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 textual affective structure analysis*

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

College of Information and Management Science, 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 May 21, 2025

Abstract: Chinese textual affective structure analysis (CTASA) is a sequence labeling task that often relies on supervised deep learning methods. However, acquiring a large annotated dataset for training can be costly and time-consuming. Active learning offers a solution by selecting the most valuable samples to reduce labeling costs. Previous approaches focused on uncertainty or diversity but faced challenges such as biased models or selecting insignificant samples. To address these issues, multilevel active learning (MAL) is introduced, which leverages 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 bidirectional encoder representations from Transformers (BERT) embeddings and the word-level probability distributions obtained through a conditional random field (CRF) model, MAL comprehensively captures the Chinese textual affective structure (CTAS). Experimental results demonstrate that MAL significantly reduces annotation costs by approximately 70% and achieves more consistent performance compared to baseline methods.

Key words: Sentiment analysis; Sequence labeling; Active learning (AL); Bidirectional encoder representations from Transformers (BERT)

https://doi.org/10.1631/FITEE.2400242 **CLC number:** TP311

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 before conducting sentiment analysis involves identifying and labeling the sentiment-related terms within the target text, known as textual affective structure analysis (TASA), which essentially falls under the purview of sequence labeling tasks. TASA aims to extract complete sentiment tuples from sentences. It is an eight-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 Section 4.1. Our task setting focuses 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

‡ Corresponding author

* Project supported by the Ministry of Education (MOE) Project of Humanities and Social Science, China (No. 19YJCZH198) and the Key Research and Development Project of Henan Province (No. 231111211300)

ORCID: Shufeng XIONG, https://orcid.org/0000-0001-5727-1766; Haiping SI, https://orcid.org/0000-0001-8430-149X

is crucial to filter out samples that hold higher value for labeling purposes.

Active learning (AL) has demonstrated its effectiveness in reducing labeling costs by selecting the most valuable samples from a pool of unlabeled data. It has been widely employed in various natural language processing (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 core idea of AL lies in designing appropriate query functions. Currently, the prevailing AL query strategies include uncertainty-based, diversity-based, and hybrid query strategies. However, uncertainty-based query strategies often overlook the challenging yet crucial samples for the model, leading to data bias. Conversely, diversity-based query strategies may select meaningless samples, resulting in wastage of resources. To overcome these limitations, the hybrid query strategies combine the strengths of uncertainty-based and diversity-based strategies while incorporating information richness. By integrating uncertainty, diversity, and information richness, the hybrid query strategies provide a more comprehensive framework for guiding the model's learning process. Our method falls into this category.

In Chinese texts, particularly in social texts where there are no unified or 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 his 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 the 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 high annotation value and the diversity of vocabulary at the word level in AL methods. Existing query strategies typically consider only a single type of information from either the sentence level or the word level when selecting annotated samples, neglecting the other types of information. Experimental results in Section 4 indicate that the performances of different baseline methods using a single query strategy are unsatisfactory.

Moreover, it is crucial to note that finding a classifier coupled with an AL 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-trained models has significantly improved the performance of numerous NLP tasks as highlighted by Qiu et al. (2020), with bidirectional encoder representations from Transformers (BERT) (Devlin et al., 2019) receiving significant attention. This can be attributed to BERT's exceptional feature extraction capability and its ability to effectively capture sentence-level information. Therefore, our primary objective is to design an effective AL strategy that synergizes with BERT for the recognition of affective structures in Chinese text.

Taking inspiration from query contrastive samples (Margatina et al., 2021a), 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 method enables a more precise selection of representative and crucial samples. We introduce a novel method called multilevel AL (MAL). The overall architecture of MAL is illustrated in Fig. 1. In essence, MAL aims to select the samples that are similar to the annotated ones in terms of sentiment expression and structure, at both the sentence and word levels. In the following sections, we provide a detailed description of the proposed method. Experimental results demonstrate that MAL achieves superior performance while requiring less labeling costs compared to the baseline methods. Our contributions are as follows:

1. Our proposed AL strategy is used for the Chinese TASA (CTASA) task and 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, as it

