An Effective Deployment of Contrastive Learning in Multi-label Text Classification

Nankai Lin[1], Guanqiu Qin[1], Jigang Wang[1], Dong Zhou[2] * and Aimin Yang[1,3] *
[1] School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, Guangdong, 510006, China
[2] School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, Guangdong, 510006, China
[3] School of Computer Science and Intelligence Education, Lingnan Normal University, Zhanjiang 524000, Guangdong, China

Abstract

The effectiveness of contrastive learning technology in natural language processing tasks is yet to be explored and analyzed. How to construct positive and negative samples correctly and reasonably is the core challenge of contrastive learning. It is even harder to discover contrastive objects in multi-label text classification tasks. There are very few contrastive losses proposed previously. In this paper, we investigate the problem from a different angle by proposing five novel contrastive losses for multi-label text classification tasks. These are Strict Contrastive Loss (SCL), Intra-label Contrastive Loss (ICL), Jaccard Similarity Contrastive Loss (JSCL), Jaccard Similarity Probability Contrastive Loss (JSPCL), and Stepwise Label Contrastive Loss (SLCL). We explore the effectiveness of contrastive learning for multi-label text classification tasks by the employment of these novel losses and provide a set of baseline models for deploying contrastive learning techniques on specific tasks. We further perform an interpretable analysis of our approach to show how different components of contrastive learning losses play their roles. The experimental results show that our proposed contrastive losses can bring improvement to multi-label text classification tasks. Our work also explores how contrastive learning should be adapted for multi-label text classification tasks.

1 Introduction

Multi-label text classification is an important branch of text classification technology (Chalkidis and Søgaard, 2022; Zhang et al., 2022b). Different from binary classification tasks or multi-class classification tasks, multi-label classification tasks need to assign at least one label to a piece of text. Since the number of labels the text belongs to is not fixed, it greatly increases the difficulty of the model prediction. Specifically, the uncertainty in the number of labels poses two challenges to the training of multi-label text classification models: the output logic of the model and the semantic representation space of the model. In recent years, most multi-label text classification research has focused on designing better output logic to solve the uncertainty of the number of labels, such as transforming the multi-label text classification problem into a multi-task problem (Lin et al., 2022). However, for another challenge, how to construct a better semantic representation space for multi-label text classification models, little research attention has been paid.

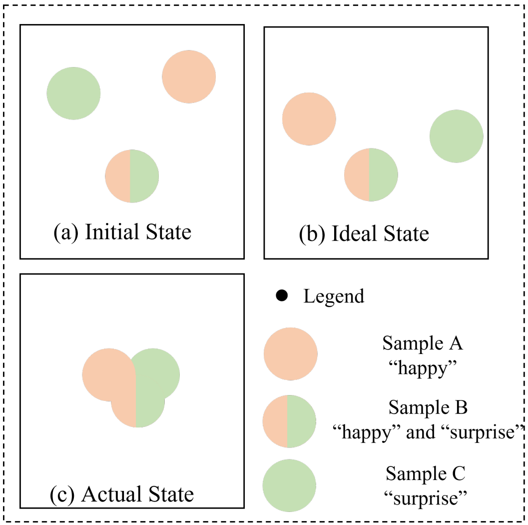

Figure 1: Example of Multi-label classification task.

The existence of multi-label samples can easily confound the semantic representation space, thereby posing a challenge in data analysis and modeling. When confronted with multi-label samples, the semantic representation space becomes susceptible to distractions, where the boundaries between different classes become blurred. This blurring effect stems from the inherent ambiguity

*Corresponding Author. E-mail: dongzhou@gdufs.edu.cn, amyang@gdut.edu.cn.

Findings of the Association for Computational Linguistics: ACL 2023, pages 8730–8744
July 9-14, 2023

that arises when multiple labels coexist within a single sample, causing uncertainty in the multi-label classification tasks. Take the multi-label emotion classification task as an example (shown in Figure 1), in which the "happy" sample (assumed to be sample A) shares a label with the "happy, surprise" sample (assumed to be sample B), and at the same time, the "surprise" sample (assumed to be sample C) also shares a label with the sample B (shown in Figure 1 (a)). Therefore, in the ideal state, the multi-label classification model assumes that sample A and sample B are located in a similar semantic space, and that sample B and sample C are located in another similar semantic space (shown in Figure 1 (b)). Figure 1 (c) shows how the samples of the "happy" category and the samples of the "surprise" category are confounded in the semantic space. This will cause sample A and sample C to be brought closer indirectly, even if their labels are completely different. As far as we know, the semantic representation of multi-label samples is still an open issue in the multi-label text classification task. Therefore, this paper focuses on using contrastive learning to improve the semantic representation of multi-label text classification models.

As an emerging technology, contrastive learning has achieved good performance in various fields of natural language processing (Khosla et al., 2020; Gao et al., 2021). How to construct positive and negative samples correctly and reasonably is the core challenge of contrastive learning. In multi-label text classification tasks, it is a great challenge to incorporate the contrastive learning module. It is more difficult for contrastive learning to perform well in multi-label text classification tasks than in other text classification tasks because implicit information representation of multi-label text is richer in the semantic space, which makes it more difficult to define positive and negative samples. Existing studies have proposed unsupervised contrastive learning methods to improve the performance of the model on multi-label text classification tasks (Khosla et al., 2020), and there are also working to improve supervised contrastive learning (Gao et al., 2021). However, the exploration of contrastive learning in multi-label text classification tasks is still very limited.

