Active learning with complementary sampling for instructing class-biased multi-label text emotion classification

Xin Kang, *Member, IEEE,* Xuefeng Shi, *Non-Member,* Yunong Wu, *Non-Member,* Fuji Ren, *Senior Member, IEEE,*

Abstract—High-quality corpora have been very scarce for the text emotion research. Existing corpora with multi-label emotion annotations have been either too small or too class-biased to properly support a supervised emotion learning. In this paper, we propose a novel active learning method for efficiently instructing the human annotations for a less-biased and high-quality multi-label emotion corpus. Specifically, to compensate annotation for the minority-class examples, we propose a complementary sampling strategy based on unlabeled resources by measuring a probabilistic distance between the expected emotion label distribution in a temporary corpus and an uniform distribution. Qualitative evaluations are also given to the unlabeled examples, in which we evaluate the model uncertainties for multi-label emotion predictions, their syntactic representativeness for the other unlabeled examples, and their diverseness to the labeled examples, for a high-quality sampling. Through active learning, a supervised emotion classifier gets progressively improved by learning from these new examples. Experiment results suggest that by following these sampling strategies we can develop a corpus of high-quality examples with significantly relieved bias for emotion classes. Compared to the learning procedures based on traditional active learning algorithms, our learning procedure indicates the most efficient learning curve and estimates the best multi-label emotion predictions.

Index Terms—Active Learning, Complementary Sampling, Class-biased Multi-Label Classification, Text Emotion.

1 Introduction

As a joint research of affective computing and natural language processing, text emotion classification aims at helping machines to understand human emotions through natural language expressions. The most widely studied emotion classification problem for natural language was to separate the positive and negative emotion polarities for a piece of text [1], [2], [3], [4], [5], while more sophisticated studies tried to distinguish different emotion categories [1], [2], [3], [5], [6], [7], [8], [9], such as Anger, Disgust, Fear, Happiness, Sadness, Surprise [10], Love, and Expectation [11]. More recent studies have taken the fusion of different emotions in the linguistic expression as a multi-label classification problem [12], [13], which aimed at a more comprehensive understanding of the text emotions. Because emotion understanding and language comprehension are strongly correlated with each other [14], [15], [16], understanding the emotional states in natural language promotes the understanding of people's thought as the intelligence of many intention-understanding-related applications, such as market opinion analysis [17], [18], [19], public opinion analysis [20], [21], [22], [23], and mental health diagnosis [24], [25], [26], [27], [28]. We focus on the *multi-label* emotion classification problem and hope this work will benefit the text emotion research for all different granularities.

- *X. Kang and F. Ren were with Faculty of Engineering, Tokushima University, 2-1, Minamijyousanjima-cho, Tokushima 770-8506 Japan. E-mail: kang-xin@is.tokushima-u.ac.jp, ren@is.tokushima-u.ac.jp.*
- *X. Shi was with School of Computer and Information, Hefei University of Technology, Danxia Road, Hefei, Anhui 230601 China. E-mail: 2020010107@mail.hfut.edu.cn.*
- *Y. Wu was with Chengdu Senton Netease Co., Ltd. Chengdu, Sichuan 610000 China. E-mail: raino.wu@gmail.com.*

Previous research has suggested that learning a text emotion classification model requires large amounts of labeled examples [29], [30], [31], [32]. However, building a high-quality emotion corpus entirely by manpower could be very challenging. The direct reason is that inferring an author's emotions requires the annotators to empathetically experience every word in that message [33], [34], which might cause a heavy mental strain to these annotators [1], [35], [36]. Many previous works suggest that annotating an emotion corpus with only a few thousand sentences would be very labor intensive [37], [38]. In addition, the nature of imbalance in the emotion distribution has made annotating a high-quality emotion corpus even more difficult. All existing moderate-sized corpora with more than 10,0000 examples have suffered severely from the class-imbalance problem [11], [39], [40], [41], [42], [43]. For example, in the Tweet-based corpus CrowdFlower [42], the classes for Joy and Neutral have covered 9,370 and 9,220 examples respectively, which are 51.51 and 52.34 times larger than that of Disgust. This problem is even more severe for some real-life dialogue-based corpora, such as the DailyDialog corpus [41], in which the class for Neutral is 491.79 times larger than that of Fear. These problems have severely restricted the development of a corpus of considerable size and balanced labels to advance the text emotion research, especially in the case of multi-label emotion classification.

In this paper, we propose a novel active learning method to efficiently instruct the human annotation for a less biased, multi-label emotion corpus, by which the most informative raw examples for model update with the most needed labels for class-balance are preferentially picked from an unlabeled resource for the human annotators. Active learning [44] is a compound sampling procedure, which iteratively queries an unlabeled resource for

the potentially valuable training examples. For active learning in a *binary-class classification* problem, estimating the informative value for an unlabeled example can be very straightforward by following the uncertainty of model prediction [45], the version space reduction [46], the expected model change [47] or error reduction [48], [49], and can be refined with a weighted density [50]. Favorably selecting an unlabeled example for the binary-minority class is also straightforward, by drawing samples towards the opposite side of class distribution in the existing corpus [51], [52], [53], [54], [55], [56], [57], if a prior estimation could be made for these examples. For a problem of *multi-class classification*, in which each example belongs to one of multiple categories, active learning methods could be adjusted by integrating the categorical maximum or mean of the informative values [58], [59], [60], [61] for an informative sampling and by integrating the rest categories as one class [62], [63], [64], [65], [66] for a class-balance sampling. However, to the best of our knowledge, there has been no research of active learning for a class-biased *multi-label classification* problem.

In the proposed method, we first estimate the informative values of unlabeled examples for human annotation instruction by integrating the prediction uncertainties in these examples given a multi-label emotion classifier and two distinct properties of their syntactic features among the labeled and unlabeled feature spaces. Specifically, the uncertainty criterion evaluates a batch-wise maximum integration of the emotion prediction entropies for unlabeled examples and finds those with the most potential for unraveling the prediction uncertainties for the current emotion classifier. The representativeness criterion evaluates a mean integration of the pairwise syntactic similarity among the pool of unlabeled examples to avoid the outlier-querying problem in the uncertainty sampling [44], [49], [67], [68]. The diverseness criterion evaluates a maximum-minimum integration of the pairwise syntactic similarity between the pool of unlabeled examples and the set of labeled examples to avoid the duplicative querying problem [69] in the uncertainty sampling.

Second, we propose a novel complementary sampling method to compensate a biased multi-label emotion corpus through active learning, by which more human efforts could be led to the annotation of examples of minority categories. The sampling method is based on a complementariness criterion which measures a probabilistic distance between the expected distribution of emotion labels in a temporary corpus and the uniform distribution of emotion labels. Each temporary corpus corresponds to the integration of a new example to the currently labeled emotion corpus, with a probabilistic emotion prediction. The normalized count of labels in this temporary corpus then corresponds to a new distribution of emotion labels under the model expectation, if this unlabeled example were finally selected by active learning. Because balanced labels in a training set are generally important to the learning of a supervised classification model [70], [71], [72], choosing an uniform distribution over all emotion categories as the optimal distribution also benefits the learning of a multi-label emotion classification model. By minimizing this complementariness criterion, the active learning algorithm could find the most potential example to compensate the current training data for the minority labels. We employ the Kullback-Leibler (KL) divergence, the cross entropy (CE) distance, and the earth mover's (EM) distance for measuring the probabilistic distances from separate perspectives, and compare their results for compensating an emotion corpus through active learning.

The major contributions of this paper are listed as below:

- A complementary sampling method in active learning is proposed by compensating a class-biased training data for the minority-class labels.
- Both the leaning efficiency and the ceiling performance for text emotion classification will be significantly improved with the proposed method.
- The main reason of failing to retain a label balance in active learning given an unlabeled but extremely biased data resource, is clarified.

The rest of this paper is arranged as follows. The related work of text emotion classification and active learning is reviewed in Section 2. We illustrate the proposed active learning method for the class-biased multi-label text emotion classification problem in Section 3. In Section 4, we describe our experiment of text emotion classification, compare the different sample selection strategies in active learning for constructing the training data, and analyze the reason for the retained class-bias problem in active learning, based on the social network messages from Sina Weibo. Finally, Section 5 concludes this paper.

2 Related Work

2.1 Active Learning

Active learning [44] is a procedure for efficiently and progressively improving a supervised machine learning model, in the cases where the number of labeled examples are limited and difficult to obtain but unlabeled examples are abundant and easy to acquire. Although most research of active learning has focused on the classification task [29], [44], [45], [47], [48], [50], [58], [59], [60], [61], [67], [68], [73], [74], [75], [76], [77], [78], [79], [80], [81], [82], there were quite a few that focused on the regression problem [51], [52], [53], [54], [55], [56], [57], [62], [63], [64], [65], [66], [83], [84], [85]. In the active learning algorithms for *binary* or *multi-class* classifications, the unlabeled examples were selected from either a pool [45], [50], [81], a stream [48], [86], or a synthesis generator [87], [88], based on a potential measurement for the unlabeled examples. Popular methods for the potential measurement included the uncertainty of model prediction [45], [50], [89], [90], the expected reduction of version space for the model [46], [91], [92], the expected model change [47], [93], [94], the expected reduction of model variance [48], [95], the expected reduction of future error [48], [49], [67], and the information density of the unlabeled example [50], [69], [79], [96]. The selected examples were incorporated into the training data, together with the labels annotated by human experts to progressively improve a supervised learning model through the positive feedback-loop.

The active learning procedure could be extended to a *multi-label* classification problem [97], in which each example would be associated to several class labels. It was argued that developing the training data for a multi-label classification problem was more difficult, expensive, and time-consuming than that for a binary or multi-class classification problem [59], [61], [76], [78], [98]. The key idea behind the multi-label extensions in active learning was to integrate the measurement of sample potentials over multiple class labels. For example, Zhang et al. [58] employed the maximum of uncertainties over all class labels as an integrated potential for the unlabeled image examples in a multi-view multi-label active learning algorithm. Reyes et al. [59] ranked the prediction uncertainty of all unlabeled examples in a pool for each class label,

and took the mean of rank ids for all class labels as an integrated potential measurement for each unlabeled example. Brinker [60] proposed a best worst-case approach for potential measurement based on the one-versus-all multi-label support-vector machines (SVM) classifier, in which the maximum of expected volume reductions of the version spaces for each single SVM classifier was taken as the integrated potential for an unlabeled example. According to the theory of convex set [99], the volume of an SVM version space would be reduced exponentially in terms of the number of related chosen examples in an ideal situation through active learning. Yang et al. [61] estimated the binary SVM classification errors on each class by the size of version space [82] and employed the summation of version space reductions for each binary SVM classifier as the expected error reduction, that is, the integrated potential for the unlabeled examples. And the rest of the active learning procedures for a multi-label classification were exactly the same as that for a binary or multi-class classification.