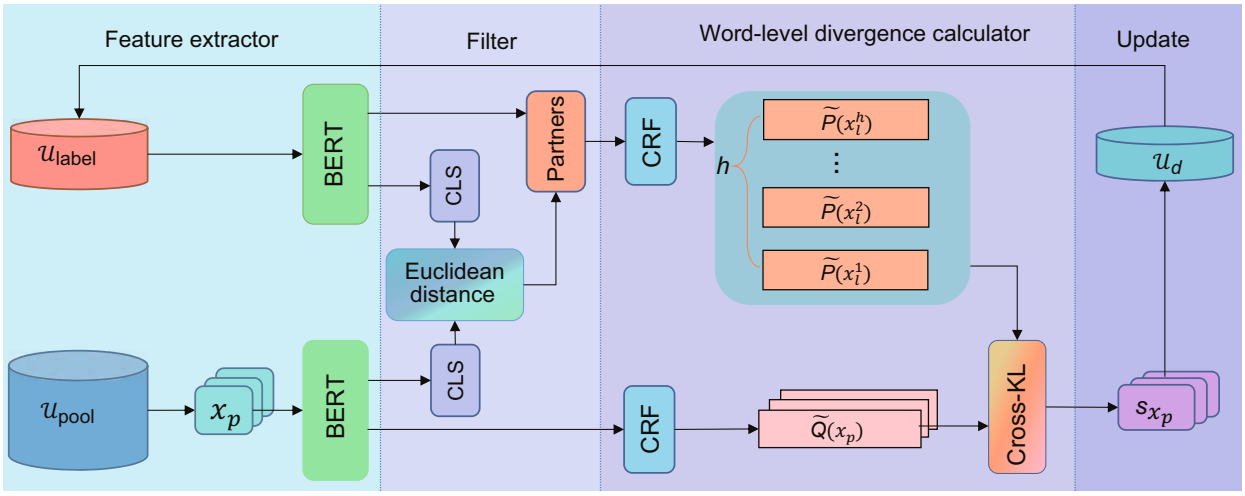

Fig. 1 Structure of multilevel active learning (MAL). BERT: bidirectional encoder representations from Transformers; CLS: classification; CRF: conditional random field; KL: Kullback–Leibler

consistently outperforms the baseline methods or achieves comparable performance.

2 Related works

The effectiveness of AL has been demonstrated in numerous studies (Settles, 2010; Dasgupta, 2011; Hanneke, 2014). Existing AL 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 iteratively trains them 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 strategies (Lewis, 1995; Cohn et al., 1996; Gal et al., 2017; Kirsch et al., 2019; Zhang MK and Plank, 2021), diversity-based query strategies (Brinker, 2003; Bodó et al., 2011; Sener and Savarese, 2018), and hybrid query strategies (Ducoffe and Precioso, 2018; Ash et al., 2020; Yuan et al., 2020; Margatina et al., 2021a) in AL research. These methods aim to strike a balance between uncertainty and diversity while selecting informative samples for annotation.

2.1 Uncertainty-based query strategies

Uncertainty-based query strategies 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 samples 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 field (CRF). It selects samples with the lowest labeling probability, capturing the model's comprehensive understanding of input sequences and choosing samples with rich information content. In contrast to LTP, our method initially focuses on sample augmentation at the sentence level while accounting for the diversity of expression at the word level. Specifically, whereas LTP emphasizes solely sample selection methods at the word level, our work represents the first AL 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 that used in 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 AL by disagreement (BALD) (Houlsby et al., 2011) is a notable uncertainty-based query method. BALD selects samples that maximize the difference between the model's prediction and the posterior probability, effectively identifying samples with the richest mutual information. This method considers both the model's uncertainty and the posterior probability, enabling the selection of informative samples (Shen et al., 2017; Siddhant and Lipton, 2018; Margatina et al., 2021b; Shelmanov et al., 2021).

2.2 Diversity-based query strategies

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

Maximizing margin distance (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 the maximum margin distance, 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 strategies

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

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

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

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

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

2.4 Chinese textual affective structure analysis (CTASA)

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 method and applied the 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 AL method under low-resource constraints within the context of CTASA.

3 Methodology

In this section, we provide a detailed description of our proposed AL strategy, MAL. It comprises four key modules: feature extraction, filter, word-level divergence calculator, and update.

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 capability to capture rich semantic information from the text.

The filter module primarily conducts the first selection of samples at the sentence level. This module evaluates the similarity between the feature representations of the selected samples and the rest samples at the sentence level. This step aims to increase the pool of samples to be labeled by identifying samples that are similar to the selected sample in terms of sentiment expression and structure. By considering the similarity between samples, we can ensure that the selected samples are representative and diverse.