As the typical task in multi-label text classification, multi-label emotion classification task (Li et al., 2022; Ju et al., 2020; Ameer et al., 2023) and multi-label news classification task (Wang et al., 2021) have received extensive attention. In this paper, we propose five contrastive losses for multi-label text classification tasks and verify the performance of our method with the multi-label emotion classification task and multi-label news classification task as the representative tasks. More specifically, they are Strict Contrastive Loss (SCL), Intra-label Contrastive Loss (ICL), Jaccard Similarity Contrastive Loss (JSCL), Jaccard Similarity Probability Contrastive Loss (JSPCL), and Stepwise Label Contrastive Loss (SLCL). These five different strategies define the positive samples and negative samples of contrastive learning from different perspectives to pull the distance among different types of samples into the semantic space. To compare the effects of the five strategies, we further conduct an interpretable analysis to investigate how the different contrastive learning methods play their roles. The experimental results show that our proposed contrastive losses can bring improvement for multi-label text classification tasks. In addition, our methods could be considered as a set of baseline models of viable contrastive learning techniques for multi-label text classification tasks. This series of contrastive learning methods are plug-and-play losses, which can be applied to any multi-label text classification model, and to a certain extent, bring effective improvements to the multi-label text classification model.

The major contributions of this paper can be summarized as follows:

(1) For multi-label text classification tasks, we propose five novel contrastive losses from different perspectives, which could be regarded as a set of baseline models of contrastive learning techniques on multi-label text classification tasks.

(2) To the best of our knowledge, this is the first work that proposes a series of contrastive learning baselines for multi-label text classification tasks. At the same time, we also explore in detail the impact of different contrastive learning settings on multi-label text classification tasks.

(3) Through interpretable analysis, we further show the effectiveness of different contrastive learning strategies in transforming the semantic representation space.

2 Related Work

2.1 Multi-label Text Classification

In the field of text classification, multi-label text classification (MLTC) is always a challenging problem (Lin et al., 2022). A sample of multi-label text classification consists of a text and a set of labels. There is a correlation among labels. For this, some research transforms the multi-label classification problem into the seq2seq problem and learns the potential correlation among labels with the sequence generation model (Nam et al., 2017; Yang et al., 2018; Xiao et al., 2021). Yang et al. (2019) proposed a reinforcement learning-based seq2set framework, which can capture the correlation among tags and reduce the dependence on tag order. In addition, there is some research introducing label embedding so that the model can simultaneously learn the feature information of text and the co-occurrence information of labels. Ma et al. (2021) proposed to learn statistical label co-occurrence via GCN. LELC (Joint Learning from Label Embedding and Label Correlation) simultaneously learned labels attention and label co-occurrence matrix information (Liu et al., 2021). Zhang et al. (2021) ensembled the MLTC and the label co-occurrence task to enhance label correlation feedback.

Most dataset of MLTC has the data distribution imbalance problem: imbalance within labels, among labels, and among label-sets. The studies we have discussed above, which use label embedding, have alleviated the impact of label imbalance to some extent while learning label association. Some research solves the problem of data imbalance by resampling. For example, based on the edited nearest neighbor rule, Charte et al. (2014) proposed a multi-label undersampling algorithm. They defined a measure of the differential distance between label sets in order to heuristically remove unnecessary samples during resampling. Considering the problem in terms of object functions, Ridnik et al. (2021) proposed an asymmetric loss that dynamically adjusts the asymmetry levels to balance the effect of positive and negative samples in training.

2.2 Multi-label Emotion Classification

Sentiment analysis (Xu et al., 2016) is of great significance to society, economy and security. In early studies sentiment analysis (Mohammad and Turney, 2013; Turney, 2002) is implemented based on the sentiment polarity dictionary. These methods utilize unsupervised methods such as point mutual information (PMI) to construct an emotional dictionary based on the basic emotional word set, and then calculate the emotional weight value and emotional polarity of the text according to the viewpoint words with different intensity of positive, neutral and negative emotional tendencies in the dictionary. While some studies (Socher et al., 2013; Nakov et al., 2013) transform sentiment analysis into binary or mutil-classification problems, which leads to many subsequent supervised learning studies based on machine learning and neural networks.

In recent years, more and more scholars (Shmueli et al., 2021; Mohammad et al., 2018) regarded the sentiment analysis task as a multi-label problem, and accordingly, Yilmaz et al. (2021) introduced it into multi-label sentiment analysis by adapting the focal loss and proposed a dynamic weighting method to balance each label's contribution in the training set. Alhuzali and Ananiadou (2021) transformed the problem of multi-label sentiment classification into span-prediction by means of prompt learning, and proposed a label relationship perception loss. They converted labels into tokens and inputted them into BERT together with the original input text, and used the attention module of the Transformer and the knowledge learned in the pre-train stage to learn the correlation of emotional labels. In addition to encoding labels and sentences with BERT at the same time, EduEmo (Zhu and Wu, 2022) also introduced the encoder of Realformer (He et al., 2021) to model the association between each elementary discourse unit and sentiment labels.

