Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

MAL: multilevel active learning with BERT for Chinese affective structure analysis*

Shufeng XIONG†, Guipei ZHANG, Xiaobo FAN, Wenjie TIAN, Lei XI, Hebing LIU, Haiping SI†‡

Henan Agricultural University, Zhengzhou 450002, China

†E-mail: xsf@whu.edu.cn; haiping@henau.edu.cn

Received Mar. 31, 2024; Revision accepted Sept. 22, 2024; Crosschecked

Abstract: Chinese textual affective structure analysis is a sequence labeling task that often relies on supervised deep learning methods. However, acquiring a large annotated dataset for training can be expensive and time-consuming. Active learning offers a solution by selecting the most valuable samples to reduce labeling costs. Previous approaches have focused on uncertainty or diversity but faced challenges such as biased models or selecting insignificant samples. To address these issues, this paper introduces multilevel active learning (MAL), which leverages the power of deep textual information at both the sentence and word levels, taking into account the complex structure of the Chinese language. By integrating the sentence-level features extracted from BERT embeddings and the word-level probability distributions obtained through a CRF model, MAL comprehensively captures the affective structure of Chinese text. Experimental results demonstrate that MAL significantly reduces annotation costs by approximately 70% and achieves more consistent performance compared to baseline strategies.

Key words: Sentiment analysis; Sequence labeling; Active learning; BERT
https://doi.org/10.1631/FITEE.2400242 **CLC number:** TP

1 Introduction

The volume of data on social media is experiencing exponential growth, and numerous business applications have reaped the benefits of this increased data power, particularly in sentiment analysis (Medhat et al., 2014; Venugopalan and Gupta, 2015; Alamoodi et al., 2021; Basiri et al., 2021). The initial step prior to conducting sentiment analysis involves identifying and labeling the sentiment-related terms within the target text, known as textual affective structure identification (TASI), which essentially falls under the purview of sequence labeling tasks. TASI aims to extract complete sentiment tuples from sentences. This task is an 8-tuple identification task, which involves marking eight categories of target spans related to sentiment descriptions in the text. These eight elements include cause, degree, holder, negation, property, trigger, compared_entity, and sent_entity, with detailed explanations provided in Table 1. Our task setting is focused on affective structure analysis in Chinese, specifically analyzing the affective structures expressed in Chinese text. Machine learning (Bishop, 2006) is frequently employed for sequence labeling tasks, and this supervised learning approach necessitates a substantial amount of high-quality training data to develop a robust classifier. However, acquiring such data requires significant investments in terms of human resources with domain expertise, and the resulting outcomes may not always meet expectations. Hence, it becomes crucial to filter out samples that hold higher value for labeling purposes.

Active learning has demonstrated its effectiveness in reducing labeling costs by selecting the most valuable samples from a pool of unlabeled data. It

‡ Corresponding author

* Project supported by the MOE (Ministry of Education of China) Project of Humanities and Social Sciences (No. 19YJCZH198), and Henan Province key research and development project, China (No. 231111211300)

ORCID: Shufeng XIONG, https://orcid.org/0000-0001-5727-1766

has been widely employed in various NLP tasks, including text classification (Hu et al., 2016), biomedical text mining (Zhang HT et al., 2012), clinical annotation (Chen et al., 2015), and sentiment analysis (Smailović et al., 2014). The crux of active learning lies in designing appropriate query functions. Currently, the prevailing active learning query strategies comprise uncertainty-based, diversity-based, and hybrid acquisition approaches. However, uncertainty-based query strategies often overlook challenging yet crucial samples for the model, leading to data bias issues. Conversely, diversity-based query strategies may select meaningless samples, resulting in wastage of resources. To overcome these limitations, the hybrid strategy combines the strengths of uncertainty-based and diversity-based approaches while incorporating information richness. By integrating uncertainty, diversity, and information richness, the hybrid strategy provides a more comprehensive framework for guiding the model's learning process. Our approach falls into this category.

In Chinese texts, particularly in social texts where there are no unified and strict norms, the same emotion can be expressed in multiple ways. For example, consider the sentence: "谁能给我推荐几本书呢？准备下了丧课看，内实在太痛苦鸟！一定要想个办法～" (Who can recommend a few books to me? Ready to read after the mourning class, inside is too painful bird! Must think of a way), In this sentence, the writer expresses their feelings of pain and seeks help. However, the characters "内" (inside) and "鸟" (bird) are irregular expressions of "那" (that) and "了" (modal particle), respectively. From this example, it can be observed that key emotional information is manifested through the informal application of Chinese vocabulary. Therefore, when processing Chinese, it is essential to consider both selecting sentence samples with higher annotation value and the diversity of vocabulary at the word level in active learning strategies. Existing query strategies typically only consider single information from either the sentence level or the word level when selecting annotated samples, neglecting the other type of information. The experimental result in section 6 also indicates that the performance of different baseline methods using a single query strategy is unsatisfactory.

Moreover, it is crucial to note that finding a classifier coupled with an active learning strategy is of utmost importance. Although combining a successful AL strategy with a simple Bayesian classifier may yield some positive results, it may not be as effective as using convolutional neural networks, as suggested by Dor et al. (2020). Additionally, the introduction of pre-training models has significantly improved the performance of numerous natural language processing (NLP) tasks as highlighted by Qiu et al. (2020), with BERT (Devlin et al., 2019) receiving significant attention. This can be attributed to BERT's exceptional feature extraction capabilities and its ability to effectively capture sentence-level information. Therefore, our primary objective is to design an effective AL strategy that synergistically aligns with BERT for the recognition of affective structures in Chinese text.

Taking inspiration from query contrastive query samples (Margatina et al., 2021b), we propose a hybrid strategy that synergistically incorporates both sentence- and word-level information. By leveraging the power of these two levels of information in conjunction, our approach enables more precise selection of representative and crucial samples. We introduce a novel approach called multilevel active learning (MAL). The overall architecture of MAL is illustrated in Fig. 1. In essence, MAL aims to select samples for the model that are similar to the annotated samples in terms of sentiment expression and structure, both at the sentence and word levels. In the subsequent sections, we will provide a detailed description of our proposed approach. The experimental results demonstrate that MAL achieves superior performance while requiring less labeling cost compared to the baseline strategies. Our contributions are the following:

1. Our proposed active learning strategy is used for the Chinese textual affective structure recognition task, and this study is the first work to focus on this problem.

2. Our proposed method enhances the model's capability to accurately recognize the affective structure of Chinese text, encompassing both the sentence and word levels.

3. The results obtained on the Chinese weibo dataset validate the effectiveness of MAL, because it consistently outperforms or achieves comparable performance to the baseline methods.