Although the class-imbalance problem was of crucial importance in machine learning, there were very few active learning researches for this problem, not to mention the multi-label active learning researches. Ertekin et al. [74] first demonstrated that active learning through a simple uncertainty sampling was able to provide more balanced data for *multi-class classification* than a random sampling. Doyle et al. [73] tailored the sampling procedure in active learning for a *binary-class classification* problem, by repeatedly querying annotators for the unlabeled examples until a specified number of majority-class and minority-class examples were found. Because the imbalance of class distribution was not changed in the sampling, many majority-class annotations had to be discarded in the procedure. Li et al. [77] proposed a co-selecting approach for automatically annotating majority-class examples for a *binary-class* corpus. The majority-class examples were selected and determined based on the certainty and uncertainty predictions given by separate classifiers over the unlabeled resource, while the minority-class examples were similarly selected but manually determined. Because predictions on the majority-class examples were not guaranteed to be correct, the result corpus could be falsely balanced with many incorrect majority-class patterns. Hospedales et al. [75] employed active learning to discover new rare classes for the highly imbalanced *multi-class classification* problems. However, because the learning target was only rewarded by the discovery of new classes rather than new examples of the existing minority-classes, the class-imbalance problem was never solved. Khanchi et al. [63], [65], [66] and Brust et al. [62] revised the binary active learning algorithm for the class-biased *multi-class classification* problems, in which the inverse proportion of label counts was employed to prioritize the sampling of minority-class examples. Zhang et al. [64] further adapted this algorithm for the streaming data with both class-imbalance and concept drift problems. A dynamic classifier and a stable classifier were learned with the previous data block and the entire data stream respectively, for making joint predictions on the unlabeled examples. To the best of our knowledge, there has been no research of active learning for a class-biased *multi-label classification* problem.

2.2 Text Emotion Classification

The social information processing (SIP) theory [100] suggested that for conveying emotions in non-face-to-face communications, such as sending social network messages, people tended to more frequently use their language, tones, and other verbal clues. And the experimental studies of SIP theory [101], [102] further demonstrated that text was as effective in conveying emotions as other face-to-face communication approaches, which proved the validity of the text emotion classification research on the basis of psychological principles [1]. Hancock et al. [102] examined the process of emotion expression in text-based systems and found that some basic patterns in the texts, such as negations and exclamation points, were significant strategies for people's emotion judgement. These results on the other side proved that human emotions could be detected by the analysis of emotion-related linguistic features through natural language processing.

In affective neuroscience, emotion was defined as the discrete and consistent responses to internal or external events that have a particular significance for the organism, with a short-term duration [103]. However, there have been no unified models of emotion in theoretical studies. Ekman [10] proposed six basic emotions, that is Anger, Disgust, Fear, Happiness, Sadness, and Surprise as the universally recognizable set of emotions by humans regardless of race, culture, and language. The Ekman emotion model was also the most widely used in computer science research [1]. Plutchik and Kellerman [104] employed four bipolar axes, that is Joy vs. Sadness, Anger vs. Fear, Trust vs. Disgust, Surprise vs. Anticipation for modeling emotions in a multi-dimensional space. Because of the difficulty of precisely arranging human emotions in the multi-dimensional space, this model was not as widely used in computer science. Shaver et al. [105] proposed a tree-structured model in which Anger, Fear, Joy, Love, Sadness, and Surprise were the emotions on the main branches, each of which had subordinate categories such as Affection, Lust, and Longing. More recently, Quan et al. [11] proposed an extended model based on the Ekman's six emotions, specifically for the text emotion analysis. Two basic emotions of Love and Expect were added to the set with Anxiety taking the original place of Fear. Together with the release of a large and thoroughly annotated emotion corpus Ren-CECps[1], this model had also attracted many related researches [18], [32], [80], [106], [107].

Most text emotion classification research considered word as an important feature in building the classification models [6], [9], [30], [36], [108], [109], [110]. Besides, many of the manually built emotion lexicons [8], [111], [112], [113], [114] which encoded the relationship between words and emotion labels were also widely employed for text emotion classification. Although the development of emotion lexicons required a lot of time and patience, the encoded word-emotion relationship still contained large amounts of conflicts and had been criticized for the lack of context sensitiveness in many real world problems [11], [30], [36], [115]. To generate features in an efficient and effective manner, Li et al. [115] proposed a regression method to infer the multi-dimensional affective representation of words based on the semantic word embeddings. The extended emotion lexicons were proved effective for improving the sentiment classification results based on multiple public data sets. Other word emotion recognition models with context sensitivity were also proposed [116], [117], [118], in which the syntactic modifications with respect to negations, degree adverbs, and special punctuations in a sentence were considered as indicative emotional features.

Supervised learning methods were widely adopted for text emotion classification. For example, Colneriĉ and Demsar [6] em-

1. Ren-CECps is a Chinese emotion corpus (http://a1-www.is.tokushima-u.ac.jp/member/ren/Ren-CECps1.0/DocumentforRen-CECps1.0.html).

ployed the word and character-based recurrent and convolutional neural networks for Twitter emotion classification, and proposed a training heuristic for the unison model to transfer the last hidden layers of all emotion prediction networks. Yang et al. [24] proposed a hybrid framework with a paragraph vector and SVM based model for inferring the emotional conditions of individuals from their interview transcripts, and for assessing their depression symptoms with the assistance of an audio-visual multi-model and a random forest model. Kang et al. [108] proposed a kernel function based on the SVM classifier to evaluate the emotion similarity of two sentences with high dimensional emotion features of words, for the sentence emotion classification.

Because the size of available training data for emotion classification was usually small, many semi-supervised methods were also considered for inferring emotions through text. Phan et al. [12] proposed a semi-supervised multi-label model for emotion classification in the conversation transcripts, based on the word semantic meanings and the language structures learned with an auto-encoder neural network. Li et al. [7] incorporated the learned semantic domain knowledge through unsupervised models into a novel neural network structure, and improved the emotion classification for online comments with this model by an enhanced network interpretability. Ren and Wu [4] proposed a matrix factorization method to learn the underlying representation of language features, based on the social network contexts and the topical contexts. These unsupervised language features were then utilized for improving the prediction of tweet emotions. Kang et al. [30] proposed a semi-supervised model for online text emotion classification, in which the point-wise mutual information (PMI) of words and emotion labels were propagated through a dynamic Bayesian network. Ren et al. [13], [32] derived generative models of documents given latent topics and knowledge of word-emotion correlations based on a labeled corpus, and inferred the posterior probabilities of document emotions and word emotions through the Gibbs sampling algorithm. Rao et al. [119], [120] investigated the relation between emotions and the context-dependent topics through a Bayesian network and learned the posterior probability of document emotions through an EM algorithm. Summa et al. [121] employed a graph-based semi-supervised learning method, to encode and propagate the relations between neighboring tweets with respect to their emotion labels.

3 Class-Biased Multi-Label Emotion Classification

3.1 Supervised Model for Emotion Classification

Suppose we have an emotion corpus $\mathscr{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{M}$ of M labeled examples, in which i indexes a labeled example. $\mathbf{x}_i \in \mathbb{R}^N$ denotes the feature vector of the ith instance, and $\mathbb{R}^N$ is the real-valued N-dimensional input feature space. Given a large number of raw messages from social network services, we filter out the advertisements, remove the none textual contents, regularize the functional expressions into their functional basics, and take the segmented words in their basic forms for feature representation. $\mathbf{y}_i \in \mathbb{B}^K$ denotes the target vector of the ith instance, that is, the binary labels for K emotion categories, and $\mathbb{B}^K$ is a K-dimensional binary-valued target space. The set of target emotion categories $\mathscr{E}$ includes Anxiety, Anger, Sorrow, Hate, Joy, Love, Expect, Surprise, and Neutral. For multi-label emotion classification, each $\mathbf{y}_i$ consists of one or more positive labels.

A supervised model for predicting the emotion labels $\mathbf{y}$ for an input example $\mathbf{x}$ by

$$f : \mathbb{R}^N \to \mathbb{B}^K \tag{1}$$

is decomposed into K separate binary supervised classification functions

$$f_k : \mathbb{R}^N \to \mathbb{B}, \tag{2}$$

with each f_k predicting the existence of the kth emotion label. In this research, we employ the logistic regression algorithm for a binary classification. Specifically, each binary classification function f_k is defined by

$$\hat{\mathbf{y}}_k = f_k(\mathbf{x}) \tag{3}$$

$$= 1\{\sigma(\mathbf{w}_k \cdot \mathbf{x} + b_k) > 0.5\}, \tag{4}$$

in which $\mathbf{w}_k$ and b_k are the weight and bias parameters of the classification function f_k, and σ is the sigmoid function given by

$$\sigma(z) = \frac{1}{1+\exp(-z)}. \tag{5}$$

The sigmoid function σ generates real-valued predictions in the $(0,1)$ interval, which are usually interpreted as the probability of positive predictions. Therefore, $\sigma(\mathbf{w}_k \cdot \mathbf{x} + b_k)$ renders the model's probabilistic prediction for the existence of emotion label k in example $\mathbf{x}$, for which we use $\mathbf{p}_k$ to represent. And the outputs of the sigmoid function can be transformed into binary predictions, by feeding the truth examination $\sigma(z) > 0.5$ to an indicator function $1\{\cdot\}$. We use $\hat{\mathbf{y}}_k$ to indicate the binary prediction of the kth emotion label in example $\mathbf{x}$.

The above binary-decomposition method is able to solve our multi-label classification problem for all emotion categories except for Neutral, that is, the Kth emotion category in $\mathscr{E}$. Because an example $\mathbf{x}$ can be either Neutral or express one or more other categorical emotions, the probabilistic prediction for the Neutral class is inferred through

$$\mathbf{p}_K = 1 - \max\{\mathbf{p}_k | k = 1, \cdots, K-1\}, \tag{6}$$

in which we take the complement of the probability with respect to the most probable emotion label. And the binary prediction for Neutral is simply given by

$$\hat{\mathbf{y}}_K = 1\{\mathbf{p}_K > 0.5\}. \tag{7}$$

3.2 Active Learning for Class-Biased Multi-Label Classification

We introduce a novel active learning algorithm, which finds useful examples from the unlabeled resources and directs human annotators to these examples to build a less biased and high quality multi-label emotion corpus. We denote the unlabeled data as $\mathscr{U}$ and the labeled corpus as $\mathscr{D}$. Our algorithm repeatedly evaluates the potential of annotations from $\mathscr{U}$ for compensating the class-imbalance in $\mathscr{D}$ with a complementariness criterion and quantifies the candidate examples in model updating, outlier filtering, and diversity expanding with the uncertainty, representativeness, and diverseness criteria respectively. Throughout active learning, batches of examples are selected from the incoming new resources $\mathscr{U}$'s, annotated with multiple emotion labels, and appended to the labeled corpus $\mathscr{D}$. With these high quality examples, we can efficiently improve the multi-label emotion classification model, which in turn helps the evaluation of unlabeled examples for active learning.