2.3 Contrastive Learning

In recent years, contrastive learning has gradually become one of the important techniques in natural language processing and computer vision. In the field of natural language processing, contrastive learning is usually used to improve the quality of embedding representation by comparing feature vectors, bringing semantically similar and same label embeddings closer, and distancing semantically dissimilar and different label embeddings.

Contrastive learning could be divided into supervised contrastive learning and unsupervised contrastive learning. Khosla et al. (2020) proposed a supervised contrastive learning method, which took the original label of the sample as the anchor, and made the clusters of the same label closer to each other, and the clusters of different labels far away from each other in the embedding space. To improve the sentence-level representation, SimCSE used dropout technology for unsupervised contrastive learning and natural language inference dataset for supervised contrastive learning (Gao et al., 2021). Some research introduced supervised contrastive learning into the pre-training process of PLMs, and experiments result on their downstream tasks showed that the performance of pre-trained models was generally improved (Gunel et al., 2020; Qin et al., 2021).

2.4 Contrastive Learning for Multi-label Text Classification

At present, the application of contrastive learning in multi-label classification mainly focuses on image-related tasks. MulCon, an end-to-end framework for multilabel image classification, used image label-level embeddings with a multi-head attention mechanism to transform the multi-label classification problem into the binary classification problem for each label-level embedding (Dao et al., 2021). Małkiński and Mańdziuk (2022) proposed a supervised multi-label contrastive learning method for abstract visual reasoning. They reconstructed the contrastive loss function according to the multi-label problem, allowing sample pairs to contrast all labels. Zhang et al. (2022a) proposed a general hierarchical multi-label representation learning framework, which introduced hierarchical loss retention and hierarchical constraints.

However, different from the representation space of images, the implicit information representation of text is richer, which makes it more difficult to define positive samples and negative samples, and it is more difficult for contrastive learning to show good performance. Research of contrastive learning in multi-label text classification is focusing on unsupervised multi-label contrastive learning (Zhou et al., 2022). What's more, Su et al. (2022) attempted to improve supervised contrastive learning by using the knowledge of existing multi-label instances for supervised contrastive learning. Bai et al. (2022) proposed to take the sample features as anchor samples, and take the corresponding positive labels and negative labels as positive and negative samples for supervised contrastive learning. However, the exploration of contrastive learning in multi-label text analysis tasks is still very limited.

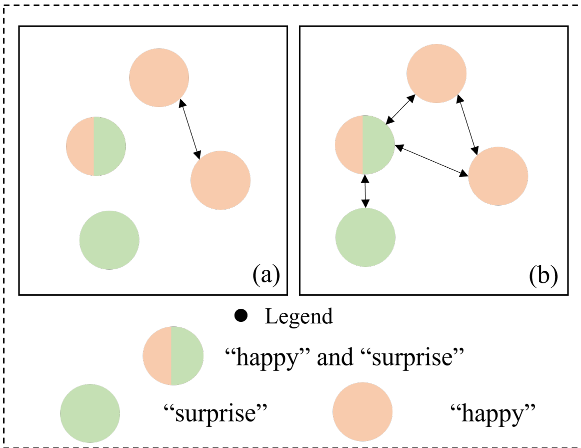

Figure 2: Example of different standards.

3 Contrastive Loss for Multi-Label Text Classification

In this section, we describe in detail the application of our proposed different contrastive learning methods on multi-label text classification tasks. We take the multi-label emotion classification task as an example to describe our method. It is worth noting that our proposed method can not only be applied to multi-label emotion classification tasks, but also can be applied to other multi-label text classification tasks.

Suppose a minibatch which contains K samples $D = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_K, Y_K)\}$ and $I = \{1, \ldots, K\}$ is the sample index set. Given a sample index i, X_i is the text sequence of samples i and its label set is denoted as Y_i. After encoding by the multi-label text classification model M, we could obtain the sentence representation vector $\boldsymbol{e_i^t}$ and the emotion representation matrix E_i^e of X_i, where $E_i^e = \{\boldsymbol{e_{i1}}, \boldsymbol{e_{i2}}, \ldots, \boldsymbol{e_{il}}\}$ and l represents the total number of emotion labels. It is worth noting that the model M here can be any deep learning multi-label language model. Y_i is the one-hot encoding of the label, i.e. $Y_i = \{y_1, y_2, ..., y_l\}$. For a given i-th emotion $y_i \in \{0, 1\}$, y_i=0 means that this type of emotion does not exist in the text, and y_i=1 means that this type of emotion exists in the text. We further define the label prediction probability distribution of the model M output as p_i. Contrastive learning aims to change the semantic representation space of the model. Since the multi-label classification tasks are more complex than the single-label classification tasks, the main exploration of our paper is how one can construct positive and negative samples for contrastive learning.