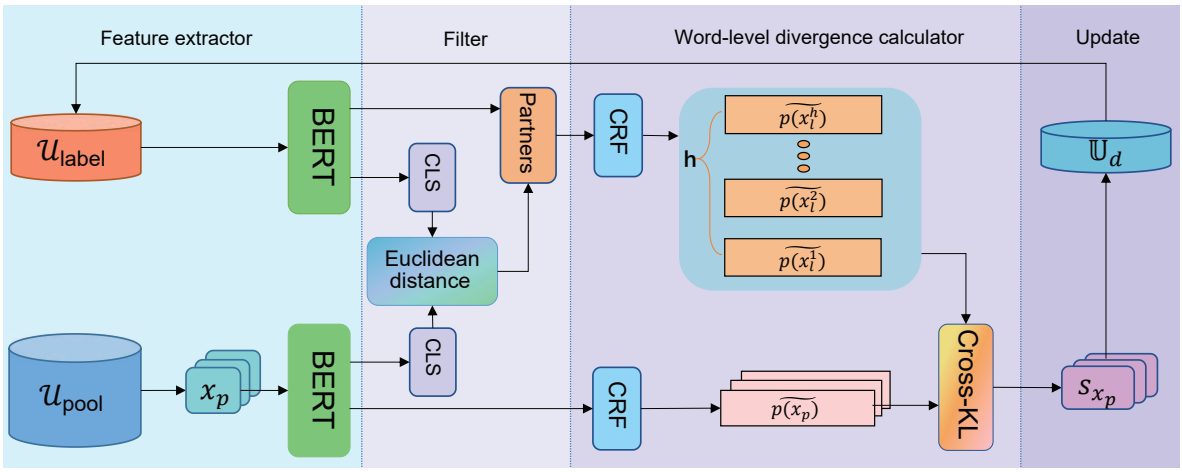


Fig. 1 Structure of multilevel active learning

2 Related works

The effectiveness of active learning has been demonstrated in numerous studies (Settles, 2010; Dasgupta, 2011; Hanneke, 2014). Existing active learning methods can be broadly categorized into pool-based (McCallum and Nigam, 1998; Shen et al., 2017), stream-based (Dagan and Engelson, 1995), and membership query synthesis (Angluin, 1988) approaches. Among these, the pool-based approach has received the most attention. It selects samples from a "pool" and trains them iteratively with labeled samples from earlier rounds, using a query function until a predefined condition is satisfied. Therefore, our focus primarily lies in the design of the query function.

Currently, there is a prevalent use of uncertainty-based query methods (Lewis, 1995; Cohn et al., 1996; Gal et al., 2017; Kirsch et al., 2019; Zhang MK and Plank, 2021), diversity-based query methods (Brinker, 2003; Bodó et al., 2011; Senerand Savarese, 2018), and hybrid methods (Ducoffe and Precioso, 2018; Ash et al., 2020; Yuan et al., 2020; Margatina et al., 2021b) in active learning research. These approaches aim to strike a balance between uncertainty and diversity while selecting informative samples for annotation.

2.1 Uncertainty-based query methods

Uncertainty-based query methods have been widely used in text classification and sequence labeling tasks. One commonly used method is the least confidence (LC) method (Lewis, 1995). It selects data points with the lowest confidence in their most likely labels, targeting the most uncertain samples for annotation. However, the LC method may prioritize longer sequences or more complex labels, potentially neglecting other important samples during training.

To address this limitation, the lowest token probability (LTP) method (Liu MY et al., 2022) was introduced. LTP considers the interrelationships between labels by leveraging the input and output of conditional random fields (CRF). It selects samples with the lowest labeling probability, capturing the model's comprehensive understanding of input sequences and choosing samples with high information content. In contrast to LTP, our approach initially focuses on sample augmentation at the sentence level while also accounting for the diversity of expression at the word level. Specifically, while LTP solely emphasizes sample selection strategies at the word level, our work represents the first active learning strategy designed in conjunction with deep learning that simultaneously considers both sentence- and word-level features, effectively incorporating the unique characteristics of Chinese expressions.

Prediction entropy, such as the token entropy (TE) method (Settles and Craven, 2008), is another uncertainty-based query method. TE calculates the predictive entropy using the model's posterior probabilities for each token, reflecting the uncertainty associated with each token. By selecting samples with the highest prediction entropy, the TE method effectively chooses the samples that are the most perplexing to the model.

Bayesian active learning by disagreement

(BALD) (Houlsby et al., 2011) is a notable uncertainty-based query method. BALD selects data points that maximize the difference between the model's prediction and the posterior probability, and identifies samples with the highest mutual information. This approach considers both the model's uncertainty and the posterior probability, enabling the selection of informative samples (Shen et al., 2017; Siddhant and Lipton, 2018; Shelmanov et al., 2021; Margatina et al., 2021a).

2.2 Diversity-based query methods

Representative sampling is a widely used method that selects unlabeled sample points based on their representativeness within the dataset. Diverse core-set is a technique that combines diversity and representativeness (Geifman and El-Yaniv, 2017; Ash et al., 2020). Representative sampling is a widely used method that selects unlabeled sample points based on their representation within the dataset, and diverse core-set is a technique that combines diversity and representativeness.

Maximizing margin distances (Tong and Koller, 2001) aims to improve the model's generalization ability by selecting samples farthest from the decision boundary between different categories. By choosing samples with maximum margin distances, the model can better distinguish between distinct categories and enhance its classification performance.

Maximizing class coverage focuses on selecting samples that cover different classes to ensure balanced learning across all categories (Settles et al., 2007; Huang et al., 2016). By including representative samples from each class, the model gains a comprehensive understanding of all categories during the learning process.

2.3 Hybrid query methods

Hybrid query methods combine uncertainty sampling and diversity sampling to overcome their limitations. These methods aim to strike a balance in the sampling strategy and adapt better to evolving data distributions (Liu M et al., 2018). For example, active learning with imitation learning mitigates the impact of changes in data distribution on heuristic-based active learning methods by incorporating imitation learning. This approach reduces reliance on heuristics and adapts better to evolving data distributions.

Batch active learning by diverse gradient embeddings (BADGE) (Ash et al., 2020) combines prediction uncertainty and sample diversity in the selection of each batch. It does not require manual tuning of hyperparameters, making it a robust and user-friendly approach.

Contrastive active learning (CAL) (Margatina et al., 2021b) focuses on selecting a set of contrastive examples. CAL identifies data points that are similar in the model feature space, yet yield maximally different predictive likelihoods. By leveraging contrastive examples, CAL aims to enhance the model's understanding of complex decision boundaries and improve its generalization capabilities.

Adaptive hybrid sampling for active learning (Wu et al., 2021) adjusts the weights of uncertainty sampling and diversity sampling to select the best sampling strategy. It dynamically adapts to the current model performance and the demand for labeled data, combining the strengths of both sampling strategies.