The proposed active learning procedure is depicted in Algorithm 1. The **complementariness** criterion evaluates the potential of an unlabeled example $\mathbf{x} \in \mathscr{U}$ for compensating class-imbalance in $\mathscr{D}$ by

$$\mathrm{c}(\mathbf{x}) = D_*(P(\tilde{\mathscr{D}}(\mathbf{x})), Q), \tag{8}$$

in which D_* measures the distance between two distributions P and Q, and P is the expected emotion-label distribution for a temporarily integrated corpus

$$\tilde{\mathscr{D}}(\mathbf{x}) = \mathscr{D} \cup (\mathbf{x}, \hat{\mathbf{y}}), \tag{9}$$

and $Q \sim \mathrm{unif}\{1, K\}$ is an uniform emotion-label distribution. By integrating the model prediction $\hat{\mathbf{y}}$ and unlabeled example $\mathbf{x}$ into the current training data $\mathscr{D}$, the expected emotion-label distribution $P(\tilde{\mathscr{D}})$ can be further expanded as

$$\begin{aligned} P(\tilde{\mathscr{D}}(\mathbf{x})) &= \mathrm{E}\left[\frac{\sum_{k=1}^{M} \mathbf{y}_i + \hat{\mathbf{y}}}{\sum_{i=1}^{M}\sum_{k=1}^{K} \mathbf{y}_{ik} + \sum_{k=1}^{K} \hat{\mathbf{y}}_k}\right] \\ &= \frac{\sum_{k=1}^{M} \mathbf{y}_i + \mathbf{p}}{\sum_{i=1}^{M}\sum_{k=1}^{K} \mathbf{y}_{ik} + \sum_{k=1}^{K} \mathbf{p}_k}, \end{aligned} \tag{10}$$

in which $\mathbf{y}_i$ is a label vector for the ith instance in $\mathscr{D}$ and $\mathbf{y}_{ik}$ is the kth binary label in the vector. We move the outside expectation to the nominator and denominator respectively and estimate the emotion labels with probabilistic predictions $\mathbf{p}_k$ from the model. The setting of an uniform distribution as the optimal emotion label distribution ultimately benefits our supervised learning.

The unlabeled example $\mathbf{x} \in \mathscr{U}$ which maximizes the complementariness criterion should have the most significant potential for class compensation in $\mathscr{D}$. Three probabilistic distance measurements, that is, the Kullback-Leibler (KL) divergence, the cross entropy (CE) distance, and the earth mover's (EM) distance are considered for D_* in Eq. 8. Specifically, KL distance evaluates the log difference between two emotion label distributions P and Q under the expectation of P

$$D_{\mathrm{KL}}(P, Q) = \sum_{k=1}^{K} P_k \log \frac{P_k}{Q_k}. \tag{11}$$

CE distance evaluates the distance of P and Q in the sense of encoding sizes in the information theory

$$D_{\mathrm{CE}}(P, Q) = -\sum_{k=1}^{K} Q_i \log P_i. \tag{12}$$

And EM distance of P and Q, which is given by

$$D_{\mathrm{EM}}(P, Q) = \frac{\sum_{i=1}^{K}\sum_{j=1}^{K} f_{ij} d_{ij}}{\sum_{i=1}^{K}\sum_{j=1}^{K} f_{ij}}, \tag{13}$$

evaluates the weighted transportation of probability amounts from P to Q through an optimized process until P is geometrically reshaped as Q. f_{ij} indicates the amount of transported probability from the ith bin of P to the jth bin of Q, and d_{ij} indicates the transportation weight. The transportation detail for EM distance is out of the scope of this paper, though, it has been considered as a well-defined metric to evaluate the geometric distance between two distributions, with important properties such as symmetry and triangle inequality.

Fig. 1 shows an example of the complementary sampling procedure. Before complementary sampling, the emotion corpus consists of biased labels as indicated by the empty bars in the left-hand side subplots. An unlabeled example $\mathbf{x}$ is then selected based on the complementariness criterion, with the categorical and probabilistic emotion predictions shown in the middle and right-hand side subplots. The selected examples are different for different distance measurements, but all examples share the same categorical predictions and the highest probabilistic predictions for emotion Sorrow. Only the KL based sampling selects the example with a high probability for Expect. The expected label distributions are shown by the stacked bars in the left-hand side subplots. In this case, all selected examples have the potential to relieve the class-bias problem in the corpus.

Furthermore, candidate examples are qualitatively evaluated based on their potential for unraveling prediction uncertainties for the emotion classifier, their syntactic representativeness of the other candidate examples to avoid querying outliers, and their syntactic diverseness to the labeled examples to avoid the duplicative querying problem. The procedure of active learning based on these evaluations are shown in Algorithm 1. An **uncertainty** criterion is given by

$$\mathrm{u}(\mathbf{x}) = \max\{\mathrm{H}(\mathbf{p}_k) \,|\, k = 1, \cdots, K\}, \tag{14}$$

which evaluates an unlabeled example $\mathbf{x}$ with the maximum integration of entropies for its emotion predictions in $\mathbf{p}$. The entropy is given by

$$\mathrm{H}(\mathbf{p}_k) = -\mathbf{p}_k \log \mathbf{p}_k - (1 - \mathbf{p}_k) \log (1 - \mathbf{p}_k), \tag{15}$$

which monotonically increases with the model uncertainty. Eq. 14 indicates that if the multi-label classification model is very uncertain in predicting at least one emotion category for the unlabeled example $\mathbf{x}$, we would have a very large uncertainty value for $\mathrm{u}(\mathbf{x})$. Because these examples provide informative patterns for learning a supervised emotion classification model, we first select them from the unlabeled resource as the candidates for future human annotation.

Querying an unlabeled resource with the uncertainty criterion can cause an outlier-querying problem for active learning, in which abnormally behaved outliers are selected for updating the training corpus. Because in most cases the abnormally behaved examples are syntactically different from the other examples, we employ a **representativeness** criterion in active learning to filter out those syntactically different examples. The representativeness criterion is given by

$$\mathrm{r}(\mathbf{x}) = \frac{1}{|\mathscr{U}| - 1} \sum_{\mathbf{x}' \in \mathscr{U}_{-\mathbf{x}}} \mathrm{sim}(\mathbf{x}, \mathbf{x}'), \tag{16}$$

in which $\mathbf{x}$ is a candidate example from the unlabeled resource for evaluation, $\mathbf{x}' \in \mathscr{U}_{-\mathbf{x}}$ indicates any unlabeled example except $\mathbf{x}$, and sim evaluates a pairwise similarity of two examples. We employ the opposite of Euclidean distance for sim calculation by

$$\mathrm{sim}(\mathbf{x}, \mathbf{x}') = -D_{\mathrm{EU}}(\mathbf{x}, \mathbf{x}'), \tag{17}$$

$$D_{\mathrm{EU}}(\mathbf{x}, \mathbf{x}') = \sqrt{\sum_{j=1}^{N} (\mathbf{x}_j - \mathbf{x}'_j)^2}, \tag{18}$$

in which $j \in \{1, \cdots, N\}$ denotes the index of the syntactic features in $\mathbf{x}$ and $\mathbf{x}'$. Because outliers are syntactically distant from the normal examples, their sim values and the representativeness scores should be correspondingly small compared to the normal examples. In the active learning algorithm, we encourage selecting examples with large representativeness scores to avoid the outlier-query problem.

In active learning, finding syntactically diverse examples is equally important as finding the class-balanced ones, which allows the supervised models to observe different possible feature

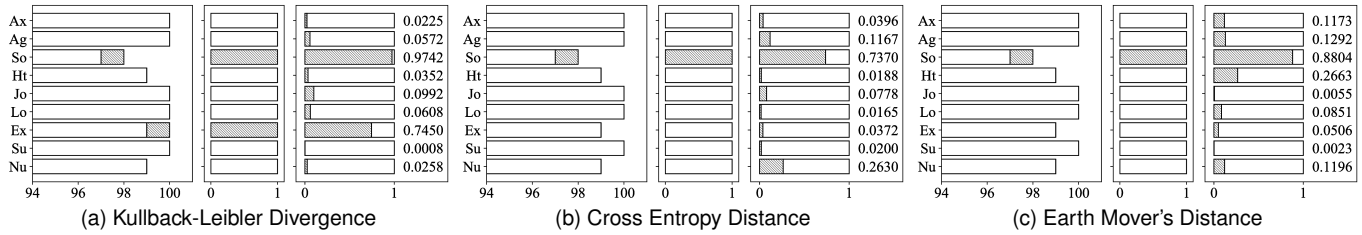


Fig. 1. Example of compensating the training data with an unlabeled example under different probabilistic distance measurements.

combinations and therefore avoids over-fitting. The **diverseness** criterion measures the potential of an unlabeled example $\mathbf{x} \in \mathscr{U}$ for compensating the syntactic diversity of the corpus $\mathscr{D}$ by

$$\mathrm{d}(\mathbf{x}) = \min_{\mathbf{x}' \in \mathscr{D}} D_{\mathrm{EU}}(\mathbf{x}, \mathbf{x}'), \tag{19}$$

that is, the minimum of the Euclidean distance D_{EU} between $\mathbf{x}$ and all labeled examples $\mathbf{x}' \in \mathscr{D}$. By measuring such distance, we understand the potential of diversity increments for $\mathscr{D}$ given by each unlabeled example $\mathbf{x}$.

Algorithm 1: Active learning for class-biased multi-label emotion classification.

input : $\mathscr{D}$ - original training data
$\mathscr{U}$ - unlabeled incoming resource
λ_{u} - selection ratio for uncertainty criterion
λ_{r} - selection ratio for representativeness criterion
λ_{d} - selection ratio for diverseness criterion
λ_{c} - selection ratio for complementariness criteria
output: $\mathscr{D}$ - updated training data

repeat m **times**
  learn a multi-label emotion classifier f from $\mathscr{D}$;
  pick $\mathscr{U}_{\mathrm{i}}$ as a batch of unlabeled data;
  predict $\hat{\mathscr{Y}}_{\mathrm{i}}$ and $\mathscr{P}_{\mathrm{i}}$ for $\mathscr{U}_{\mathrm{i}}$ with f;
  repeat n **times**
    $\mathscr{U}_{\mathrm{u}} \leftarrow \mathrm{maxpartition}(\mathscr{U}_{\mathrm{i}}, \mathrm{u}(\mathscr{U}_{\mathrm{i}}), \lambda_{\mathrm{u}}|\mathscr{U}_{\mathrm{i}}|)$;
    $\mathscr{U}_{\mathrm{r}} \leftarrow \mathrm{maxpartition}(\mathscr{U}_{\mathrm{u}}, \mathrm{r}(\mathscr{U}_{\mathrm{u}}), \lambda_{\mathrm{r}}|\mathscr{U}_{\mathrm{u}}|)$;
    $\mathscr{U}_{\mathrm{d}} \leftarrow \mathrm{maxpartition}(\mathscr{U}_{\mathrm{r}}, \mathrm{d}(\mathscr{U}_{\mathrm{r}}), \lambda_{\mathrm{d}}|\mathscr{U}_{\mathrm{r}}|)$;
    $\mathscr{U}_{\mathrm{c}} \leftarrow \mathrm{maxpartition}(\mathscr{U}_{\mathrm{d}}, \mathrm{c}(\mathscr{U}_{\mathrm{d}}), \lambda_{\mathrm{c}}|\mathscr{U}_{\mathrm{d}}|)$;
    acquire ground truth labels $\mathscr{Y}_{\mathrm{c}}$ for $\mathscr{U}_{\mathrm{c}}$;
    $\mathscr{D} = \mathscr{D} + (\mathscr{U}_{\mathrm{c}}, \mathscr{Y}_{\mathrm{c}})$;
    $\mathscr{U}_{\mathrm{i}} = \mathscr{U}_{\mathrm{i}} - \mathscr{U}_{\mathrm{c}}$;
  end
end