When contrastive learning is applied to multi-label text classification, for an anchor, the definition of its positive samples can be diversified. For example, when a strict standard is implemented, positive samples are defined as samples with exactly the same label set (shown in figure 2 (a)), when a loose standard is implemented, positive samples are defined as samples with partly the same label set (shown in figure 2 (b)). For different positive samples, contrastive learning pulls different samples closer in semantic space for a given anchor. In the strict standard, we could find that for an anchor point, there are fewer positive samples, and samples containing some similar features cannot be pulled closer. In the loose standard, there are more positive samples for an anchor point, which may indirectly bring samples of different labels closer. Therefore, different positive and negative sample construction methods affect the optimization goal of the model. What's more, there are two different types of contrastive learning, Feature-based Contrastive Learning (FeaCL) (Fu et al., 2022) and Probability-based Contrastive Learning (ProCL) (Li et al., 2021). FeaCL uses semantic representations of sentences as the basic component to build the contrastive objective function. ProCL constructs the contrastive objective function from the perspective of probability distributions instead of semantic representations. Using different features for contrastive learning will also affect the optimization of the model. In order to explore how contrastive learning can be better applied to multi-label text classification tasks, we introduce five different contrastive learning methods SCL, ICL, JSCL, JSPCL, and SLCL, as below.

3.1 Strictly Contrastive Loss

As a strict standard method, SCL requires that only when the label set of the sample is exactly the same as the label set of the anchor point can it be used as a positive contrastive sample of the anchor point. Therefore, SCL does not consider samples that partially overlap with the anchor label set. In addition, SCL is also a method of FeaCL type, which uses the semantic representation of samples obtained from model encoding as the contrastive feature. In the SCL, for a given sample i, all other samples that share the same label set with it in the batch form the set $S = \{s : s \in I, Y_s = Y_i \wedge s \neq i\}$. Then we could define the SCL function for each entry i across the batch as

$$L_{SCL} = -\frac{1}{|S|}\sum_{s\in S}\log\frac{\exp(\frac{sim(\boldsymbol{e}_i^t,\boldsymbol{e}_s^t)}{\tau})}{\sum_{k\in I\setminus\{i\}}\exp(\frac{sim(\boldsymbol{e}_i^t,\boldsymbol{e}_k^t)}{\tau})} \tag{1}$$

where $sim(\cdot)$ indicates the cosine similarity function.

3.2 Jaccard Similarity Contrastive Loss

SCL is a strict contrastive learning method, which only pulls the samples with the exact same label closer, while JSCL operates on the samples to different degrees according to the similarity of the labels of the samples. We use Jaccard coefficient (Jaccard, 1912) to calculate the label similarity between samples. Similar to SCL, JSCL uses the semantic representation of samples obtained from model encoding as the contrastive feature. For a given sample, JSCL will zoom in as close as possible on samples with the exact same label while only slightly zooming in on samples that have some of the same labels. In the JSCL, for a given sample i, we could define the JSCL function across the batch as

$$L_{JSCL} = -\frac{1}{|I|}\sum_{s\in I}\log\frac{\frac{|Y_i\cap Y_s|}{|Y_i\cup Y_s|}\cdot\exp(\frac{sim(\boldsymbol{e}_i^t,\boldsymbol{e}_s^t)}{\tau})}{\sum_{k\in I\setminus\{i\}}\exp(\frac{sim(\boldsymbol{e}_i^t,\boldsymbol{e}_k^t)}{\tau})} \tag{2}$$

3.3 Jaccard Similarity Probability Contrastive Loss

Li et al. (2021) suggested that ProCL can produce more compact features than feature contrastive learning, while forcing the output probabilities to be distributed around class weights. Based on JSCL, we try to use probability for contrastive learning. In the JSPCL, for a given sample i, we could define the JSPCL function across the batch as

$$L_{JSPCL} = -\frac{1}{|I|}\sum_{s\in I}\log\frac{\frac{|Y_i\cap Y_s|}{|Y_i\cup Y_s|}\cdot\exp(\frac{sim(p_i,p_s)}{\tau})}{\sum_{k\in I\setminus\{i\}}\exp(\frac{sim(p_i,p_k)}{\tau})} \tag{3}$$

3.4 Stepwise Label Contrastive Loss

SLCL is another way to consider contrastive learning among samples with labels that are not exactly the same. The previous three contrastive learning methods mainly consider the situation when multiple emotions are considered at the same time, while SLCL considers different emotions separately, calculates the contrast loss separately, and then combines the losses of each emotion. In the JSPCL, for a given sample i, all other samples that share the same label y_j with it in the batch form the positive sample set S_j. The set of positive samples under each emotion label is $S = \{S_1, S_2, ..., S_q\}$ and q is the emotions' number of sample i. Then we could define the SLCL function for each entry i is across the batch as

$$L_{SLCL} = -\frac{1}{q}\sum_{S_j \in S}\frac{1}{|S_j|}\sum_{s \in S_j}\log\frac{\exp(\frac{sim(e_i^t, e_s^t)}{\tau})}{\sum_{k \in I\setminus\{i\}}\exp(\frac{sim(e_i^t, e_k^t)}{\tau})} \tag{4}$$