Adaptive hybrid active Learning with reinforcement learning (Konyushkova et al., 2017) models the active learning problem as a reinforcement learning problem. The model dynamically chooses the optimal sampling strategy based on the current state and environment, improving the sampling strategy through interaction and feedback.

In summary, uncertainty-based, diversity-based, and hybrid query methods are extensively used in active learning. These methods have demonstrated significant achievements in natural language processing tasks, reduced annotation costs, and enhanced model performance. Choosing the appropriate sampling method depends on task requirements and data characteristics, and facilitates an efficient active learning process.

2.4 Chinese affective structure analysis

The objective of affective structure analysis is to extract complete sentiment tuples from sentences. Barnes et al. (2021) introduced the concept of sentiment structure analysis, framing it as a dependency graph parsing task. They employed a "head/tail" transformation approach and applied first-order parsing methods. Building upon their modeling framework, Shi et al. (2022) proposed a

novel labeling strategy and used graph attention networks for aggregative decoding of span boundaries. Samuel et al. (2022) employed transformers to directly predict dependency graphs from text; Zhai et al. (2023) introduced new labels to simulate the boundaries of discontinuous spans and applied axial attention encoders along with table-filling schemes to decode relationships. Zhou et al. (2024) further examined the internal structure of spans, proposing a two-stage parsing method that leverages TreeCRFs and a novel internal constraint algorithm to explicitly model latent structures, while exploiting the advantages of joint scoring graph arcs and the head of spans for global optimization and inference. Our goal is to develop an efficient active learning method under low-resource constraints within the context of Chinese affective structure analysis.

3 Methodology

In this section, we provide a detailed description of our proposed active learning strategy, MAL. It comprises four key modules: feature extraction, filter, word-level divergence calculator, and update. The main task of the feature extraction module is to represent samples using a uniform encoding method (BERT), which serves as the foundational representation layer for the entire framework. The filter module primarily conducts the first selection of samples at the sentence level, aiming to increase the pool of samples to be labeled by including those with emotional structures similar to the already labeled samples. The word-level divergence calculator implements the second selection strategy, focusing on the diversity of Chinese expressions at the word level. The final update module functions as the executor in changing the sample pool in the active learning process.

The feature extraction module is responsible for extracting sentence-level feature representations from the input samples using BERT, a powerful pre-trained language model. This module leverages BERT's advanced feature extraction capabilities to capture rich semantic information from the text.

The filter module evaluates the distance between the feature representations of the selected sample and the remaining samples at the sentence level. This step aims to identify samples that are similar to the selected sample in terms of sentiment expression and structure. By considering the distance between samples, we can ensure that the selected samples are representative and diverse.

The word-level divergence calculator module calculates the divergence between the predicted label probability distributions of different samples, enabling a more detailed assessment of the variation in affective structure at the word-level.

The update module is the final step in the MAL strategy. It adds the selected samples to the training set and updates the model accordingly. This process continues iteratively until the pool of samples to be selected is empty or a predefined stopping criterion is met. By iteratively selecting informative samples, MAL effectively reduces the labeling effort while maximizing the learning performance of the model.

Overall, MAL combines sentence-level and word-level information to comprehensively understand the affective structure of Chinese text. It leverages the power of BERT for feature extraction, evaluates the similarity between samples at different levels, and selects samples that provide high labeling value. Through this iterative process, MAL achieves superior performance compared to baseline strategies while minimizing the annotation cost.

3.1 Feature extraction module

The feature extraction module plays a crucial role in the affective structure recognition pipeline as it extracts valuable and detailed sentence-level feature representations from the input samples. In our approach, we leverage BERT, a highly powerful pre-trained language model, to conduct advanced feature extraction. This module takes advantage of BERT's cutting-edge language understanding capabilities to capture comprehensive semantic information from the text. These extracted sentence-level features then serve as inputs for subsequent stages of the affective structure recognition pipeline. By incorporating BERT into the feature extraction process, the Feature Extraction module gains the advantage of capturing fine-grained semantic information and contextual dependencies. This capability greatly enhances the overall effectiveness of the affective structure recognition system.

$$\text{Feature}_{\text{sentence}} = \text{BERT}_{[\text{CLS}]}. \tag{1}$$

For the feature representation of all sample points, we use the [CLS] token embedding of BERT as the representation, denoted as Eq. (1). This approach allows for the comprehensive fusion of semantic information from individual words or characters in the text.

3.2 Filter module

The same emotion can be expressed in various ways in Chinese. For example, the sentences "忽然觉得很孤单。连个说话的人都没有。" (Suddenly I felt very lonely. Not even a person to talk to.) and "世界杯结束了，我感觉到空虚、寂寞、有点冷，我的生活失去了奔头！哎！" (The World Cup is over, I feel void, lonely, a little cold, my life lost the run! Oops!) both convey the feeling of loneliness. However, the expressions used are different. By selecting samples that are similar to the current training samples at the sentence level, our model can effectively learn the various expressions of the same emotion, thereby enhancing understanding and recognition capabilities.

We employ the semantic distance as the measure of similarity between sample points. Specifically, we consider a sample to be selected if its distance to a training sample is less than a small threshold value, denoted as ϵ. Euclidean distance was chosen as the semantic similarity measure in our model. Given two sentences s^1 and s^2, the distance is calculated as in Eq. (2).

$$\text{Distance}(s^1, s^2) = \sqrt{\Sigma(s_i^1 - s_i^2)^2}, \tag{2}$$

where $i = 1, 2, ..., n$, s_i^1 denotes the i-dimensional coordinates of the first sentence embedding, and s_i^2 denotes the i-dimensional coordinates of the second sentence embedding.

However, there are certain risks associated with this approach, such as the possibility of not finding sample points that satisfy the condition. To mitigate these risks, we select the h partners with the shortest distance from the queried sample. These partners are referred to as such because they often provide valuable assistance.

Considering the diverse expressions in the Chinese language, where different lexical choices can convey the same semantic meaning, relying solely on sentence-level semantics is not sufficient. It is important to also consider the word-level information.

3.3 Word-level divergence calculator

In a sequence labeling task, each word in a sentence is associated with a specific label, indicating its probability distribution among different categories. Taking the example from the previous subsection, words such as "孤单" (lonely) and "空虚" (void), "有点冷" (a little cold) can all express the emotion of loneliness in this context, and they may even be interchangeable in certain scenarios. Hence, it is important to capture sample points that exhibit similarity in both the feature space at the sentence level and the probability distribution at the word level.

$$\text{KL}(p||q) = \sum p\,(x) \log \frac{p(x)}{q(x)}. \tag{3}$$

We use the Kullback–Leibler (KL) divergence to measure the similarity of probability distributions between two sample points. The KL divergence is calculated using Eq. (3), where a smaller KL divergence indicates a higher similarity in probability distributions between the sample points.