Algorithm 1 depicts the proposed active learning algorithm for a class-biased multi-label text emotion classification problem, which takes the original training data $\mathscr{D}$, an unlabeled incoming resource $\mathscr{U}$, and a set of selection ratio parameters in $\lambda_{\{\mathrm{u}|\mathrm{r}|\mathrm{d}|\mathrm{c}\}}$ as input, progressively selects the most informative, high-quality, and the most class-compensative examples into $\mathscr{U}_{\mathrm{c}}$, querying them for the true emotion labels $\mathscr{Y}_{\mathrm{c}}$, and updates the training data $\mathscr{D}$ with these new pairs $(\mathscr{U}_{\mathrm{c}}, \mathscr{Y}_{\mathrm{c}})$ as output. The sample selection procedure consists of two loops. The outer loop repeatedly selects batch-wise samples for m times, by learning a supervised classifier f from the current training data $\mathscr{D}$, picking an incoming batch $\mathscr{U}_{\mathrm{i}}$ from the unlabeled resource $\mathscr{U}$, and predicting the emotion labels $\hat{\mathscr{Y}}_{\mathrm{i}}$ for the incoming batch by f. The inner loop repeatedly selects a specified number $\mathscr{N}$ of examples for n times from the input resource $\mathscr{U}$ which have the correspondingly maximum scores in $\mathscr{C}$. The selection is taken by a maximum-partition function $\mathrm{maxpartition}(\mathscr{U}, \mathscr{C}, \mathscr{N})$, in which data $\mathscr{U}_{\mathrm{u}}$ can be first selected from the incoming data batch $\mathscr{U}_{\mathrm{i}}$ based on the uncertainty evaluation u and the selection ratio λ_{u}. Then data $\mathscr{U}_{\mathrm{r}}$, $\mathscr{U}_{\mathrm{d}}$, and $\mathscr{U}_{\mathrm{c}}$ can be consecutively selected based on the corresponding criterion and selection ratio, with all the desired properties. At the end of the inner loop, the selected examples of $\mathscr{U}_{\mathrm{c}}$ are appended to the training data $\mathscr{D}$ with the acquired ground truth labels $\mathscr{Y}_{\mathrm{c}}$ and are removed from the incoming batch $\mathscr{U}_{\mathrm{i}}$.

4 EXPERIMENT

Based on a large number of timeline messages retrieved from of Sina Weibo[2] for six months in 2013, we report our findings on the active learning for the class-biased multi-label emotion classification. An average number of 24,705 messages were retrieved for every hour. The data was divided into four sets, that is, a training set, a validation set, a test set, and an unlabeled set, as shown in Fig. 2.

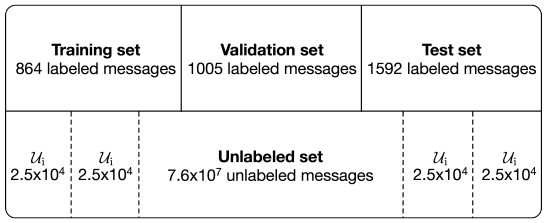


Fig. 2. Data segmentation of the timeline Weibo messages.

The training and validation sets consist of around 100 labels for each emotion category, while the test set consists of around 200 labels for each emotion category. The numbers of emotion labels are kept in this way because we want to learn a balanced multi-label emotion classifier f at the beginning of active learning and to evaluate multi-label emotion classification results evenly for all emotion categories. The unlabeled set consists of large amounts of messages and potentially highly biased emotion labels. We divide it into smaller batches, each of which corresponds to around 2.5×10^4 timeline messages retrieved per hour, and feed them as the incoming batches $\mathscr{U}_{\mathrm{i}}$ in the outer loop of active learning. All text messages are segmented into lists of words with a Chinese word segment package[3]. For some functional expressions in the social network messages, such as *@name*, *urls*, and *numbers*, we regularize them into their functional basics. Because the quality of timeline messages varies drastically, we also trained a spam filter

2. Sina Weibo (http://weibo.com/) is a Chinese microblogging weibsite.
3. THULAC(https://github.com/thunlp/THULAC-Python)

to remove the advertisements and none textual contents from the unlabeled set.

The values of hyper-parameters $\lambda_{\{\mathrm{u|r|d|c}\}}$ in Algorithm 1 are selected according to the performance of active learning on the validation set, with a total of $m = 6$ incoming batches from the unlabeled set. Because the unlabeled data is easy to retrieve, we keep the number of inner loops to $n = 1$, that is, the algorithm selects only one set of unlabeled examples $\mathscr{U}_\mathrm{c}$ for each incoming batch $\mathscr{U}_\mathrm{i}$. And we further put a constraint on the number of output examples $\lambda_\mathrm{c}|\mathscr{U}_\mathrm{d}|$ so that each loop updates the training data $\mathscr{D}$ with exactly the same number of new examples, although the results should nearly be the same since the incoming batch sizes are very similar. The selected values for λ_u, λ_r, λ_d, and $\lambda_\mathrm{c}|\mathscr{U}_\mathrm{d}|$ are 0.2, 0.5, 0.5, and 40, respectively.

For the evaluation of active learning on test set, we take $m = 40$ incoming batches from the unlabeled set and keep the same number of inner loops, that is, $n = 1$ as the above model selection procedure. For each incoming batch, the algorithm first selects the top 20% uncertain examples, then the top 50% representative examples from the selected data before. Next the top 50% diverse examples from the selected data before, and finally the top 40 complementary examples from the selected data before are selected. We compare the effectiveness of complementary sampling with respect to three probabilistic distance measurements, that is, KL for the Kullback-Leibler divergence, CE for the cross entropy distance, and EM for the earth mover's distance.

Two baseline active learning algorithms are also employed in which the first algorithm considers all criteria except complementariness for example sampling while the second algorithm considers a simple class complementation method for corpus construction by following the algorithm for a Japanese Twitter emotion classification problem [29]. Because the baseline two method only considered a KL divergence from the probabilistic model prediction to the label proportion in the training set, direct estimation of the class balancing effect by integrating a candidate sample to the existing corpus was impossible. Besides, zero and extremely small probability values from model prediction could cause computation errors for KL divergence, and minimizing this KL divergence for sample selection does not directly lead to the goal of class balance.

The evaluation metrics for multi-label emotion classification are briefly explained as follows. The Macro and Micro F_1 metrics evaluate the harmonic means of the precision and recall of emotion classification by

$$\mathrm{F}_1^{\{\mathrm{macro|micro}\}} = \frac{2 \times \mathrm{P}^{\{\mathrm{macro|micro}\}} \times \mathrm{R}^{\{\mathrm{macro|micro}\}}}{\mathrm{P}^{\{\mathrm{macro|micro}\}} + \mathrm{R}^{\{\mathrm{macro|micro}\}}}, \tag{20}$$

in which the Macro precision and recall evaluate the class-wise averaged ratio of true positive predictions to positive predictions and the ratio of true positive predictions to positive ground truth labels by

$$\mathrm{P}^{\mathrm{macro}} = \frac{1}{K}\sum_{k=1}^{K}\frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k}, \tag{21}$$

$$\mathrm{R}^{\mathrm{macro}} = \frac{1}{K}\sum_{k=1}^{K}\frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k}. \tag{22}$$

The Micro precision and recall evaluate the same ratios except that the predictions and ground truth labels are aggregated for all classes before calculation by

$$\mathrm{P}^{\mathrm{micro}} = \frac{\sum_{k=1}^{K}\mathrm{TP}_k}{\sum_{k=1}^{K}\mathrm{TP}_k + \mathrm{FP}_k}, \tag{23}$$

$$\mathrm{R}^{\mathrm{micro}} = \frac{\sum_{k=1}^{K}\mathrm{TP}_k}{\sum_{k=1}^{K}\mathrm{TP}_k + \mathrm{FN}_k}. \tag{24}$$

The Accuracy metric evaluates the ratio of an exact match of the predictions to the ground truth labels for all emotion classes by

$$\mathrm{Acc}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{M}\sum_{i}^{M} 1\{\mathbf{y}_i = \hat{\mathbf{y}}_i\}, \tag{25}$$

in which M denotes the test set size, and the indicator function $1\{\mathbf{y}_i = \hat{\mathbf{y}}_i\}$ renders 1 if and only if $\mathbf{y}_{ik} = \hat{\mathbf{y}}_{ik}$ is true for all $k \in [1, K]$. The Jaccard Similarity metric evaluates the ratio of the correct positive predictions to the union of the positive predictions and the ground truth labels, regardless of their emotion classes by

$$\mathrm{JSim}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\sum_{i=1}^{M}\sum_{k=1}^{K} 1\{\mathbf{y}_{ik} = 1 \wedge \hat{\mathbf{y}}_{ik} = 1\}}{\sum_{i=1}^{M}\sum_{k=1}^{K} 1\{\mathbf{y}_{ik} = 1 \vee \hat{\mathbf{y}}_{ik} = 1\}}. \tag{26}$$

The Hamming Loss metric evaluates the ratio of incorrect predictions, despite the reality of prediction values, to the total number of test set labels by

$$\mathrm{HLoss}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{M \times K}\sum_{i=1}^{M}\sum_{k=1}^{K} 1\{\mathbf{y}_{ik} \neq \hat{\mathbf{y}}_{ik}\}. \tag{27}$$

And the Log Loss metric evaluates the mean of cross-entropy loss for the probabilistic predictions $\mathbf{p}$, across all test samples and emotion classes by

$$\mathrm{LLoss}(\mathbf{y}, \mathbf{p}) = -\frac{1}{M \times K}\sum_{i=1}^{M}\sum_{k=1}^{K} \mathbf{y}_{ik}\log \mathbf{p}_{ik} + (1 - \mathbf{y}_{ik})\log(1 - \mathbf{p}_{ik}). \tag{28}$$

Fig. 3 shows the batch-wise improvement of emotion classification results, in which we compare the active learning algorithms of B1, B2, KL, CE, and EM, according to the above six evaluation metrics. The proposed complementary sampling with a KL probabilistic distance measurement has shown the most efficient learning curve for emotion classification in the experiment. For all active learning algorithms, the selected new samples have helped to train the emotion classification model with an overall increment in Macro F1, Micro F1, Accuracy, and Jaccard Similarity, and an overall decrement in Hamming Loss and Log Loss. Both the increment and the decrement for the KL-based active learning algorithm are more significant than the others, while the increments and decrements for the B1 and B2 algorithms indicate obvious fluctuations. These observations suggest that the proposed complementary sampling is effective and steady for improving the training data for emotion classification. In addition, the increments and decrements for the CE- and EM-based active learning algorithms are much slower than those of KL, which suggests that the KL probabilistic distance measurement is more appropriate for complementary sampling.