3.5 Intra-label Contrastive Loss

Different from several other contrastive losses to narrow the semantic representation of samples with the same labels, ICL aims to make multiple emotional representations existing in the same sample closer. That is, ICL narrows the distance among emotional representations, while not narrowing the distance among sample representations. In the ICL, for a given sample i and the indexes of i's emotion $I_Y = \{1, ..., l\}$, we could define the ICL function for the j-th emotion of each entry i as

$$L_{ICL_j} = -\frac{1}{|I_Y|}\sum_{s \in I_Y}\log\frac{\exp(\frac{sim(e_{ij}, e_{is})}{\tau})}{\sum_{k \in I_Y\setminus\{j\}}\exp(\frac{sim(e_{ij}, e_{ik})}{\tau})} \tag{5}$$

$$L_{ICL} = \frac{1}{|Y_i|}\sum_{Y_i} L_{ICL_j} \tag{6}$$

3.6 Training Objective

To train the model, we combine the contrastive loss with cross-entropy and train them jointly. This aims to use a contrastive loss to close the distance between positive samples, while maximizing the probability of correct labels through a cross-entropy loss. The overall training objective is calculated as follows:

$$L = \alpha \cdot L_{CL} + (1 - \alpha) \cdot L_{BCE} \tag{7}$$

where $L_{CL} \in \{$ *SCL, ICL, JSCL, JSPCL, SLCL*$\}$.

4 Experiments and Analysis

4.1 Dataset

In order to investigate multi-label text classification tasks, we have selected the SemEval2018 (Mohammad et al., 2018) multi-label emotion classification (MEC) task in English, Arabic, and Spanish as an illustrative example. The MEC datasets have been annotated to identify the presence of eleven discrete emotions, namely anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and trust. In order to examine the efficacy and applicability of our approach, we have conducted experiments on a multi-label news classification (MNC) task in addition to the multi-label emotion classification task. For this purpose, we utilized an open source Indonesian multi-label news classification dataset (Wang et al., 2021), comprising 8 labels including society, politics, economy, technology, military, environment, culture, and others. Each sample in the dataset is associated with at most two category labels. The datasets were initially partitioned into three distinct subsets, namely the training set (Train), validation set (Valid), and test set (Test). For the purpose of training and testing, the default partitioning method of the dataset was directly employed. We evaluate our methods using the micro F1-score, macro F1-score, and Jaccard index score (JS) in accordance with the metrics in SemEval2018 (Mohammad et al., 2018). For each language, Table 1 summarizes the train, valid, and test sets and shows the number of instances in each set.

4.2 Experimental Settings

We use SpanEmo[1] proposed by Alhuzali and Ananiadou (2021) as the base model. SpanEmo is a SOTA model for multi-label text classification tasks proposed by Alhuzali and Ananiadou (2021), they trained the model with a loss combining the cross-entropy loss and the label-correlation aware (LCA) loss (Yeh et al., 2017). We replaced the LCA loss of this model with several of our proposed contrastive losses for comparison. In addition to the SpanEmo model, we also compared the models with superior performance under each dataset separately. For the MEC task, the English models include JBNN (He and Xia, 2018), DATN (Yu et al., 2018), NTUA (Baziotis et al., 2018), LEM (Fei et al., 2020), and ReRc (Zhou et al., 2018). On the Arabic dataset, we compare our method with EMA (Badaro et al., 2018), Tw-StAR (Mulki et al., 2018a), HEF (Alswaidan and Menai, 2020) and BERT-base (Xu et al., 2020). On the Spanish dataset, we used

[1]Since our proposed method is based on SpanEmo for experiments, we also reproduce the experimental results of the method.

Info./Lang.	English	Arabic	Spanish	Indonesian
Train (#)	6,838	2,278	3,561	3373
Valid (#)	886	585	679	860
Test (#)	3,259	1,518	2,854	1841
Total (#)	10,983	4,381	7,094	6074
Classes (#)	11	11	11	8
Type	MEC	MEC	MEC	MNC

Table 1: Data Statistics.

Method	F_{Macro}	F_{Micro}	JS
BNN	52.80	63.20	-
ReRc	53.90	65.10	-
DATN	55.10	-	58.30
NTUA	52.80	70.10	58.80
LEM	56.70	67.50	-
SpanEmo	57.00	70.32	58.30
JSCL	**57.68**	**71.01**	**59.05**
JSPCL	57.42	70.75	58.58
SLCL	56.62	70.9	58.9
ICL	57.59	70.49	58.6
SCL	57.63	70.8	58.89

Table 2: Experimental results on English dataset.

Tw-StAR (Mulki et al., 2018b), ELiRF (González et al., 2018), MILAB (Mohammad et al., 2018) and BERT-base (Xu et al., 2020) as comparison models. To address the MNC task, we have identified and selected the state-of-the-art (SOTA) methods that have demonstrated superior performance on this dataset. The chosen methods comprise SGM (Yang et al., 2018), SU4MLC (Lin et al., 2018), mBERT (Xu et al., 2020), Indonesian-BERT (Wang et al., 2021), and Indonesian-BERT+Sim (Wang et al., 2021).