To assess the annotation value of unannotated samples at both the sentence and word levels, we introduce the concept of Cross-KL divergence. Given two probability distributions $P = \{p_1, p_2, \ldots, p_m\}$, $Q = \{q_1, q_2, \ldots, q_m\}$ for two sentences, where m, n represent the length of the two sentences, respectively. For each word of each sentence, the model $\mathcal{M}$ outputs a probability distribution P and Q for all possible labels. The Cross-KL divergence between P and Q is defined in Eq. (4).

$$\text{KL}(P||Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} \text{KL}(p_i||q_j)}{m * n}. \tag{4}$$

3.4 Update module

The function of the update module is to redefine the samples in the data pool according to the query strategy to enhance their labeling value and improve model performance. First, our framework identifies representative samples at the sentence level that also exhibit diversity at the word level based on the evaluation results of the aforementioned three modules. These sample points are then sent to the labeling system for manual annotation. After obtaining the annotation results, the newly labeled samples are incorporated into the training set. Finally, by retraining the model, a new round of the sample selection

evaluation process is carried out until the termination condition is met.

3.5 Multilevel active learning

In this section, we provide a detailed description of the proposed multilevel active learning (MAL) approach, which is outlined in Algorithm 1.

Algorithm 1 Multilevel active learning

Require:
Labeled data set $\mathbb{U}_{\text{label}}$
Unlabeled data set $\mathbb{U}_{\text{pool}}$
Iteration size d
Number of partners h
Model $\mathcal{M}$

Ensure:
Queried data set $\mathbb{U}_d$

1: **for** x_p in $\mathbb{U}_{\text{pool}}$ **do**
2: $\text{BERT}(x_p, x_l) \rightarrow \text{CLS}(x_p, x_l)$, $x_l \in \mathbb{U}_{\text{label}}$
3: $Ed\left(x_p, x_l^k\right)_{\text{top}-h_{\min}}, k = 1, 2, ..., h$
4: $\mathcal{M}(x_p) \rightarrow \widetilde{P(x_p)}$
5: **for all** $x_l^k, k = 1, 2, ..., h$ **do**
6: $\mathcal{M}(x_l^k) \rightarrow \widetilde{P(x_l^k)}$
7: $\text{KL}(\widetilde{P\left(x_l^k\right)}||\widetilde{P\left(x_p\right)}) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\text{KL}(\widetilde{P(x_l^{k\,i})}||\widetilde{P(x_p{}^j)})}{m*n}$
8: **end for**
9: $s_{x_p} = \sum_{k=1}^{h}\text{KL}(\widetilde{P(x_l^k)}||\widetilde{P(x_p)})$
10: **end for**
11: $\mathbb{U}_d = \text{argmins}_{x_p}, |\mathbb{U}_d| = d$

Given the labeled dataset $\mathbb{U}_{\text{label}}$ for training the model, we want to iteratively draw d samples from the unlabeled data pool $\mathbb{U}_{\text{pool}}$ into $\mathbb{U}_{\text{label}}$ for training. Specifically, we first use the [CLS] token embedding of BERT as the sentence-level feature representation of x_p and x_l, and then compute the Euclidean distance between x_p and x_l based on the feature representation to find hx_l^k with the shortest distance from x_p, $k = 1, 2, ..., h$ (cf. line 3). In this way, we match h partners for each $x_p \in \mathbb{U}_{\text{pool}}$, and the next computations revolve around them. We use a classifier called $\mathcal{M}$ to calculate the probability distributions of x_p and its partner x_l^k. These probability distributions are denoted as $\widetilde{P(x_p)}$, $\widetilde{P(x_l^k)}$ respectively. We then compute the cross-KL divergence between them to measure the difference in their probability distributions. We repeat the computation operation for a total of h times, where h is a predetermined constant. Finally, we sum up the h cross-KL divergences to obtain the final score of x_p, denoted as s_{x_p}. Finally, we select d samples with the lowest scores among all x_p to add to $\mathbb{U}_{\text{label}}$ for iterative training, and remove these d samples from $\mathbb{U}_{\text{pool}}$. These x_p samples not only have sentence-level feature representations similar to x_l, but also have similar word-level probability distributions.

To summarize, the process of iteratively drawing samples from the unlabeled data pool $\mathbb{U}_{\text{pool}}$ into the labeled dataset $\mathbb{U}_{\text{label}}$ can be described as follows:

For each sample x_p in $\mathbb{U}_{\text{pool}}$:

1. Compute the sentence-level feature representation of x_p using the [CLS] token embedding of BERT.

2. Calculate the Euclidean distance between the feature representation of x_p and the feature representations of the samples x_l in $\mathbb{U}_{\text{label}}$.

3. Select the h samples x_l^k from $\mathbb{U}_{\text{label}}$ that have the shortest distances to x_p.

For each pair (x_p, x_l^k):

1. Use the classifier $\mathcal{M}$ to compute the probability distributions $\widetilde{P(x_p)}$ and $\widetilde{P(x_l^k)}$.

2. Calculate the Cross-KL divergence between the probability distributions of x_p and x_l^k to measure the difference in their probability distributions.

3. Sum the h Cross-KL divergences to obtain the final score s_{x_p} for x_p.

Select the d samples with the lowest scores s_{x_p} from all the samples x_p in $\mathbb{U}_{\text{pool}}$:

1. Add these d selected samples to $\mathbb{U}_{\text{label}}$ for training the model.

2. Remove the selected samples from $\mathbb{U}_{\text{pool}}$.

By considering both the sentence-level feature representations and the word-level probability distributions, the proposed approach aims to select samples from $\mathbb{U}_{\text{pool}}$ that exhibit both sentence-level features and similar probability distributions similar to those in $\mathbb{U}_{\text{label}}$. This selection process enables the capture of samples with diverse expressions, enhancing the model's ability to handle variations in language usage. Moreover, by leveraging unlabeled data in this manner, the approach maximizes the utilization of available resources for model training, potentially improving overall performance.

4 Experiments

4.1 Dataset

We evaluate our method on a Chinese dataset called CTAS (Xiong et al., 2023), which consists of 8 categories. The data is labeled in the BIO format. The training, development, and testing sets are approximately divided in a ratio of 8:1:1 for all datasets. Detailed statistics of the datasets are shown in Table 1.

4.2 Baselines

To validate the performance of our proposed MAL, we compare our approach to the following baselines.