The word-level divergence calculator module implements the second selection strategy. This module calculates the divergence among 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 of the MAL. It adds the newly selected samples to the training set and updates the model accordingly. This process iterates 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 both 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 among samples at different levels, and selects samples that provide high labeling value. Through this iterative process, MAL achieves superior performance compared to baselines while minimizing the annotation costs.

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 method, we leverage BERT 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 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.

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

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

3.2 Filter module

The same emotion can be expressed in various ways in Chinese. For example, the sentences "忽然我觉得很孤单。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 sets at the sentence level, our model can effectively learn the various expressions of the same emotion, thereby enhancing the understanding and recognition capabilities.

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

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

where s_i^1 and s_i^2 denote the i^{th} dimensional coordinates of the first and second sentence embedding, respectively.

However, there are certain risks associated with this method, such as the possibility of not finding samples that satisfy the condition. To mitigate these risks, we select h partners with the shortest distance from the queried sample. These partners are selected 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 also important to consider the word-level information.

3.3 Word-level divergence calculator module

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), “空虚” (void), and “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 samples that exhibit similarity in both the feature space at the sentence level and the probability distribution at the word level.

We use the Kullback–Leibler (KL) divergence to measure the similarity of probability distributions between two samples. The KL divergence is calculated in Eq. (3):

$$\text{KL}(P||Q) = \sum P(x)\log_2\frac{P(x)}{Q(x)}. \tag{3}$$

Given P as the true probability distribution and Q as the approximating distribution, the KL divergence sums over all possible samples x in the shared sample space. For each sample x, where the true distribution $Q(x) > 0$ weights the log-ratio of $P(x)$ to the approximating distribution $Q(x)$ by the probability $P(x)$ itself. A smaller KL divergence indicates a higher similarity in probability distributions between the samples.

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 $P = \{p_1, p_2, \ldots, p_m\}$ and $Q = \{q_1, q_2, \ldots, q_n\}$ for two sentences, where m and n are the lengths of two sentences, separately. For each word in each sentence, the classifier $\mathcal{M}$ outputs P and Q for all possible labels. The cross-KL divergence between P and Q is defined in Eq. (4):

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

3.4 Update module

The update module redefines 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 samples 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 (MAL)

In this subsection, we provide a detailed description of the proposed MAL, which is outlined in Algorithm 1.

Given the labeled dataset $\mathcal{U}_{\text{label}}$, we aim to iteratively draw d samples from the unlabeled data pool

Algorithm 1 Multilevel active learning (MAL)

Input: labeled dataset $\mathcal{U}_{\text{label}}$, unlabeled dataset $\mathcal{U}_{\text{pool}}$, number of samples selected per iteration d, number of partners h, and classifier $\mathcal{M}$

Output: queried dataset $\mathcal{U}_d$

for $x_p \in \mathcal{U}_{\text{pool}}$ **do**
  $\text{BERT}(x_p, x_l) \to \text{CLS}(x_p, x_l)$, $x_l \in \mathcal{U}_{\text{label}}$
  $\text{argmin Ed}(x_p, x_l^k)_h, k = 1, 2, ..., h$
  $\mathcal{M}(x_p) \to \widetilde{Q}(x_p)$
  for all $x_l^k, k = 1, 2, ..., h$ **do**
    $\mathcal{M}(x_l^k) \to \widetilde{P}(x_l^k)$
    $\text{cross} - \text{KL}(\widetilde{P}\left(x_l^k\right) || \widetilde{Q}\left(x_p\right)) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\text{KL}(\widetilde{P}\left(x_l^{k\,i}\right)||\widetilde{Q}\left(x_p{}^j\right))}{mn}$
  end for
  $s_{x_p} = \sum_{k=1}^{h} \text{cross} - \text{KL}(\widetilde{P}(x_l^k)||\widetilde{Q}(x_p))$
end for
$\mathcal{U}_d = \text{argmin } s_{x_p}, |\mathcal{U}_d| = d$

$\mathcal{U}_{\text{pool}}$ into $\mathcal{U}_{\text{label}}$ for training. Specifically, we first use the [CLS] token embedding of BERT as the sentence-level feature representation of unlabeled sample x_p and labeled sample x_l, and then compute the Euclidean distance between x_p and each $x_l \in \mathcal{U}_{\text{label}}$ based on the feature representation to find h samples x_l^k with the shortest distance from x_p, $k = 1, 2, ..., h$. In this way, we match h partners for each $x_p \in \mathcal{U}_{\text{pool}}$, and the next computations revolve around them. We use the classifier $\mathcal{M}$ to calculate the probability distributions of x_p and its partner x_l^k. These probability distributions are denoted as $\widetilde{P}(x_p)$ and $\widetilde{P}(x_l^k)$. We then compute the cross-KL divergence between them to measure the difference in their probability distributions. We repeat the computation operation for h times, where h is predetermined. 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, add them to $\mathcal{U}_{\text{label}}$ for iterative training, and remove them from $\mathcal{U}_{\text{pool}}$. These x_p samples not only have sentence-level feature representations similar to those of x_l, but also have similar word-level probability distributions.