Fig. 4 shows the metric scores for the last five active learning loops in the box plots, which reveals the ceiling performance of active learning algorithms for improving emotion classification given a limited number of incoming batches. Specifically, each box indicates the minimum, the first quarter, the median, the third quarter, and the maximum of the scores for a metric. In this experiment we take $m = 40$ incoming batches from the unlabeled

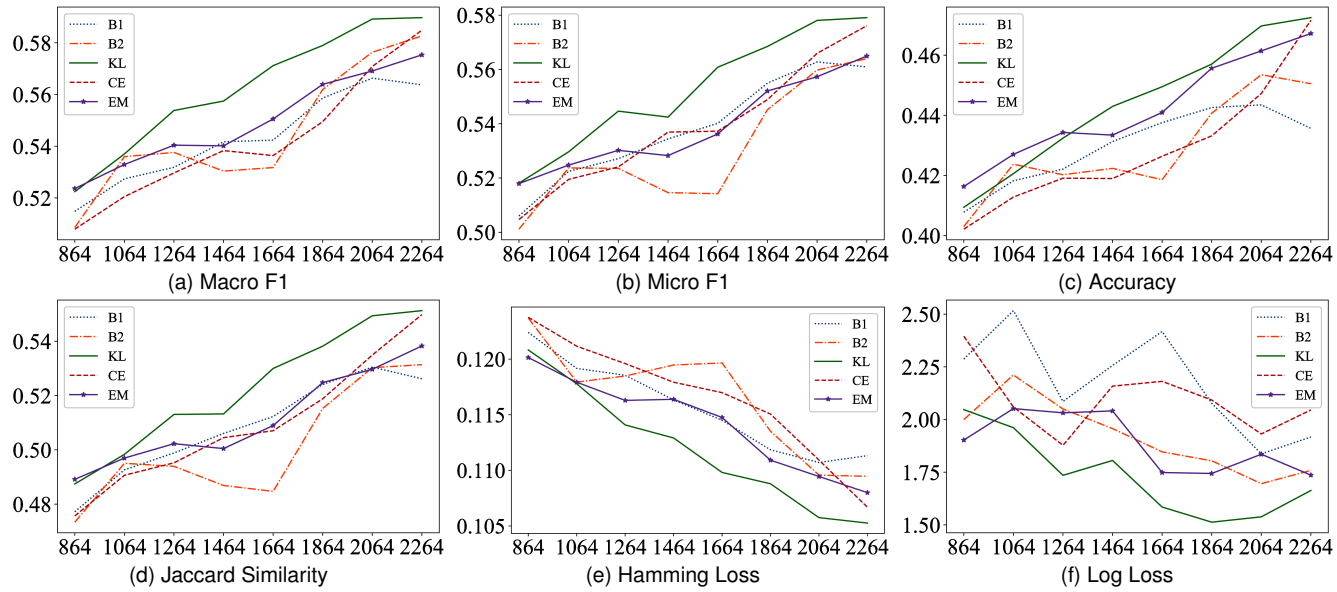


Fig. 3. Batch-wise evaluation of text emotion classification results.

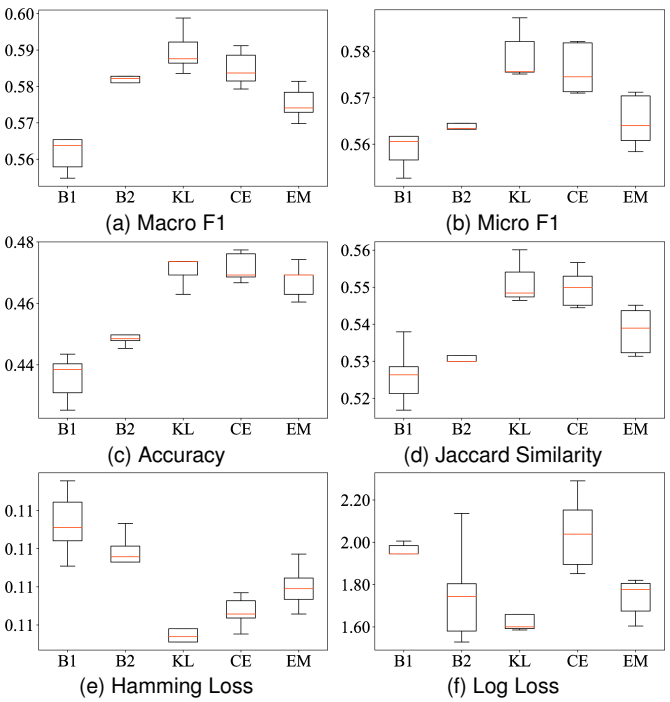
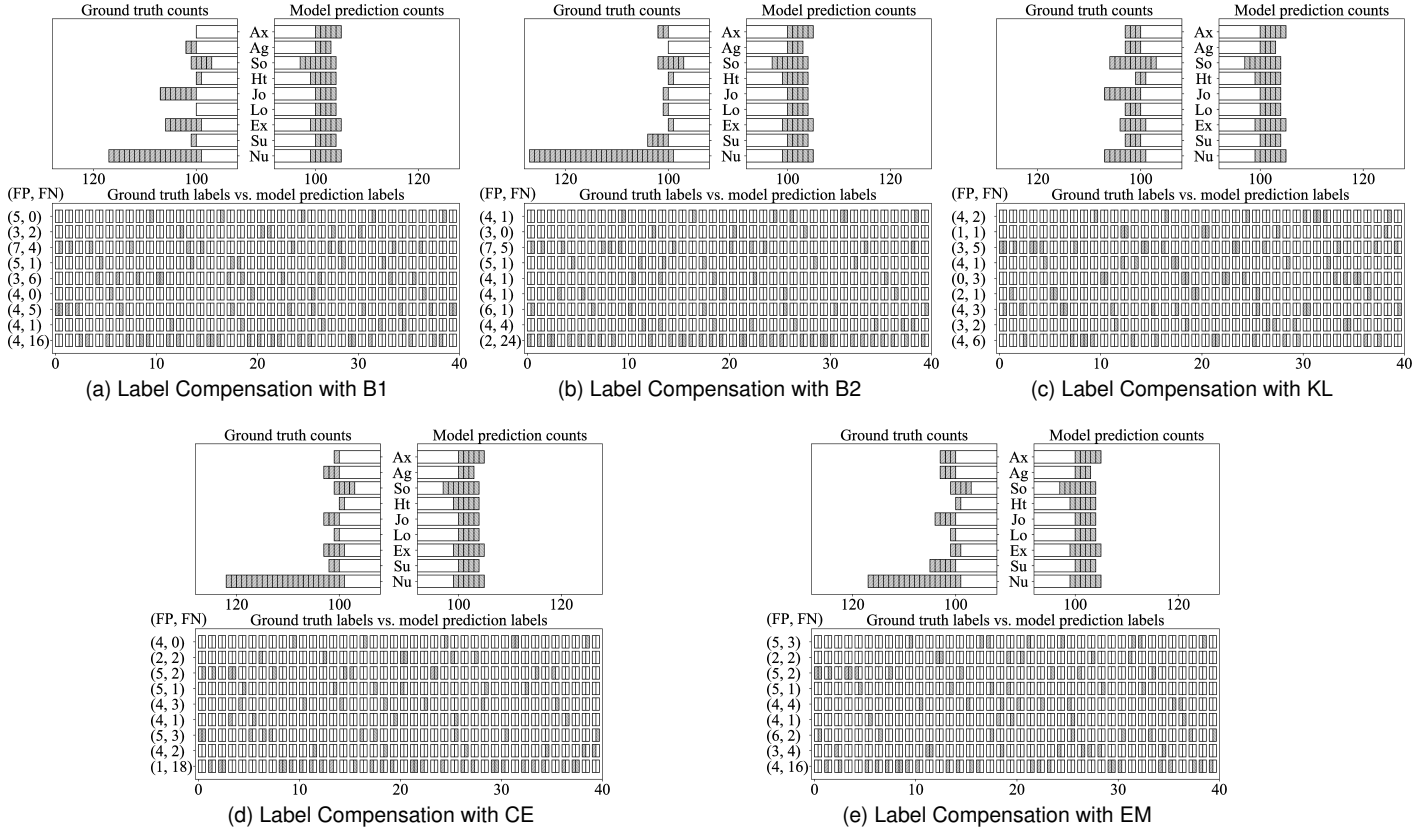


Fig. 4. Classification result evaluations for the last five active learning loops.

set, from which the 35th to 40th intermediately updated models are evaluated. The proposed KL-based active learning achieves the highest maximum scores for Macro F1, Micro F1, Jaccard Similarity, and the lowest minimum score for Hamming Loss. It is only slightly outperformed by the CE-based algorithm in Accuracy and by the B2 algorithm in Log Loss. In the meantime, the KL-based scores at the minimum, the first quarter, the median, and the third quarter of the boxes also indicate a generally better classification performance than the others. These observations suggest that the proposed complementary sampling with a KL probabilistic distance measurement could improve emotion classification to the best extent through active learning. In addition, the KL-, CE-, and EM-based algorithms render generally better scores in the boxes than B1 and B2, which again proves the effectiveness of the proposed method.

Next, we report a case study of the class compensation effect endowed by the active learning algorithm. Fig. 5 shows the results of compensating labels of a class-biased training data through active learning. The counts of which are indicated by the empty bars in the upper-left and upper-right subplots. Specifically, emotion labels of Sorrow (97), Hate (99), Expect (99), Neutral (99) are fewer than the other emotion labels (100). With different sampling strategies and separate probabilistic distance measurements, each active learning algorithm selects 40 examples from the same incoming batch of 25,000 unlabeled timeline messages. The predicted labels and the ground-truth labels for these selected examples are shown in the lower subplots, in which the predicted and the ground-truth labels are represented by pairs for each example. We use the slashed bars for the positive ground-truth labels, the backslashed bars for the positive predicted labels, and the empty bars for the negative labels for either ground truth or model prediction. The updated counts of the ground-truth labels are represented as the stacks of slashed and empty bars in the upper-left subplots, while the updated label counts under the model expectation are represented as the stacks of backslashed and empty bars in the upper-right subplots.

It is interesting to see that although each algorithm selects a different set of 40 examples, we get similar counts of expected emotion labels but very different counts of ground-truth emotion labels in the updated data. For instance, the B1 and B2 algorithms update the training data by stacking many Neutral labels. This in fact has pushed the class-imbalance problem even further. Because the unlabeled resource is extremely biased, keeping the count of different emotion labels the same for the training data has been difficult or even impossible through active learning. A better or more suitable target for complementary sampling would be to relieve the otherwise extremely biased class labels through active



(a) Label Compensation with B1 (b) Label Compensation with B2 (c) Label Compensation with KL

(d) Label Compensation with CE (e) Label Compensation with EM

Fig. 5. Label compensation for a class-biased training data through active learning.

learning. By comparing the upper-left subplots of Fig. 5b and 5c, we can observe a significant drop in the number of Neutral labels and sufficient increments in the numbers of the minority-class labels, such as Anxiety and Hate. This suggests that the KL-based active learning has effectively relieved the class-imbalance problem in the updated training data. Besides, we observe that the EM-based algorithm slightly outperforms the CE-based one for relieving the class-imbalance problem. However, none of them outperform the KL-based active learning algorithm.

We find that prediction mistakes in the selected examples could greatly affect the final label balance. To demonstrate this effect, we plotted the counts of false positive predictions and false negative predictions for the 40 selected examples in the $(\mathrm{FP}, \mathrm{FN})$ columns of the lower subplots in Fig. 5, for each active learning algorithm. These mistakes have pushed the true label distribution to deviate from the expected balanced distribution for the training data update. For example, a false positive Anxiety will reduce the count of true Anxiety labels by one in the training data, while a false negative Neutral will increase the count of true Neutral labels by one on the opposite side. To look further into this problem, we evaluated the false positive rates and false negative rates of the intermediately learned emotion classifiers for all active learning algorithms, with the results shown in Fig. 6.