1. Contrastive active learning (Margatina et al., 2021b). CAL aims to select a set of contrastive examples, that is, data points that are similar in the model feature space and yet the model outputs maximally different predictive likelihoods. By comparing the prediction results of similar samples, CAL aims to identify challenging samples and enable the model to better understand decision boundaries.

2. Lower token probability (Liu MY et al., 2022). LTP selects the tokens whose probability under the most likely tag sequence y is lowest. LTP aims to select samples for which the model is most uncertain or ambiguous about its predictions, as these samples may contain valuable information that requires further labeling to improve the model's understanding.

$$\phi_{\mathrm{LTP}}(x) = 1 - \min_{y_i \in y} P(y_i|x_i), \tag{5}$$

where y_i and x_i denote the i^{th} token in the text sequence and its corresponding label, respectively.

3. Least confidence (Culotta et al., 2005). LC sorts the examples in ascending order according to the probability that the model assigns to the most likely sequence of labels. LC selects the samples with the lowest confidence, indicating that the model is most uncertain about its predicted outcome. By querying samples for which the model is uncertain about the outcome, LC aims to obtain more information-rich data points and reduce classification errors.

$$\phi_{\mathrm{LC}}(x) = 1 - P(y|x). \tag{6}$$

4. Normalized least confidence. NLC is an extension of the LC strategy that takes into account the effect of sample length and normalizes the confidence score by dividing it by the length of the sequence. The normalized confidence helps prevent preference for longer sequences and ensures fair comparisons between samples of different lengths.

$$\phi_{\mathrm{NLC}}(x) = 1 - \frac{1}{n} P(y|x). \tag{7}$$

5. Maximum token entropy (Settles and Craven, 2008). MTE builds on the TE strategy by removing the restriction of querying only shorter sequences. It allows querying of longer sequences if they contain more information. MTE captures potentially information-rich patterns in longer sequences by considering the overall entropy, and encourages the active learning process to explore and query diverse and potentially complex samples.

$$\phi_{\mathrm{MTE}}(x) = -\sum_{n=1}^{N}\sum_{m=1}^{M} P(y_n = m)\log P(y_n = m), \tag{8}$$

where N is the length of x, m ranges over all possible token labels, and $P(y_n = m)$ is shorthand for the marginal probability that m is the label at position n in the sequence, according to the model.

6. Minimum token probability (Liu MY et al., 2022). MTP focuses on selecting samples with the highest information content, regardless of the effect of conditional random field (CRF) decoding. It selects samples based on the probability of individual tokens and aims to query samples with low probability of information-rich tokens, which may challenge the current understanding of the model and provide valuable insights.

$$\phi_{\mathrm{MTP}}(x) = 1 - \min_i \max_j P(y_i = j|x_i), \tag{9}$$

where $\max_j P(y_i = j|x_i)$ is the highest probability assigned to any label j for the i^{th} token x_i, and $\min_i \max_j P(y_i = j \mid x_i)$ is the minimum of these maximum probabilities over all tokens i in the input x.

In addition to the baselines used above, we also chose Random selection as a baseline method.

4.3 Experimental setup

We used the Early-Stopping technique to determine the optimal model from the development set, with a maximum limit of 100 epochs. If there was

Table 1 Data split statistics in CTAS and description of labels

	Train	Dev	Test	Paraphrase
Number of sentences	5440	683	684	–
# Cause	910	95	118	The thing that makes emotions happen
# Degree	4192	538	537	The level or amount of emotions
# Holder	1813	214	220	The person or people who hold the emotions
# Negation	325	34	48	To express the emotions that do not exist
# Property	192	31	32	A characteristic of an entity
# Trigger	7500	971	953	The expression of emotions
# Compared_entity	64	4	5	The entity being compared or related to emotions
# Sent_entity	2803	373	358	The entity being sent emotions or receiving emotions

no improvement in the score within 60 epochs, we terminated the training process. Subsequently, we saved the optimal model and evaluated its performance on the testing set. Consistently, we employed Bert-CRF as the base model and used the bert-base-Chinese pre-training model. Our learning rate was set at $5e^5$, with a batch size of 64. To facilitate efficient batch training, we imposed a sample length limit of 256. For initial training, we selected a mere 0.05% of the available data, while the iteration size d remained fixed at 0.05%. All experiments were conducted on an Nvidia GeForce RTX3090 GPU. Our code is available at https://github.com/henau-nlp/MAL.

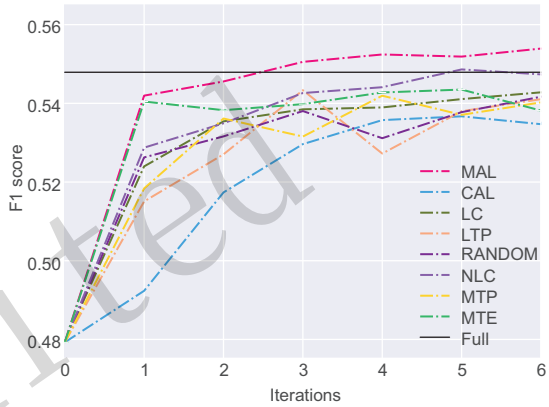


Fig. 2 Performance of MAL and baseline strategies on CTAS

4.4 Experimental results

We assessed the entity-level F1-scores of both the baseline strategies and MAL approach using the CTAS dataset. The baseline score, as established by Xiong et al. (2023), was 0.5479. To ensure consistency and accuracy, we used fixed random seeds for initialization, effectively reducing any potential random effects.

We present the performance of all strategies on the CTAS dataset in Fig. 2. It is evident that the MAL strategy was the most competitively and outperforms all baseline strategies. The confidence-based NLC strategy avoided querying longer sentences, as opposed to LC, and was the top performer among the baseline strategies. Although the LC and NLC strategies are straightforward and intuitive, the confidence-based query strategy remains highly competitive.

The MTE strategy overcame the limitation of the TE strategy by allowing the querying of longer sentences. Nevertheless, sentence length should not be a major concern for the sentiment structure recognition task. Hence, the performance of these two strategies in this task was more general. Similarly, the MTP strategy did not yield better results in selecting the most informative sentences. The LTP strategy, which considers both global and local information simultaneously, did not achieve competitive performance.

CAL, which was the worst-performing query strategy, also took into account samples that were similar in feature space. However, the maximum prediction gap among similar samples interfered with model learning.

In conclusion, for our task, MAL stands out as the most effective choice, while confidence-based LC and NLC continue to exhibit strong performance. When comparing the MAL and CAL strategies, it becomes clear that selecting samples that are closely aligned in terms of features and have smaller divergence enhances model learning. These results serve as valuable guidance for selecting an active learning

strategy tailored to a specific task, ultimately aiding in the optimization of labeled data utilization and the improvement of model performance.

5 Ablation study

We conducted an ablation study to evaluate the contribution of the filter fraction and the impact of the number of partners on performance.