To summarize, the iterative process of selecting samples from the unlabeled data pool $\mathcal{U}_{\text{pool}}$ into the labeled dataset $\mathcal{U}_{\text{label}}$ can be described in three steps.

1. For each sample $x_p \in \mathcal{U}_{\text{pool}}$:

(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 every sample $x_l \in \mathcal{U}_{\text{label}}$.

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

2. 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 up h cross-KL divergences to obtain the final score s_{x_p} for x_p.

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

(1) Add these d samples to $\mathcal{U}_{\text{label}}$ for training the model.

(2) Remove these selected samples from $\mathcal{U}_{\text{pool}}$.

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

4 Experiments

4.1 Dataset

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

4.2 Baselines

To validate the performance of our proposed MAL, we compare the method to the following baselines. In the representation formulas of all baselines,

Table 1 Data split statistics in CTAS and description of labels

Label of CTAS	Number of sentences			Description of label
	Training set	Development set	Test set	
Cause	910	95	118	Things that make emotions happen
Degree	4192	538	537	Levels of emotions
Holder	1813	214	220	Person or people who hold the emotions
Negation	325	34	48	Expression of emotions that do not exist
Property	192	31	32	Characteristic of entity
Trigger	7500	971	953	Expression of emotions
Compared_entity	64	4	5	Entities that are compared or related to emotions
Sent_entity	2803	373	358	Entities that are sending or receiving emotions

CTAS: Chinese textual affective structure

x denotes the input sentence, where the i^{th} token in the sentence is represented as x_i, its corresponding label is y_i, and the entire label sequence is denoted as y.

1. CAL (Margatina et al., 2021a). CAL focuses on selecting a set of contrastive samples, i.e., samples that are similar in the model feature space but have maximally different predictive likelihoods in the model outputs. By comparing the prediction results of similar samples, CAL aims to identify challenging samples and enable the model to better understand decision boundaries.

2. LTP (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.

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

3. LC (Culotta and McCallum, 2005). LC sorts the examples in ascending order according to the probability that the model assigns to the most likely label sequence. 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 informative samples and reduce classification errors.

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

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

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

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

$$\phi_{\text{MTE}}(x) = -\sum_{i=1}^{n} \sum_{m=1}^{M} P(y_i = m) \log_2 P(y_i = m), \tag{8}$$

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

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

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

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 $\min_i \max_j P(y_i = j|x_i)$ is the minimum of these maximum probabilities over all tokens x_i in the input x.

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

4.3 Experimental setup

We used an early-stopping technique to determine the optimal model from the development set, with a maximum training duration of 100 epochs. If no improvement in the score was observed within 60 epochs, the training process was terminated. Subsequently, we saved the optimal model and evaluated its performance on the test set. Consistently, we employed BERT-CRF as the base model and BERT-base-Chinese as the pre-training model. Our learning rate was set to 5×10^5, with a batch size of 64. To facilitate efficient batch training, we imposed a maximum sample length of 256. For initial training, we selected a mere 5% of the available data throughout the iteration process. All experiments were conducted on an Nvidia GeForce RTX3090 GPU.

4.4 Experimental results

We assessed the entity-level F1 scores of the baselines and the MAL method using the CTAS dataset. The baseline score was reported by Xiong et al. (2023), who trained a fully supervised model and obtained an F1 score of 0.5479. The goal of the AL method is to reach or exceed this baseline score while using as few training samples as possible. To ensure consistency and accuracy, we used fixed random seeds for initialization, effectively reducing the potential random effects.

We present the performance of all methods on the CTAS dataset, as shown in Fig. 2. The MAL method was the most competitive and outperformed all baselines. It achieves a performance comparable to that of the full model as early as the third iteration, and its performance continues to improve in subsequent iterations. The confidence-based NLC method, which avoided querying longer sentences (unlike the LC method), was the top performer among the baselines. Although the LC and NLC methods are straightforward and intuitive, the confidence-based query strategy remains highly competitive.