In Fig. 6, each box aggregates the minimum, the first quarter, the median, the third quarter, and the maximum of the false positive or false negative rates for a binary emotion prediction over the $m = 40$ loops of active learning. We can observe that the false positive rates in Fig. 6c are much lower than those in Fig. 6a, 6b, 6d, and 6e, which indicates that the KL-based active learning has more accurate positive predictions than the baselines or the CE- and EM-based active learning algorithms. More importantly, the false positive rates in Fig. 6c for different emotion categories are closer to each other than those in the other subplots. This ensures that although the real emotion labels are reduced for these false positive mistakes, the number of reduced labels is around the same for these emotion categories, which helps relieving the class-bias problem once again.

In sub-figures of Fig. 6, the false negative rates for the Neutral prediction are significantly higher than those for the other emotion predictions. This implies that the active learning algorithms tend to select the examples with unrecognized Neutral labels, which causes the most significant increment of Neutral in all emotion categories. An emotion classifier which has been trained on such corpus would favor Neural in its predictions for the incoming unlabeled examples. Since the proportion of Neutral is extremely high for the Weibo timeline messages, and probably the same for the other Internet text resources, the false negative mistakes for Neutral are inevitable. This turns to be the major reason for all active learning algorithms to inevitably break the balance of emotion labels in our results.

Finally, we report the increment of emotion labels in result corpora for all active learning algorithms in Fig. 7. The KL-based active learning in Fig. 7c and the B2 algorithm in Fig. 7b have shown more balanced increments in the minority-class labels, through $m = 40$ loops, than those of the other active learning algorithms. Concurrently, both the KL-based algorithm and the B2 algorithm have shown better controls in the increment of the majority-class label, that is, Neutral than the other algorithms. We also find that the KL-based active learning has steadily outperformed the B2 algorithm, with a more restrained label increment in

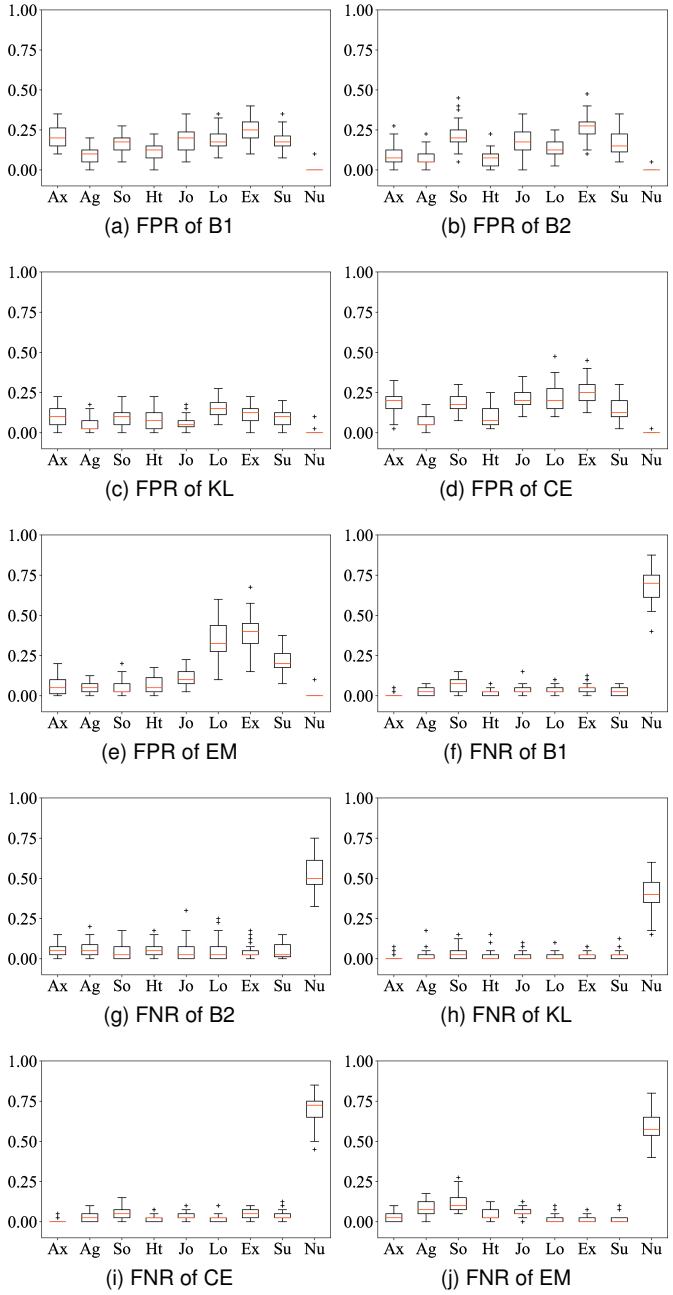


Fig. 6. False positive rates (FPR) and false negative rates (FNR) of the intermediately learned emotion classifiers.

Neutral. All these results suggest that through active learning, the KL-based complementary sampling strategy could retrieve new text examples from an unlabeled resource with the potentially best label balancing property.

5 Conclusion

Class bias in a multi-label emotion corpus has been a severe problem for the supervised emotion classification, in which a vast majority of the emotion labels are Neutral. In this paper, we propose a novel complementary sampling method for selecting training examples from a large unlabeled text resource for multi-label emotion classification with active learning. The proposed sampling method can find qualified examples to compensate a multi-label corpus of biased classes, by minimizing the probabilistic distance between the expected label distribution in a temporary corpus and a uniform label distribution. Through active learning, we also evaluate the candidate examples for model updating, outlier filtering, and diversity expanding with the uncertainty, representativeness, and diverseness criteria, respectively. Experiment results suggest that a multi-label emotion classifier could be improved along a more efficient curve and with a higher ceiling performance than those learned through the traditional active learning procedures. The minority-class counts of the training data significantly increased while the majority-class counts remain steadily controlled through all loops of active learning. With a thorough analysis of the intermediate classification results, we find that compared to false positive mistakes the false negative mistakes could contribute more to the failure of retaining label balance in training data. However, for the unlabeled resources with an extremely biased label distribution, it is still difficult or even impossible for the current active learning method to balance the class labels. Complementary sampling is a general sampling strategy for compensating corpus annotations, in which the label distribution can be extended to any reasonable dimensions. We hope to explore semantic features, label connections, and powerful neural networks, based on larger and more general data sets, to improve active learning for the class compensation problem in the future.

Acknowledgments

This research has been supported by JSPS KAKENHI Grant Number 19K20345 and Grant Number 19H04215.

References

[1] A. Yadollahi, A. G. Shahraki, and O. R. Zaiane, "Current state of text sentiment analysis from opinion to emotion mining," *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, p. 25, 2017.

[2] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102–107, 2016.

[3] S. M. Mohammad, "Sentiment analysis: Detecting valence, emotions, and other affectual states from text," in *Emotion measurement*. Elsevier, 2016, pp. 201–237.

[4] F. Ren and Y. Wu, "Predicting user-topic opinions in twitter with social and topical context," *IEEE Transactions on Affective Computing*, vol. 4, no. 4, pp. 412–424, 2013.

[5] H. Liu, H. Lieberman, and T. Selker, "A model of textual affect sensing using real-world knowledge," in *Proceedings of the 8th international conference on Intelligent user interfaces*. ACM, 2003, pp. 125–132.

[6] N. Colneriĉ and J. Demsar, "Emotion recognition on twitter: comparative study and training a unison model," *IEEE Transactions on Affective Computing*, 2018.

[7] X. Li, Y. Rao, H. Xie, R. Y. K. Lau, J. Yin, and F. L. Wang, "Bootstrapping social emotion classification with semantically rich hybrid neural networks," *IEEE Transactions on Affective Computing*, vol. 8, no. 4, pp. 428–442, 2017.

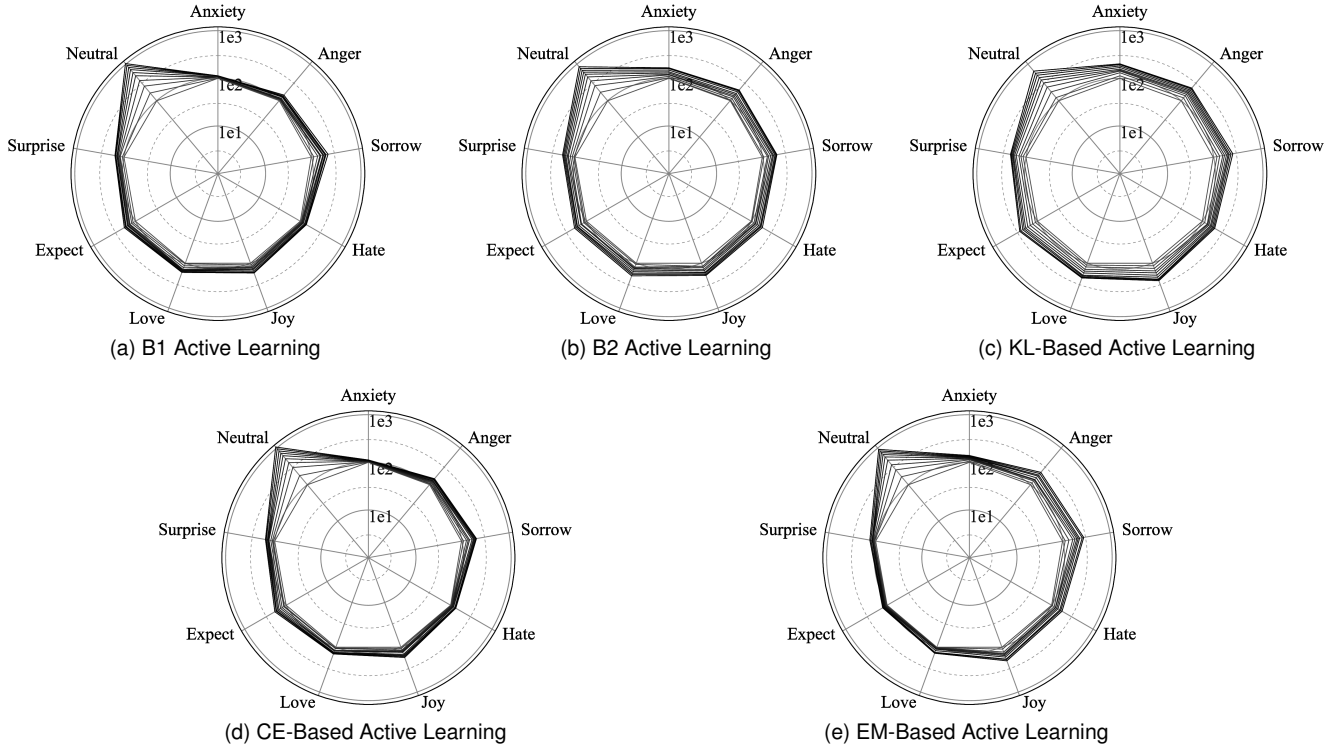
[8] X. Li, Q. Peng, Z. Sun, L. Chai, and Y. Wang, "Predicting social emotions from readers' perspective," *IEEE Transactions on Affective Computing*, 2017.