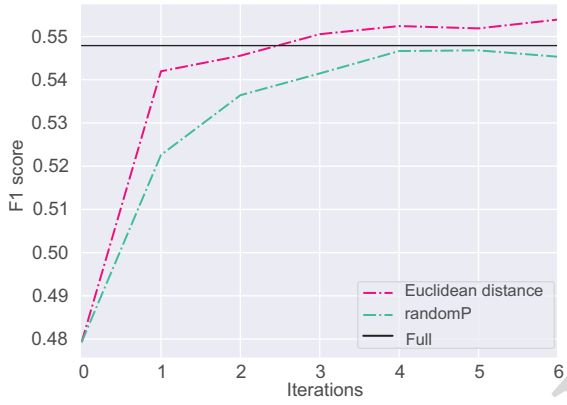


Fig. 3 Impact of partner selection on performance, where randomP refers to the selection of a partner in a random manner

Filter: we determine partners for x_p by calculating the Euclidean distance between their feature representations. To assess the effectiveness of the Filter module, we randomly select h partners for x_p. The results of this random selection are depicted in Fig. 3. It is evident that the Random approach, which randomly selects partners, exhibits poorer performance during the initial iterations and displays significant fluctuations. This observation suggests that considering the distance between the sentence-level feature representations of labeled samples and the samples to be selected is a meaningful approach for sample selection. This consideration can enhance the model's learning ability at the sentence level.

Partners: we also examined the influence of the number of partners on MAL performance, as depicted in Fig. 4. Interestingly, an increase in the number of partners does not lead to improved performance. Instead, the model using more partners shows greater fluctuations. This phenomenon can be attributed to the inherent complexity of the sentiment structure recognition task in Chinese text, as explored in this paper. When dealing with a larger number of partners, increased uncertainty is introduced, and the selection of more partners for x_p can potentially result in the inclusion of data points that are situated farther from other unselected data points, despite their relative proximity.

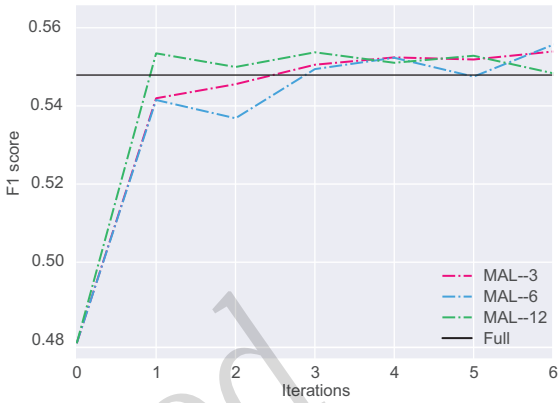


Fig. 4 Impact of the number of partners on performance

6 Analysis

In the preceding section, we established the efficacy of the filter. To further explore the efficiency of MAL, Fig. 5 visually depicts the evolution of token selection by MAL-3, MAL-12, and randomP sampling after the initial two iterations. Due to the heterogeneous distribution of sample labels, we have divided this into two plots for more intuitive visualization.

Evidently, regardless of whether we select 3 or 12 partners, MAL consistently selects more rare labels than randomP. As the number of iterations increases, MAL demonstrates the capability to select additional rare samples for the model, thereby enhancing the model's learning capacity for these infrequent samples.

Finally, we further investigated MAL and all the considered acquisition functions (baselines) to gain insights into the types of data each method tends to select. Specifically, we evaluated the number of labeled samples selected by each strategy. Table 2 presents the initial number of labeled samples in the training set and the number of labeled samples after obtaining 10% of the dataset using each method. This analysis provides valuable information on the

label distribution and the effectiveness of the active learning strategies in selecting informative samples.

The distribution of labels in the CTAS dataset is unbalanced, with degree, holder, trigger, and sent_entity accounting for more than 90% of the labels. In general, cause labels are different from other labels; they are longer because they tend to be a paragraph rather than a short one-word sentence, and thus the number of cause tags varies greatly from data point to data point.

In our investigation, we leveraged the power of BERT, a highly effective pre-trained language model, within the MAL framework. By incorporating BERT into the feature extraction process, we were able to harness its advanced language understanding capabilities to capture comprehensive semantic information from the text. This integration of BERT with MAL offers several advantages. First, BERT demonstrates its effectiveness in capturing feature and contextual information for categories with a limited number of instances, such as the compared_entity, even in datasets with imbalanced distributions. Second, BERT enhances the Filter module by offering a more comprehensive sentence-level feature representation.

Among the samples selected by all the strategies, the number of labels for the samples selected using MAL is lowest except for the compared_entity. The number of labels did not double when the size of the training set was doubled. This indicates that in the Chinese textual affective structure recognition task, increasing the amount of information in the training data does not improve the training effect. On the contrary, too much redundant information may interfere with the model, whereas our proposed method can find the truly valuable samples for the model.

In summary, our experimental results show that the MAL strategy performs well in picking the number of sample labels, and the number the selected sample labels is less compared with other strategies. This further demonstrates the ability of our approach to discover samples that are of real value for model training. In the affective structure recognition task, too much redundant information does not necessarily help model training, whereas our method can provide more targeted and efficient sample selection, thus improving the model performance.

7 Discussion

Our proposed MAL strategy combines sentence-level and word-level information. First, we use BERT to extract sentence-level feature representations of the samples. Then, we selected h partners for the unlabeled sample x_p by measuring the Euclidean distance between samples, such that x_p is similar to these partners in terms of sentence-level features. Next, we obtained the word-level probability distributions of x_p and its partners using the CRF model. To measure the differences in word-level probability distributions between x_p and its partners, we designed the cross-KL method. By minimizing the differences, we were able to better capture the information differences between samples and thus select samples with smaller differences for annotation.

For the affective structure identification task of Chinese social media corpus, the task is rich in diverse and non-standardized expressions. Therefore, we considered the information of samples from both sentence-level and word-level perspectives. The sentence-level feature representation can capture the overall semantic information, while the word-level probability distribution can focus on the fine-grained annotation information of the samples. This approach of fusing sentence-level and word-level information can capture the features of the samples more comprehensively and improve the performance of the model on Chinese social media corpus.

8 Conclusions

This study introduces the multilevel active learning (MAL) strategy, which significantly enhances textual affective structure identification (TASI) for Chinese social media texts. By integrating BERT for sentence-level feature extraction with CRF for word-level probability distributions, MAL effectively combines these two layers of information to improve sample selection for annotation. Our approach addresses the limitations of traditional query strategies, which often overlook critical yet challenging samples or select less relevant ones, by providing a more comprehensive framework that balances uncertainty, diversity, and information richness.