All experiments were carried out using PyTorch[2] and an RTX TITAN with 24 GB of memory. Using the open-source Hugging-Face implementation[3], we fine-tuned "bert-base"[4] (Wolf et al., 2020) for English. What's more, we selected "bert-base-arabic" [5] constructed by Safaya et al. (2020). for Arabic and "bert-base-spanish-uncased"[6] constructed by Canete et al. (2020) for Spanish. We set the same hyper-parameters with a fixed initialization seed for three models training, where the batch size is 32 and the feature dimension is 768. The dropout rate is 0.1 and the early stop patience we set as 10 and 20 epochs. With a learning rate of 1e-3 for the FFN and 2e-5 for the BERT encoder, Adam was chosen for optimization. For the loss weight α, we use the Hyperopt[7] hyperparameter selection method (Bergstra et al., 2011) to search for the optimal parameters under each contrastive learning method. For each model, we used five different random seeds to carry out experiments, and the scores of five experiments were averaged as the final score.

4.3 Results and Analysis

Main Performance for MEC. As shown in Table 2 to Table 4, all five of our contrastive learning strategies essentially delivered improvement to the model for the MEC task, with the JSCL approach performing best on the English dataset, reaching 57.68, 71.01 and 59.05 for F_{Macro}, F_{Micro} and JS respectively, an improvement of 0.68, 0.69 and 0.75 over the SpanEmo model. The performance improvement of our method is more obvious on the Arabic dataset, where the F_{Micro} value of the SLCL method is 1.25 higher than that of SpanEmo, and the F_{Micro} and JS of the SCL method are improved by 0.90 and 1.49 respectively. The SCL method also performed well on the Spanish language dataset, achieving the highest JS value of 53.52.

Main Performance for MNC. As shown in Table 5, in the task of MNC, the SCL method exhibits superior performance, achieving noteworthy scores of 74.29, 85.27, and 84.06 for F_{Macro}, F_{Micro} and JS metrics respectively. These remarkable results substantiate the efficacy of the SCL approach in addressing the MNC challenge. Our five contrastive learning methods have a significant improvement effect on the model on F_{Macro},, indicating that our methods can improve the categories with poor per-

[2]https://pytorch.org/
[3]https://huggingface.co/
[4]https://huggingface.co/bert-base-uncased
[5]https://huggingface.co/asafaya/bert-base-arabic
[6]https://huggingface.co/dccuchile/bert-base-spanish-wwm-uncased
[7]http://hyperopt.github.io/hyperopt/

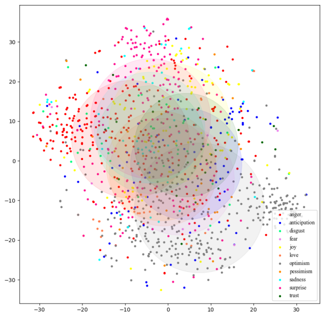

Figure 3: 2D visualization of SpanEmo's semantic space.

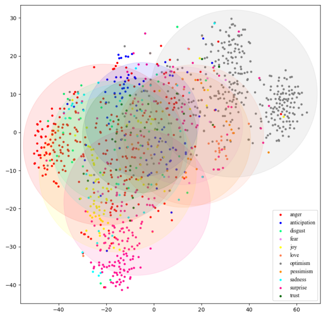

Figure 4: 2D visualization of JSPCL's semantic space

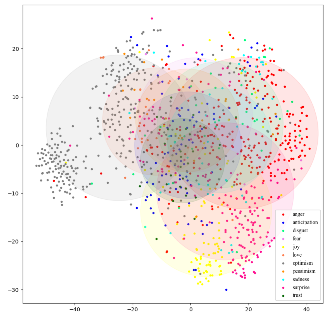

Figure 5: 2D visualization of ICL's semantic space

formance, thereby alleviating the class imbalance problem of MNC tasks to a certain extent.

Comparison between Out-sample and In-sample. In general, one particular method, referred to as ICL, exhibits comparatively less improvement. This approach primarily emphasizes contrasting labels within a single sample, considering the labels present in the text as positive examples and those absent as negative examples. However, due to its limited ability to pay attention to label relationships across different texts, ICL fails to effectively capture the inherent distinctions among labels.

Comparison between Strict Standard and Loose Standard. Through the comparison between the loose standard loss JSCL and the strict standard loss SCL, we can find that the overall performance of SCL on the four datasets is better, that is, to a certain extent, strict standard contrastive learning methods are more suitable for multi-label text classification tasks than loose standard contrastive learning methods.

Method	F_{Macro}	F_{Micro}	JS
Tw-StAR	44.60	59.70	46.50
EMA	46.10	61.80	48.90
BERT_{base}	47.70	65.00	52.30
HEF	50.20	63.10	51.20
SpanEmo	53.63	65.81	53.94
JSCL	54.08	66.00	54.14
JSPCL	53.70	65.86	53.98
SLCL	**54.88**	66.37	54.65
ICL	54.26	66.13	54.17
SCL	54.27	**66.71**	**55.43**

Table 3: Experimental results on Arabic dataset.

Comparison between ProCL and FeaCL. Through the results of JSCL and JSPCL, we could find that the method of the ProCL type does not perform as well as the method of the FeaCL type in terms of performance. We believe that because the semantic space of the multi-label text model is too complex, it is more effective to directly focus on the semantic space of the model than the probability distribution.