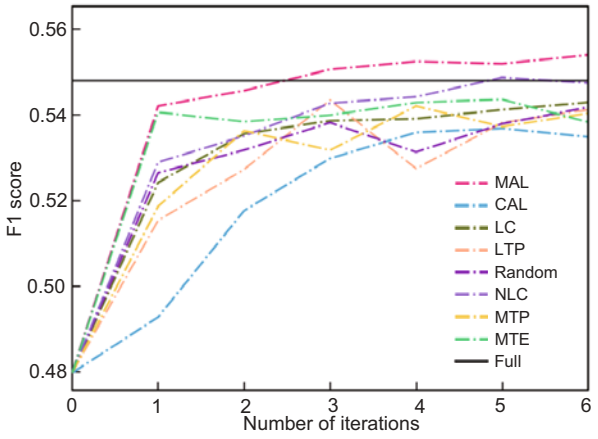

Fig. 2 Performance of MAL and baselines on the CTAS dataset. CTAS: Chinese textual affective structure; MAL: multilevel active learning; CAL: contrastive active learning; LC: least confidence; LTP: lowest token probability; Random: select samples in a random manner; NLC: normalized LC; MTP: minimum token probability; MTE: maximum token entropy; Full: training model with all of the available data. References to color refer to the online version of this figure

The MTE method overcame the limitation of the TE method by querying longer sentences. Nevertheless, sentence length should not be a major concern for the sentiment structure recognition task. Hence, the performances of these two methods in the task were unremarkable. Similarly, the MTP method did not yield better results in selecting the most informative sentences. The LTP method, which considers both global and local information simultaneously, failed to achieve competitive performance either.

CAL, which was the worst-performing query strategy, also took into account samples that were similar in the 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 MAL and CAL, it becomes clear that selecting samples that are closely aligned in the feature space and have smaller divergence enhances model learning. These findings serve as valuable guidance for selecting an AL method tailored to a specific task, ultimately aiding in the optimization

of labeled data utilization and the improvement of model performance.

5 Ablation studies

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

Partners for x_p are determined 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 named as RandomP are depicted in Fig. 3. It is evident that the RandomP 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.

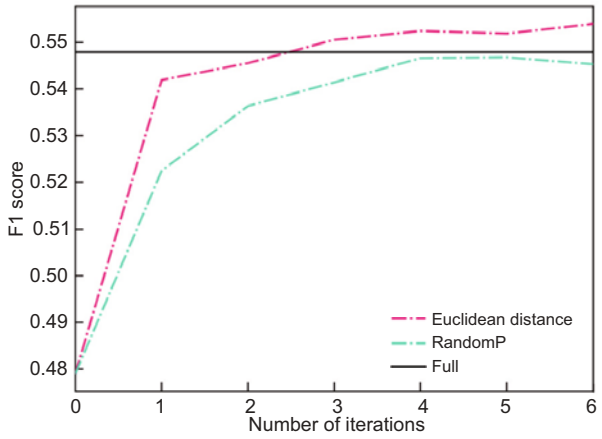

Fig. 3 Impact of partner selection on the performance, where RandomP refers to selecting partners in a random manner. References to color refer to the online version of this figure

We also examined the influence of the number of partners on MAL performance, as shown 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. As the number of partners processed increases, so does the uncertainty. Selecting more partners for x_p may result in including more distant partners, even though these partners are relatively closer to x_p than other unselected samples.

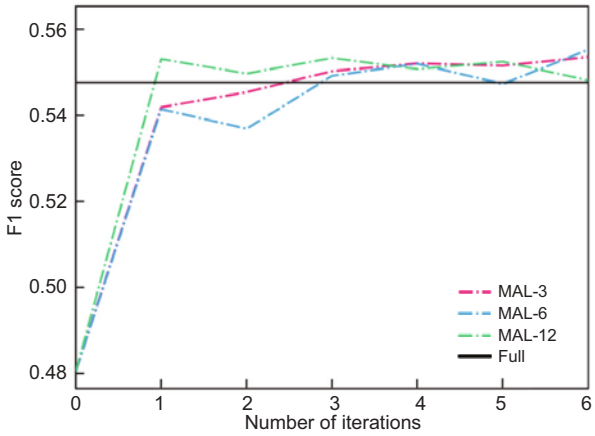

Fig. 4 Impact of the number of partners on performance. MAL-3, MAL-6, and MAL-12: multilevel active learning method with 3, 6, and 12 partners, respectively. References to color refer to the online version of this figure

6 Analysis

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

Evidently, regardless of whether we select 3 or 12 partners per iteration, MAL consistently selects samples with less labels than RandomP. As the number of iterations increases, MAL demonstrates the capability to select additional samples with rare labels 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 samples each method tends to select. Specifically, we evaluated the number of labeled samples selected by each method. 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.