The experimental results demonstrate that MAL not only outperforms existing baseline methods but also achieves superior performance while re-

Table 2 Number of labels in the training sets

	Initial	LC	NLC	MTP	MTE	LTP	Random	CAL	MAL
# Cause	46	73	203	104	73	155	93	199	70
# Degree	211	397	431	426	383	476	413	490	327
# Holder	86	175	204	194	159	204	187	217	129
# Negation	19	39	38	47	37	39	33	45	27
# Property	10	22	18	27	21	26	20	30	13
# Trigger	373	744	707	809	729	752	750	826	626
# Compared_entity	3	4	9	6	3	6	6	12	6
# Sent_entity	134	255	257	288	238	326	265	405	204

ducing labeling costs. This study represents pioneering research in applying active learning specifically to Chinese affective structure identification, marking a significant advancement in the field. Moving forward, our focus will be on refining the MAL strategy and exploring its application to other sequence labeling tasks. We also aim to integrate MAL with advanced models to further enhance its effectiveness, thereby contributing valuable solutions to the broader NLP landscape.

Contributors

Shufeng XIONG, Guipei ZHANG, and Haiping SI designed the research. Guipei ZHANG, Xiaobo FAN, and Wenjie TIAN processed the data. Guipei ZHANG and Xiaobo FAN drafted the paper. Lei XI and Hebing LIU helped organize the paper. Shufeng XIONG and Haiping SI revised and finalized the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are openly available in github at https://github.com/henau-nlp/MAL.

References

Alamoodi AH, Zaidan BB, Zaidan AA, et al., 2021. Sentiment analysis and its applications in fighting covid-19 and infectious diseases: a systematic review. *Expert Syst Appl*, 167:114155. https://doi.org/10.1016/j.eswa.2020.114155

Angluin D, 1988. Queries and concept learning. *Mach Learn*, 2(4):319-342. https://doi.org/10.1023/A:1022821128753

Ash JT, Zhang CC, Krishnamurthy A, et al., 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. Proc 8th Int Conf on Learning Representations.

Barnes J, Kurtz R, Oepen S, et al., 2021. Structured sentiment analysis as dependency graph parsing. Proc 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int Joint Conf on Natural Language Processing, p.3387-3402. https://doi.org/10.18653/v1/2021.acl-long.263

Basiri ME, Nemati S, Abdar M, et al., 2021. ABCDM: an attention-based bidirectional CNN-RNN deep model for sentiment analysis. *Fut Gener Comput Syst*, 115:279-294. https://doi.org/10.1016/j.future.2020.08.005

Bishop CM, 2006. Pattern Recognition and Machine Learning. Springer, New York, USA.

Bodó Z, Minier Z, Csató L, 2011. Active learning with clustering. Active Learning and Experimental Design Workshop, in Conjunction with AISTATS 2010, p.127-139.

Brinker K, 2003. Incorporating diversity in active learning with support vector machines. Proc 20th Int Conf on Machine Learning, p.59-66.

Chen YK, Lasko TA, Mei QZ, et al., 2015. A study of active learning methods for named entity recognition in clinical text. *J Biomed Inf*, 58:11-18. https://doi.org/10.1016/j.jbi.2015.09.010

Cohn DA, Ghahramani Z, Jordan MI, 1996. Active learning with statistical models. *J Artif Intell Res*, 4:129-145. https://doi.org/10.1613/jair.295

Culotta A, McCallum A, 2005. Reducing labeling effort for structured prediction tasks. Proc 20th National Conf on Artificial Intelligence and the 17th Innovative Applications of Artificial Intelligence Conf, p.746-751.

Dagan I, Engelson SP, 1995. Committee-based sampling for training probabilistic classifiers. Proc 12th Int Conf on Machine Learning, p.150-157. https://doi.org/10.1016/B978-1-55860-377-6.50027-X

Dasgupta S, 2011. Two faces of active learning. *Theor Comput Sci*, 412(19):1767-1781. https://doi.org/10.1016/j.tcs.2010.12.054

Devlin J, Chang MW, Lee K, et al., 2019. BERT: pre-training of deep bidirectional transformers for language understanding. Proc Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, p.4171-4186. https://doi.org/10.18653/v1/N19-1423

Dor LE, Halfon A, Gera A, et al., 2020. Active learning for BERT: an empirical study. Proc Conf on Empirical Methods in Natural Language Processing, p.7949-7962. https://doi.org/10.18653/v1/2020.emnlp-main.638

Ducoffe M, Precioso F, 2018. Adversarial active learning for deep networks: a margin based approach. https://arxiv.org/abs/1802.09841

Gal Y, Islam R, Ghahramani Z, 2017. Deep Bayesian active learning with image data. Proc 34th Int Conf on Machine Learning, p.1183-1192.

Geifman Y, El-Yaniv R, 2017. Deep active learning over the long tail. https://arxiv.org/abs/1711.00941

Hanneke S, 2014. Theory of disagreement-based active learning. *Found Trends Mach Learn*, 7(2-3):131-309. https://doi.org/10.1561/2200000037

Houlsby N, Huszár F, Ghahramani Z, et al., 2011. Bayesian active learning for classification and preference learning. https://arxiv.org/abs/1112.5745

Hu R, Mac Namee B, Delany SJ, 2016. Active learning for text classification with reusability. *Expert Syst Appl*, 45:438-449. https://doi.org/10.1016/j.eswa.2015.10.003

Huang TK, Li LH, Vartanian A, et al., 2016. Active learning with oracle epiphany. Proc 30th Int Conf on Neur Information Processing Systems, p.2828-2836

Kirsch A, Van Amersfoort J, Gal Y, 2019. BatchBALD: efficient and diverse batch acquisition for deep Bayesian active learning. Proc 33rd Int Conf on Neural Information Processing Systems, article 631.

Konyushkova K, Sznitman R, Fua P, 2017. Learning active learning from data. Proc 31st Int Conf on Neural Information Processing Systems, p.4228-4238.

Lewis DD, 1995. A sequential algorithm for training text classifiers: corrigendum and additional data. *ACM SIGIR Forum*, 29(2):13-19. https://doi.org/10.1145/219587.219592

Liu M, Buntine W, Haffari G, 2018. Learning how to actively learn: a deep imitation learning approach. Proc 56th Annual Meeting of the Association for Computational Linguistics, p.1874-1883. https://doi.org/10.18653/v1/P18-1174

Liu MY, Tu ZY, Zhang T, et al., 2022. LTP: a new active learning strategy for CRF-based named entity recognition. *Neural Process Lett*, 54(3):2433-2454. https://doi.org/10.1007/s11063-021-10737-x

Margatina K, Barrault L, Aletras N, 2021a. On the importance of effectively adapting pretrained language models for active learning. Proc 60th Annual Meeting of the Association for Computational Linguistics, p.825-836. https://doi.org/10.18653/v1/2022.acl-short.93

Margatina K, Vernikos G, Barrault L, et al., 2021b. Active learning by acquiring contrastive examples. Proc Conf on Empirical Methods in Natural Language Processing, p.650-663. https://doi.org/10.18653/v1/2021.emnlp-main.51

McCallum A, Nigam K, 1998. Employing EM and pool-based active learning for text classification. Proc 15th Int Conf on Machine Learning, p.350-358.