[9] U. Krcadinac, P. Pasquier, J. Jovanovic, and V. Devedzic, "Synesketch: An open source library for sentence-based emotion recognition," *IEEE Transactions on Affective Computing*, vol. 4, no. 3, pp. 312–325, 2013.

[10] P. Ekman, "Facial expression and emotion." *American psychologist*, vol. 48, no. 4, p. 384, 1993.

[11] C. Quan and F. Ren, "A blog emotion corpus for emotional expression analysis in chinese," *Computer Speech & Language*, vol. 24, no. 4, pp. 726–749, 2010.

[12] D.-A. Phan, Y. Matsumoto, and H. Shindo, "Autoencoder for semisupervised multiple emotion detection of conversation transcripts," *IEEE Transactions on Affective Computing*, 2018.



(a) B1 Active Learning (b) B2 Active Learning (c) KL-Based Active Learning

(d) CE-Based Active Learning (e) EM-Based Active Learning

Fig. 7. Increment of emotion labels in the training data through active learning.

[13] X. Kang, F. Ren, and Y. Wu, "Exploring latent semantic information for textual emotion recognition in blog articles," *IEEE/CAA Journal of Automatica Sinica*, vol. 5, no. 1, pp. 204–216, 2018.

[14] V. Kuperman, Z. Estes, M. Brysbaert, and A. B. Warriner, "Emotion and language: Valence and arousal affect word recognition." *Journal of Experimental Psychology: General*, vol. 143, no. 3, p. 1065, 2014.

[15] D. A. Havas, A. M. Glenberg, and M. Rinck, "Emotion simulation during language comprehension," *Psychonomic bulletin & review*, vol. 14, no. 3, pp. 436–441, 2007.

[16] C. A. Lutz and L. E. Abu-Lughod, "Language and the politics of emotion." in *This book grew out of a session at the 1987 annual meeting of the American Anthropological Association called" Emotion and Discourse."*. Editions de la Maison des Sciences de l'Homme, 1990.

[17] L.-C. Yu, J.-L. Wu, P.-C. Chang, and H.-S. Chu, "Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news," *Knowledge-Based Systems*, vol. 41, pp. 89–97, 2013.

[18] F. Ren and C. Quan, "Linguistic-based emotion analysis and recognition for measuring consumer satisfaction: an application of affective computing," *Information Technology and Management*, vol. 13, no. 4, pp. 321–332, 2012.

[19] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of computational science*, vol. 2, no. 1, pp. 1–8, 2011.

[20] M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini, "Political polarization on twitter," in *Fifth international AAAI conference on weblogs and social media*, 2011.

[21] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer, "Predicting the political alignment of twitter users," in *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*. IEEE, 2011, pp. 192–199.

[22] D. Gayo-Avello, ""i wanted to predict elections with twitter and all i got was this lousy paper"–a balanced survey on election prediction using twitter data," *arXiv preprint arXiv:1204.6441*, 2012.

[23] V. Lampos, D. Preoţiuc-Pietro, and T. Cohn, "A user-centric model of voting intention from social media," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2013, pp. 993–1003.

[24] L. Yang, D. Jiang, and H. Sahli, "Integrating deep and shallow models for multi-modal depression analysis—hybrid architectures," *IEEE Transactions on Affective Computing*, 2018.

[25] J. C. Eichstaedt, H. A. Schwartz, M. L. Kern, G. Park, D. R. Labarthe, R. M. Merchant, S. Jha, M. Agrawal, L. A. Dziurzynski, M. Sap *et al.*, "Psychological language on twitter predicts county-level heart disease mortality," *Psychological science*, vol. 26, no. 2, pp. 159–169, 2015.

[26] F. Ren, X. Kang, and C. Quan, "Examining accumulated emotional traits in suicide blogs with an emotion topic model," *IEEE Journal of Biomedical and Health Informatics*, vol. 13, no. 9, 2015.

[27] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, R. E. Lucas, M. Agrawal, G. J. Park, S. K. Lakshmikanth, S. Jha, M. E. Seligman *et al.*, "Characterizing geographic variation in well-being using tweets," in *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[28] L. Bylsma, B. Morris, and J. Rottenberg, "A meta-analysis of emotional reactivity in major depressive disorder," *Clinical Psychology Review*, vol. 28, no. 4, pp. 676–691, 2008.

[29] X. Kang, Y. Wu, and F. Ren, "Progressively improving supervised emotion classification through active learning," in *International Conference on Multi-disciplinary Trends in Artificial Intelligence*. Springer, 2018, pp. 49–57.

[30] X. Kang, F. Ren, and Y. Wu, "Semi-supervised learning of author-specific emotions in micro-blogs," *TEEE C (Electronics, Information and Systems)*, vol. 11, no. 6, 2016.

[31] F. Ren and K. Matsumoto, "Semi-automatic creation of youth slang corpus and its application to affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 176–189, 2016.

[32] F. Ren and X. Kang, "Employing hierarchical bayesian networks in simple and complex emotion topic analysis," *Computer Speech & Language*, vol. 27, no. 4, pp. 943–968, 2013.

[33] C. Katsimerou, J. Albeda, A. Huldtgren, I. Heynderickx, and J. A. Redi, "Crowdsourcing empathetic intelligence: the case of the annotation of emma database for emotion and mood recognition," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 7, no. 4, pp. 1–27, 2016.

[34] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 2013, pp. 1–8.

[35] E. A. Kolog, C. S. Montero, and E. Sutinen, "Annotation agreement of emotions in text: The influence of counsellors' emotional state on their

emotion perception," in *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)*. IEEE, 2016, pp. 357–359.

[36] S. Aman and S. Szpakowicz, "Identifying expressions of emotion in text," in *International Conference on Text, Speech and Dialogue*. Springer, 2007, pp. 196–205.

[37] E. Kim and R. Klinger, "Who feels what and why? annotation of a literature corpus with semantic roles of emotions," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 1345–1359.

[38] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word–emotion association lexicon," *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013.

[39] S. M. Mohammad, "# emotional tweets," in *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 2012, pp. 246–255.

[40] H. Schuff, J. Barnes, J. Mohme, S. Padó, and R. Klinger, "Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus," in *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2017, pp. 13–23.

[41] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," *arXiv preprint arXiv:1710.03957*, 2017.

[42] D. Preoţiuc-Pietro, H. A. Schwartz, G. Park, J. Eichstaedt, M. Kern, L. Ungar, and E. Shulman, "Modelling valence and arousal in facebook posts," in *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, 2016, pp. 9–15.

[43] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: machine learning for text-based emotion prediction," in *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, 2005, pp. 579–586.

[44] B. Settles, "Active learning literature survey," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 2009.

[45] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, 1994, pp. 3–12.

[46] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 287–294.

[47] B. Settles, M. Craven, and S. Ray, "Multiple-instance active learning," in *Advances in Neural Information Processing Systems 20*, 2007, pp. 1289–1296.

[48] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *Journal of Artificial Intelligence Research*, vol. 4, no. 1, pp. 129–145, 1996.

[49] N. Roy and A. McCallum, "Toward optimal active learning through sampling estimation of error reduction," in *ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning*, 2001, pp. 441–448.

[50] B. Settles and M. Craven, "An analysis of active learning strategies for sequence labeling tasks," in *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2008, pp. 1070–1079.

[51] J. Shao, Q. Wang, and Y. Lin, "Skyblocking for entity resolution," *Information Systems*, vol. 85, pp. 30–43, 2019.

[52] B. Kwolek, M. Koziarski, A. Bukała, Z. Antosz, B. Olborski, P. Wąsowicz, J. Swadźba, and B. Cyganek, "Breast cancer classification on histopathological images affected by data imbalance using active learning and deep convolutional neural network," in *International Conference on Artificial Neural Networks*. Springer, 2019, pp. 299–312.

[53] A. Sharma and R. Rani, "Be-dti': Ensemble framework for drug target interaction prediction using dimensionality reduction and active learning," *Computer methods and programs in biomedicine*, vol. 165, pp. 151–162, 2018.

[54] D. Himaja, T. M. Padmaja, and P. R. Krishna, "Oversample based large scale support vector machine for online class imbalance problem," in *International Conference on Big Data Analytics*. Springer, 2018, pp. 348–362.

[55] J. Błaszczyński and J. Stefanowski, "Actively balanced bagging for imbalanced data," in *International Symposium on Methodologies for Intelligent Systems*. Springer, 2017, pp. 271–281.

[56] M. Zięba and J. M. Tomczak, "Boosted svm with active learning strategy for imbalanced data," *Soft Computing*, vol. 19, no. 12, pp. 3357–3368, 2015.

[57] J. Zhang, X. Wu, and V. S. Shengs, "Active learning with imbalanced multiple noisy labeling," *IEEE transactions on cybernetics*, vol. 45, no. 5, pp. 1095–1107, 2014.

[58] X. Zhang, J. Cheng, C. Xu, H. Lu, and S. Ma, "Multi-view multi-label active learning for image classification," in *2009 IEEE International Conference on Multimedia and Expo*, 2009, pp. 258–261.

[59] O. Reyes, C. Morell, and S. Ventura, "Effective active learning strategy for multi-label learning," *Neurocomputing*, vol. 273, pp. 494–508, 2018.

[60] K. Brinker, "On active learning in multi-label classification," in *From Data and Information Analysis to Knowledge Engineering*. Springer, 2006, pp. 206–213.

[61] B. Yang, J.-T. Sun, T. Wang, and Z. Chen, "Effective multi-label active learning for text classification," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 917–926.

[62] C.-A. Brust, C. Käding, and J. Denzler, "Active learning for deep object detection," *arXiv preprint arXiv:1809.09875*, 2018.

[63] S. Khanchi, A. Vahdat, M. I. Heywood, and A. N. Zincir-Heywood, "On botnet detection with genetic programming under streaming data label budgets and class imbalance," *Swarm and evolutionary computation*, vol. 39, pp. 123–140, 2018.

[64] H. Zhang, W. Liu, J. Shan, and Q. Liu, "Online active learning paired ensemble for concept drift and class imbalance," *IEEE Access*, vol. 6, pp. 73 815–73 828, 2018.

[65] S. Khanchi, M. I. Heywood, and A. N. Zincir-Heywood, "Properties of a gp active learning framework for streaming data with class imbalance," in *Proceedings of the Genetic and Evolutionary Computation Conference*, 2017, pp. 945–952.

[66] S. Khanchi, M. I. Heywood, and N. Zincir-Heywood, "On the impact of class imbalance in gp streaming classification with label budgets," in *European Conference on Genetic Programming*. Springer, 2016, pp. 35–50.

[67] X. Zhu, J. Lafferty, and Z. Ghahramani, "Combining active learning and semi-supervised learning using gaussian fields and harmonic functions," in *In: (Proceedings) 20th International Conference on Machine Learning workshop. (2003)*, 2003.

[68] S. C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Batch mode active learning and its application to medical image classification," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 417–424.

[69] A. Fujii, T. Tokunaga, K. Inui, and H. Tanaka, "Selective sampling for example-based word sense disambiguation," *Computational Linguistics*, vol. 24, no. 4, pp. 573–597, 1998.

[70] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016.

[71] P. O. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," *ACM Computing Surveys*, vol. 49, no. 2, p. 31, 2016.