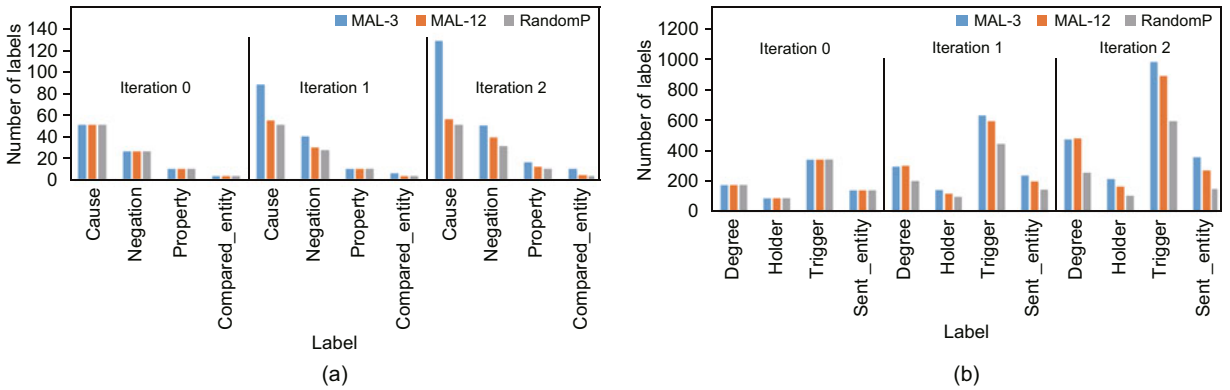

Fig. 5 Distribution trends of different labels in selected samples when using different methods. RandomP refers to selecting partners in a random manner. MAL: multilevel active learning. References to color refer to the online version of this figure

Table 2 Number of labels in the training set

Label	Number of labels								
	Initial	LC	NLC	MTP	MTE	LTP	RandomP	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

CTAS: Chinese textual affective structure; MAL: multilevel active learning; CAL: contrastive active learning; LC: least confidence; LTP: lowest token probability; NLC: normalized LC; MTP: minimum token probability; MTE: maximum token entropy

This analysis provides valuable information on the label distribution and the effectiveness of the AL methods 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 sample to sample.

In our investigation, we leveraged 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 capability 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 label 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 methods, the number of labels for the samples selected using MAL is the lowest except for the compared_entity. When the size of the training set was doubled, the number of labels associated with the selected samples did not double. This indicates that in the CTASA task, increasing the amount of information in the training data does not improve the training effectiveness. Conversely, too much redundant information may interfere with the model, whereas our proposed method can find truly valuable samples for the model.

In summary, experimental results show that MAL performs well in picking the number of sample labels, and the number of the selected sample labels is less compared with other methods. This further demonstrates the ability of our method to

discover samples 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 method combines sentence-level and word-level information. First, we used 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 based on their sentence-level features, aiming for partners similar to x_p. 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 among samples and thus select samples with smaller differences for annotation.

The affective structure identification task of the Chinese social media corpus involves expressions rich in diversity and non-standardization. Therefore, we considered the information of samples from both sentence- 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 method of fusing sentence- and word-level information can capture the features of the samples more comprehensively and improve the performance of the model on the Chinese social media corpus.

8 Conclusions

This study introduces the MAL method, which significantly enhances TASA 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 method 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 baselines but also achieves superior performance while reducing labeling costs. This study represents pioneering research in applying AL specifically to CTASA, marking a significant advancement in the field. Moving forward, our focus will be on refining the MAL method 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. *Proc Mach Learn Res*, 16:127-139.

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 Inform*, 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, 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 Neural 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. *Neur Process Lett*, 54(3):2433-2454. https://doi.org/10.1007/s11063-021-10737-x

Margatina K, Vernikos G, Barrault L, et al., 2021a. 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

Margatina K, Barrault L, Aletras N, 2021b. 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

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 Annual 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, USA.

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, 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. *Inform 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.

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