Method	F_{Macro}	F_{Micro}	JS
Tw-StAR	39.20	52.00	43.80
ELiRF	44.00	53.50	45.80
MILAB	40.70	55.80	46.90
BERT_{base}	47.40	59.60	48.70
SpanEmo	55.49	63.34	52.68
JSCL	55.62	63.45	52.94
JSPCL	**56.44**	**64.16**	53.31
SLCL	56.00	63.56	52.69
ICL	55.82	63.46	52.66
SCL	55.88	63.70	**53.52**

Table 4: Experimental results on Spanish dataset.

Interpretable Analysis. Taking the experimental results in Spanish as an example, we analyze the interpretability of our method from the multi-label dimension and the single-label dimension respectively. In the multi-label dimension, we use the entire test set for analysis, consider the samples with identical labels to be under the same cluster, and then use the T-SNE method for dimensionality reduction and visualization. At the same time, we also calculated the Calinski-Harbasz score of cluster clustering to evaluate whether the semantic representation space of each category can be well discriminated. It is worth noting that under the single-label dimension, we only use the test set

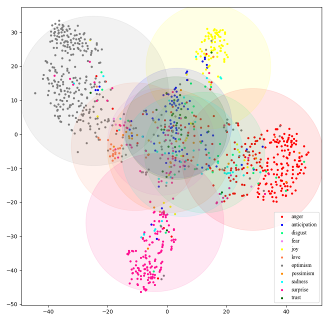

Figure 6: 2D visualization of JSCL's semantic space

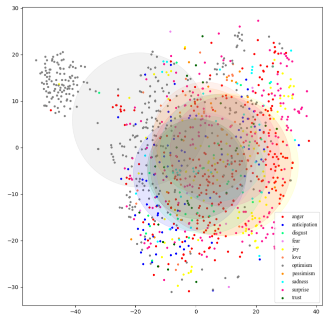

Figure 7: 2D visualization of SLCL's semantic space

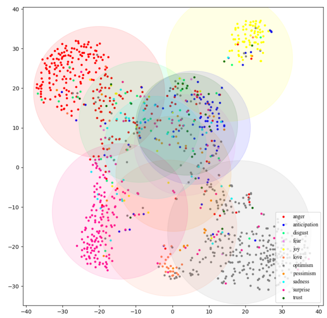

Figure 8: 2D visualization of SCL's semantic space

Method	F_{Macro}	F_{Micro}	JS
SGM	44.08	74.24	-
SU4MLC	40.38	75.66	-
mBERT	66.56	81.85	-
Indonesian-BERT	67.57	84.53	-
Indonesian-BERT +Sim	70.82	84.66	-
SpanEmo	71.62	85.09	83.66
JSCL	73.13	85.23	83.98
JSPCL	72.17	84.91	82.81
SLCL	73.47	85.19	83.86
ICL	74.22	85.15	83.82
SCL	**74.29**	**85.27**	**84.06**

Table 5: Experimental results on Indonesian dataset.

Method	Multi-label	Singel-label
SpanEmo	5.64	48.07
JSCL	24.07	**200.48**
JSPCL	17.80	131.54
SLCL	4.35	42.33
ICL	13.43	109.55
SCL	**25.14**	198.84

Table 6: Interpretable analysis results

with only one label for interpretable analysis.

The interpretable analysis results for each method in the multi-label dimension and the single-label dimension are shown in Table 6. The larger the interpretable analysis results, the higher the discrimination of samples of different categories in the semantic space, and the better the semantic representation ability of the model. It can be seen that in addition to SLCL, other contrastive learning methods can make the samples of the same category in the semantic space more compact, and the boundaries of sample clusters of different categories are more obvious. SLCL aims to narrow the representation of categories, so it cannot make the boundaries between different categories more obvious. Among them, JSCL and SCL have better effects in optimizing the semantic representation space. As a rigorous contrastive learning method, SCL achieves the best results on multi-label dimension evaluation, with a Calinski-Harbasz value of 25.14. When evaluates from a multi-label perspective, JSCL performs slightly worse than SCL, but when evaluated from a single-label perspective, JSCL achieves the highest Calinski-Harbasz score of 200.48. We also further visualize the semantic space under the single-label dimension, as shown in Figures 3 to 8. It can be clearly seen that in JSCL and SCL, each category is more closely aggregated, and the boundaries among different categories are also more obvious.

5 Conclusion

To investigate the efficacy of contrastive learning using various methodologies, we offer five effective contrastive losses for multi-label text classification tasks. The experimental results of this paper show that contrastive loss can improve the performance of multi-label text classification tasks. Furthermore, we find that strict criteria contrastive learning and feature-based contrastive learning outperform other contrastive learning methods on multi-label text classification tasks. In the future, based on these two methods, we will further explore the contrastive loss that is more suitable for multi-label text classification tasks.

Acknowledgements

This work was supported by the Guangdong Basic and Applied Basic Research Foundation of China (No. 2023A1515012718).

Limitations

This paper proposes five novel contrastive losses for multi-label text classification tasks. However, our method has the following limitations:

1. We only selected the multi-label emotion classification task and multi-label news classification as the representative of the multi-label text classification tasks.