Medhat W, Hassan A, Korashy H, 2014. Sentiment analysis algorithms and applications: a survey. *Ain Shams Eng J*, 5(4):1093-1113. https://doi.org/10.1016/j.asej.2014.04.011

Qiu XP, Sun TX, Xu YG, et al., 2020. Pre-trained models for natural language processing: a survey. *Sci China Technol Sci*, 63(10):1872-1897. https://doi.org/10.1007/s11431-020-1647-3

Samuel D, Barnes J, Kurtz R, et al., 2022. Direct parsing to sentiment graphs. Proc 60th Annu Meeting of the Association for Computational Linguistics, p.470-478. https://doi.org/10.18653/v1/2022.acl-short.51

Sener O, Savarese S, 2018. Active learning for convolutional neural networks: a core-set approach. Proc 6th Int Conf on Learning Representations.

Settles B, 2010. Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison, Madison, WI.

Settles B, Craven M, 2008. An analysis of active learning strategies for sequence labeling tasks. Proc Conf on Empirical Methods in Natural Language Processing, p.1070-1079.

Settles B, Craven M, Ray S, 2007. Multiple-instance active learning. Proc 20th Int Conf on Neural Information Processing Systems, p.1289-1296.

Shelmanov A, Puzyrev D, Kupriyanova L, et al., 2021. Active learning for sequence tagging with deep pre-trained models and Bayesian uncertainty estimates. Proc 16th Conf of the European Chapter of the Association for Computational Linguistics: Main Volume, p.1698-1712. https://doi.org/10.18653/v1/2021.eacl-main.145

Shen YY, Yun H, Lipton Z, et al., 2017. Deep active learning for named entity recognition. Proc 2nd Workshop on Representation Learning for NLP, p.252-256. https://doi.org/10.18653/v1/W17-2630

Shi WX, Li F, Li JY, et al., 2022. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. Proc 60th Annual Meeting of the Association for Computational Linguistics, p.4232-4241. https://doi.org/10.18653/v1/2022.acl-long.291

Siddhant A, Lipton ZC, 2018. Deep Bayesian active learning for natural language processing: results of a large-scale empirical study. Proc Conf on Empirical Methods in Natural Language Processing, p.2904-2909. https://doi.org/10.18653/v1/D18-1318

Smailović J, Grčar M, Lavrač N, et al., 2014. Stream-based active learning for sentiment analysis in the financial domain. *Inf Sci*, 285:181-203. https://doi.org/10.1016/j.ins.2014.04.034

Tong SM, Koller D, 2001. Support vector machine active learning with applications to text classification. *J Mach Learn Res*, 2:45-66. https://doi.org/10.1162/153244302760185243

Venugopalan M, Gupta D, 2015. Exploring sentiment analysis on twitter data. Proc 8th Int Conf on Contemporary Computing, p.241-247. https://doi.org/10.1109/IC3.2015.7346686

Wu X, Chen C, Zhong MY, et al., 2021. HAL: Hybrid active learning for efficient labeling in medical domain. *Neurocomputing*, 456:563-572. https://doi.org/10.1016/j.neucom.2020.10.115

Xiong SF, Fan XB, Batra V, et al., 2023. An entropy-based method with a new benchmark dataset for Chinese textual affective structure analysis. *Entropy*, 25(5):794. https://doi.org/10.3390/e25050794

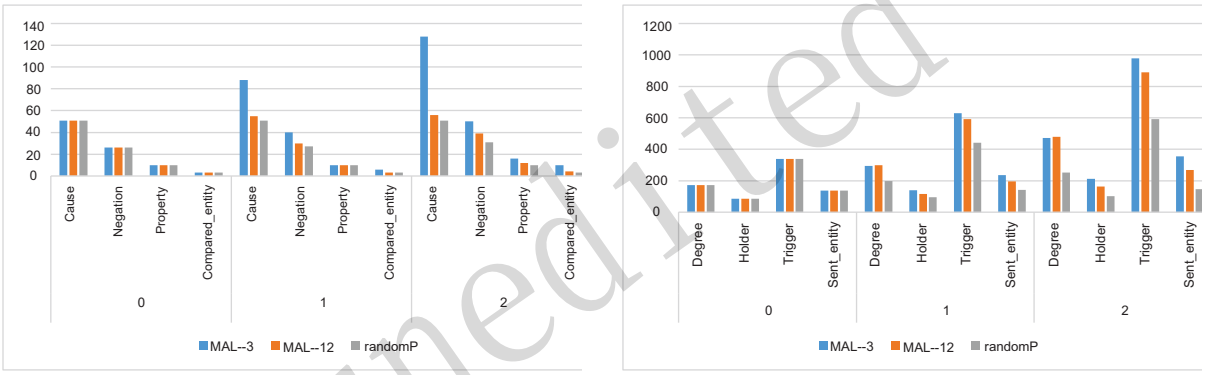
Yuan M, Lin HT, Boyd-Graber J, 2020. Cold-start active learning through self-supervised language modeling. Proc Conf on Empirical Methods in Natural Language Processing, p.7935-7948. https://doi.org/10.18653/v1/2020.emnlp-main.637

Zhai ZP, Chen H, Li RF, et al., 2023. USSA: a unified table filling scheme for structured sentiment analysis. Proc 61st Annual Meeting of the Association for Computational Linguistics, p.14340-14353. https://doi.org/10.18653/v1/2023.acl-long.802

Zhang HT, Huang ML, Zhu XY, 2012. A unified active learning framework for biomedical relation extraction. *J Comput Sci Technol*, 27(6):1302-1313. https://doi.org/10.1007/s11390-012-1306-0

Zhang MK, Plank B, 2021. Cartography active learning. Proc Findings of the Association for Computational Linguistics, p.395-406. https://doi.org/10.18653/v1/2021.findings-emnlp.36

Zhou CJ, Li BB, Fei H, et al., 2024. Revisiting structured sentiment analysis as latent dependency graph parsing. Proc 62nd Annual Meeting of the Association for Computational Linguistics, p.10178-10191. https://doi.org/10.18653/v1/2024.acl-long.548



(a) (b)

Fig. 5 Trend in the labels within the training set, randomP refers to the selection of a partner in a random manner