[72] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *intelligent data analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[73] S. Doyle, J. Monaco, M. Feldman, J. Tomaszewski, and A. Madabhushi, "An active learning based classification strategy for the minority class problem: application to histopathology annotation," *BMC bioinformatics*, vol. 12, no. 1, p. 424, 2011.

[74] S. Ertekin, J. Huang, L. Bottou, and L. Giles, "Learning on the border: active learning in imbalanced data classification," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM, 2007, pp. 127–136.

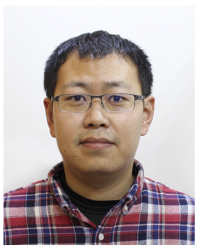
[75] T. M. Hospedales, S. Gong, and T. Xiang, "Finding rare classes: Active learning with generative and discriminative models," *IEEE transactions on knowledge and data engineering*, vol. 25, no. 2, pp. 374–386, 2013.

[76] S.-J. Huang, S. Chen, and Z.-H. Zhou, "Multi-label active learning: query type matters," in *IJCAI'15 Proceedings of the 24th International Conference on Artificial Intelligence*, 2015, pp. 946–952.

[77] S. Li, S. Ju, G. Zhou, and X. Li, "Active learning for imbalanced sentiment classification," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2012, pp. 139–148.

[78] X. Li and Y. Guo, "Active learning with multi-label svm classification," in *IJCAI '13 Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, 2013, pp. 1479–1485.

[79] H. T. Nguyen and A. Smeulders, "Active learning using pre-clustering," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 79.
[80] M. Ptaszynski, R. Rzepka, K. Araki, and Y. Momouchi, "Automatically annotating a five-billion-word corpus of japanese blogs for sentiment and affect analysis," *Computer Speech & Language*, vol. 28, no. 1, pp. 38–55, 2014.
[81] T. Reitmaier and B. Sick, "Let us know your decision: Pool-based active training of a generative classifier with the selection strategy 4ds," *Information Sciences*, vol. 230, pp. 106–131, 2013.
[82] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *Journal of machine learning research*, vol. 2, no. Nov, pp. 45–66, 2001.
[83] C. Riquelme, R. Johari, and B. Zhang, "Online active linear regression via thresholding," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
[84] S. Sabato and R. Munos, "Active regression by stratification," in *Advances in Neural Information Processing Systems*, 2014, pp. 469–477.
[85] R. Willett, R. Nowak, and R. M. Castro, "Faster rates in regression via active learning," in *Advances in Neural Information Processing Systems*, 2006, pp. 179–186.
[86] E. Lughofer and M. Pratama, "Online active learning in data stream regression using uncertainty sampling based on evolving generalized fuzzy models," *IEEE Transactions on fuzzy systems*, vol. 26, no. 1, pp. 292–309, 2018.
[87] D. Angluin, "Queries and concept learning," *Machine learning*, vol. 2, no. 4, pp. 319–342, 1988.
[88] R. Alur, R. Bodik, G. Juniwal, M. M. Martin, M. Raghothaman, S. A. Seshia, R. Singh, A. Solar-Lezama, E. Torlak, and A. Udupa, "Syntax-guided synthesis," in *2013 Formal Methods in Computer-Aided Design*. IEEE, 2013, pp. 1–8.
[89] R. Hwa, "Sample selection for statistical parsing," *Computational linguistics*, vol. 30, no. 3, pp. 253–276, 2004.
[90] M. Sharma and M. Bilgic, "Evidence-based uncertainty sampling for active learning," *Data Mining and Knowledge Discovery*, vol. 31, no. 1, pp. 164–202, 2017.
[91] P. Melville and R. J. Mooney, "Diverse ensembles for active learning," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 74.
[92] H. Wang, Y. Jin, and J. Doherty, "Committee-based active learning for surrogate-assisted particle swarm optimization of expensive problems," *IEEE transactions on cybernetics*, vol. 47, no. 9, pp. 2664–2677, 2017.
[93] W. Cai, Y. Zhang, and J. Zhou, "Maximizing expected model change for active learning in regression," in *2013 IEEE 13th International Conference on Data Mining*. IEEE, 2013, pp. 51–60.
[94] W. Cai, M. Zhang, and Y. Zhang, "Batch mode active learning for regression with expected model change," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 7, pp. 1668–1681, 2017.
[95] T. Zhang and F. Oles, "The value of unlabeled data for classification problems," in *Proceedings of the Seventeenth International Conference on Machine Learning,(Langley, P., ed.)*, vol. 20, no. 0. Citeseer, 2000, p. 0.
[96] Z. Xu, R. Akella, and Y. Zhang, "Incorporating diversity and density in active learning for relevance feedback," in *European Conference on Information Retrieval*. Springer, 2007, pp. 246–257.
[97] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 8, pp. 1819–1837, 2014.
[98] O. Reyes and S. Ventura, "Evolutionary strategy to perform batch-mode active learning on multi-label data," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 9, no. 4, p. 46, 2018.
[99] B. Grünbaum *et al.*, "Partitions of mass-distributions and of convex bodies by hyperplanes." *Pacific Journal of Mathematics*, vol. 10, no. 4, pp. 1257–1261, 1960.
[100] J. B. Walther, "Interpersonal effects in computer-mediated interaction: A relational perspective," *Communication research*, vol. 19, no. 1, pp. 52–90, 1992.
[101] J. B. Walther, T. Loh, and L. Granka, "Let me count the ways: The interchange of verbal and nonverbal cues in computer-mediated and face-to-face affinity," *Journal of language and social psychology*, vol. 24, no. 1, pp. 36–65, 2005.
[102] J. T. Hancock, C. Landrigan, and C. Silver, "Expressing emotion in text-based communication," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2007, pp. 929–932.
[103] E. Fox, *Emotion science cognitive and neuroscientific approaches to understanding human emotions*. Palgrave Macmillan, 2008.
[104] S. V. Toller, "Emotion: Theory, research, and experience. volume 3: Biological foundations of emotion : R. plutchik and h. kellerman (eds.), (academic press, new york, 1986) pp. xxiv + 423, Â£40.50, Â£25.00 (paperback)." *Biological Psychology*, vol. 23, no. 3, pp. 320–322, 1986.
[105] P. Shaver, J. Schwartz, D. Kirson, and C. O'connor, "Emotion knowledge: further exploration of a prototype approach." *Journal of personality and social psychology*, vol. 52, no. 6, p. 1061, 1987.
[106] W. Shi, H. Wang, and S. He, "Sentiment analysis of chinese microblogging based on sentiment ontology: a case study of '7.23 wenzhou train collision'," *Connection Science*, vol. 25, no. 4, pp. 161–178, 2013.
[107] B. Gunter, N. Koteyko, and D. Atanasova, "Sentiment analysis: A market-relevant and reliable measure of public feeling?" *International Journal of Market Research*, vol. 56, no. 2, pp. 231–247, 2014.
[108] X. Kang, F. Ren, and Y. Wu, "Bottom up: Exploring word emotions for chinese sentence chief sentiment classification," in *2010 International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*. IEEE, 2010, pp. 1–5.
[109] Y. Wang and A. Pal, "Detecting emotions in social media: A constrained optimization approach," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.
[110] A. Bandhakavi, N. Wiratunga, D. Padmanabhan, and S. Massie, "Lexicon based feature extraction for emotion text classification," *Pattern recognition letters*, vol. 93, pp. 133–142, 2017.
[111] C. Yang, K. Lin, and H. Chen, "Building emotion lexicon from weblog corpora," in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics, 2007, pp. 133–136.
[112] R. Tokuhisa, K. Inui, and Y. Matsumoto, "Emotion classification using massive examples extracted from the web," in *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2008, pp. 881–888.
[113] I. Maks and P. Vossen, "A verb lexicon model for deep sentiment analysis and opinion mining applications," *ACL HLT 2011*, p. 10, 2011.
[114] S. Mohammad and T. Yang, "Tracking sentiment in mail: How genders differ on emotional axes," *ACL HLT 2011*, p. 70, 2011.
[115] M. Li, Q. Lu, Y. Long, and L. Gui, "Inferring affective meanings of words from word embedding," *IEEE Transactions on Affective Computing*, vol. 8, no. 4, pp. 443–456, 2017.
[116] Y. Wu, K. Kita, F. Ren, K. Matsumoto, and X. Kang, "Modification relations based emotional keywords annotation using conditional random fields," in *2011 Fourth International Conference on Intelligent Networks and Intelligent Systems*. IEEE, 2011, pp. 81–84.
[117] ——, "Exploring emotional words for chinese document chief emotion analysis," in *25th Pacific Asia Conference on Language*, pp. 597–606.
[118] D. Das and S. Bandyopadhyay, "Word to sentence level emotion tagging for bengali blogs," in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Association for Computational Linguistics, 2009, pp. 149–152.
[119] Y. Rao, "Contextual sentiment topic model for adaptive social emotion classification," *IEEE Intelligent Systems*, vol. 31, no. 1, pp. 41–47, 2016.
[120] Y. Rao, H. Xie, J. Li, F. Jin, F. L. Wang, and Q. Li, "Social emotion classification of short text via topic-level maximum entropy model," *Information & Management*, vol. 53, no. 8, pp. 978–986, 2016.
[121] A. Summa, B. Resch, and M. Strube, "Microblog emotion classification by computing similarity in text, time, and space," in *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, 2016, pp. 153–162.



Xin Kang received his Ph.D degree from Tokushima University, Tokushima, Japan, in 2013, his M.E. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2009, and his B.E. degree from Northeastern University, Shenyang, China, in 2006. He is currently an assistant professor in Tokushima University. His research interests include machine learning, text emotion prediction, and natural language generation. Faculty of Engineering, Tokushima University, 2-1, Minamijyousanjima-cho, Tokushima 770-8506 Japan

Xuefeng Shi is pursuing Ph.D. degree in Hefei University of Technology. His research interests include active learning and emotion classification. School of Computer and Information, Hefei University of Technology, Danxia Road, Hefei, Anhui 230601 China

Yunong Wu received his Master degree and Ph.D degree from the Tokushima University, Tokushima, Japan, in 2011 and 2014 respectively. He worked as an assistant professor in Tokushima University before 2019 and became a senior researcher in Chengdu Senton Netease Co., Ltd. since then. His research interests include statistical machine learning, affective information computing, neural networks, and probabilistic graphical model. Chengdu Senton Netease Co., Ltd. Chengdu, Sichuan 610000 China

Fuji Ren was born in 1959, in China. He received his B.E. and M.E. degrees from Beijing University of Posts and Telecommunications, Beijing, China, in 1982 and 1985, respectively. He received his Ph.D. degree in 1991 from Hokkaido University, Japan. From 1991, he worked at CSK, Japan, where he was a chief researcher of NLP. From 1994 to 2000, he was an associate professor in the Faculty of Information Sciences, Hiroshima City University. He became a professor in the Faculty of Engineering of the University of Tokushima in 2001. His research interests include natural language processing, artificial intelligence, language understanding and communication, and affective computing. He is a member of IEICE, CAAI, IEEJ, IPSJ, JSAI, and AAMT, and a senior member of IEEE. He is a fellow of the Japan Federation of Engineering Societies. He is the president of the International Advanced Information Institute. Faculty of Engineering, University of Tokushima, 2-1, Minamijyousanjima-cho, Tokushima 770-8506 Japan