operation control Foundation, and the Science and Technology Funds from National State Grid Ltd., China degree and Graduate Education Fund.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Chen, Y.; Mani, S. (2010): Study of active learning in the challenge. *International Joint Conference on Neural Networks*, pp. 1-7.

Devlin, J.; Chang, M. W.; Lee, K.; Toutanova, K. (2018): Bert: pre-training of deep bidirectional transformers for language understanding. arXiv: 1810.04805.

Feng, C.; Liu, M. Y.; Kao, C. C.; Lee, T. Y. (2017): Deep active learning for civil infrastructure defect detection and classification. *Computing in Civil Engineering*, pp. 298-306.

Fu, Y.; Zhu, X.; Li, B. (2013): A survey on instance selection for active learning. *Knowledge and Information Systems*, vol. 35, no. 2, pp. 249-283.

Gal, Y.; Islam, R.; Ghahramani, Z. (2017): Deep Bayesian active learning with image data. *Proceedings of the 34th International Conference on Machine Learning*, pp. 1183-1192.

Hu, R.; Mac Namee, B.; Delany, S. J. (2016): Active learning for text classification with reusability. *Expert Systems with Applications*, vol. 45, pp. 438-449.

Lewis, D. D.; Gale, W. A. (1994): A sequential algorithm for training text classifiers. *SIGIR'94*, pp. 3-12.

Long, B.; Bian, J.; Chapelle, O.; Zhang, Y.; Inagaki, Y. et al. (2015): Active learning for ranking through expected loss optimization. *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 5, pp. 1180-1191.

Mozafari, B.; Sarkar, P.; Franklin, M.; Jordan, M.; Madden, S. (2014): Scaling up crowd-sourcing to very large datasets: a case for active learning. *Proceedings of the VLDB Endowment*, vol. 8, no. 2, pp. 125-136.

Pang, B.; Lee, L. (2005): Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 115-124.

Rodrigues, F.; Pereira, F.; Ribeiro, B. (2014): Gaussian process classification and active learning with multiple annotators. *Proceedings of the 31st International Conference on Machine Learning*, pp. 433-441.

Scheffer, T.; Decomain, C.; Wrobel, S. (2001): Active hidden markov models for information extraction. *International Symposium on Intelligent Data Analysis*, pp. 309-318.

Settles, B.; Craven, M.; Ray, S. (2008): Multiple-instance active learning. *Advances in Neural Information Processing Systems*, pp. 1289-1296.

Shannon, C. E. (1948): A mathematical theory of communication. *Bell System Technical Journal*, vol. 27, no. 3, pp. 379-423.

Shen, Y.; Yun, H.; Lipton, Z. C.; Kronrod, Y.; Anandkumar, A. (2017): Deep active learning for named entity recognition. arXiv: 1707.05928.

Sheng, V. S.; Provost, F.; Ipeirotis, P. G. (2008): Get another label improving data quality and data mining using multiple, noisy labelers. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 614-622.

Sun, S.; Hardoon, D. R. (2010): Active learning with extremely sparse labeled examples. *Neurocomputing*, vol. 73, no. 16-18, pp. 2980-2988.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L. et al. (2017): Attention is all you need. *Advances in Neural Information Processing Systems*, pp. 5998-6008.

Wan, S.; Li, B.; Zhang, A.; Wang, K.; Li, X. (2018): Vertical and sequential sentiment analysis of micro-blog topic. *International Conference on Advanced Data Mining and Applications*, pp. 353-363.

Wang, K.; Zhang, D.; Li, Y.; Zhang, R.; Lin, L. (2016): Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2591-2600.

Wang, Y.; Chen, W.; Li, B.; Boots, R. (2019): Learning fine-grained patient similarity with dynamic bayesian network embedded RNNs. *International Conference on Database Systems for Advanced Applications*, pp. 587-603.

Xu, F.; Zhang, X.; Xin, Z.; Yang, A. (2019): Investigation on the Chinese text sentiment analysis based on convolutional neural networks in deep learning. *Computers, Materials & Continua*, vol. 58, no. 3, pp. 697-709.

Yan, Y.; Rosales, R.; Fung, G.; Dy, J. G. (2011): Active learning from crowds. *Proceedings of the 28th International Conference on Machine Learning*, pp. 1161-1168.

Yue, L.; Chen, W.; Li, X.; Zuo, W.; Yin, M. (2018): A survey of sentiment analysis in social media. *Knowledge and Information Systems*, pp. 1-47.

Yue, L.; Zuo, W.; Tao, P.; Wang, Y.; Han, X. (2015): A fuzzy document clustering approach based on domain-specified ontology. *Data & Knowledge Engineering*, vol. 100, pp. 148-166.

Zhang, A.; Li, B.; Wan, S.; Wang, K. (2019): Cyberbullying detection with BiRNN and attention mechanism. *International Conference on Machine Learning and Intelligent Communications*, pp. 623-635.

Zhang, J.; Wu, X.; Shengs, V. S. (2014): Active learning with imbalanced multiple noisy labeling. *IEEE Transactions on Cybernetics*, vol. 45, no. 5, pp. 1095-1107.

Zhang, Y.; Lease, M.; Wallace, B. C. (2017): Active discriminative text representation learning. *Thirty-First AAAI Conference on Artificial Intelligence*, pp. 3386-3392.

Zhong, J.; Tang, K.; Zhou, Z. H. (2015): Active learning from crowds with unsure option. *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pp. 1061-1067.

Zhu, J.; Wang, H.; Tsou, B. K.; Ma, M. (2010): Active learning with sampling by uncertainty and density for data annotations. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1323-1331.