2. We only conduct experiments on the single modal of text, and have not extended to multi-modal tasks.

3. Our method chooses the SpanEmo model as the backbone, lacking attempts to more models.

References

Hassan Alhuzali and Sophia Ananiadou. 2021. SpanEmo: Casting multi-label emotion classification as span-prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1573–1584, Online. Association for Computational Linguistics.

Nourah Alswaidan and Mohamed El Bachir Menai. 2020. Hybrid feature model for emotion recognition in arabic text. *IEEE Access*, 8:37843–37854.

Iqra Ameer, Necva Bölücü, Muhammad Hammad Fahim Siddiqui, Burcu Can, Grigori Sidorov, and Alexander Gelbukh. 2023. Multi-label emotion classification in texts using transfer learning. *Expert Systems with Applications*, 213:118534.

Gilbert Badaro, Obeida El Jundi, Alaa Khaddaj, Alaa Maarouf, Raslan Kain, Hazem Hajj, and Wassim El-Hajj. 2018. EMA at SemEval-2018 task 1: Emotion mining for Arabic. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 236–244, New Orleans, Louisiana. Association for Computational Linguistics.

Junwen Bai, Shufeng Kong, and Carla P Gomes. 2022. Gaussian mixture variational autoencoder with contrastive learning for multi-label classification. In *International Conference on Machine Learning*, pages 1383–1398. PMLR.

Christos Baziotis, Athanasiou Nikolaos, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth Narayanan, and Alexandros Potamianos. 2018. NTUA-SLP at SemEval-2018 task 1: Predicting affective content in tweets with deep attentive RNNs and transfer learning. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 245–255, New Orleans, Louisiana. Association for Computational Linguistics.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for hyper-parameter optimization. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, page 2546–2554, Red Hook, NY, USA. Curran Associates Inc.

José Canete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020:1–10.

Ilias Chalkidis and Anders Søgaard. 2022. Improved multi-label classification under temporal concept drift: Rethinking group-robust algorithms in a label-wise setting. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2441–2454, Dublin, Ireland. Association for Computational Linguistics.

Francisco Charte, Antonio J. Rivera, María J. del Jesus, and Francisco Herrera. 2014. Mlenn: A first approach to heuristic multilabel undersampling. In *Intelligent Data Engineering and Automated Learning – IDEAL 2014*, pages 1–9, Cham. Springer International Publishing.

Son D Dao, Ethan Zhao, Dinh Phung, and Jianfei Cai. 2021. Multi-label image classification with contrastive learning. *CoRR*, abs/2107.11626.

Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. 2020. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7692–7699.

Yingwen Fu, Nankai Lin, Ziyu Yang, and Shengyi Jiang. 2022. A dual-contrastive framework for low-resource cross-lingual named entity recognition. *CoRR*, abs/2204.00796.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

José-Ángel González, Lluís-F. Hurtado, and Ferran Pla. 2018. ELiRF-UPV at SemEval-2018 tasks 1 and 3: Affect and irony detection in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 565–569, New Orleans, Louisiana. Association for Computational Linguistics.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *CoRR*, abs/2011.01403.

Huihui He and Rui Xia. 2018. Joint binary neural network for multi-label learning with applications to emotion classification. In *Natural Language Processing and Chinese Computing*, pages 250–259, Cham. Springer International Publishing.

Ruining He, Anirudh Ravula, Bhargav Kanagal, and Joshua Ainslie. 2021. RealFormer: Transformer likes residual attention. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 929–943, Online. Association for Computational Linguistics.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50.

Xincheng Ju, Dong Zhang, Junhui Li, and Guodong Zhou. 2020. Transformer-based label set generation for multi-modal multi-label emotion detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 512–520, New York, NY, USA. Association for Computing Machinery.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673.

Irene Li, Aosong Feng, Hao Wu, Tianxiao Li, Toyotaro Suzumura, and Ruihai Dong. 2022. LiGCN: Label-interpretable graph convolutional networks for multi-label text classification. In *Proceedings of the 2nd Workshop on Deep Learning on Graphs for Natural Language Processing (DLG4NLP 2022)*, pages 60–70, Seattle, Washington. Association for Computational Linguistics.

Junjie Li, Yixin Zhang, Zilei Wang, and Keyu Tu. 2021. Probability contrastive learning for domain adaptation. *CoRR*, abs/2111.06021.

Junyang Lin, Qi Su, Pengcheng Yang, Shuming Ma, and Xu Sun. 2018. Semantic-unit-based dilated convolution for multi-label text classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4554–4564, Brussels, Belgium. Association for Computational Linguistics.

Nankai Lin, Sihui Fu, Xiaotian Lin, and Lianxi Wang. 2022. Multi-label emotion classification based on adversarial multi-task learning. *Information Processing and Management*, 59(6):103097.

Huiting Liu, Geng Chen, Peipei Li, Peng Zhao, and Xindong Wu. 2021. Multi-label text classification via joint learning from label embedding and label correlation. *Neurocomputing*, 460:385–398.

Qianwen Ma, Chunyuan Yuan, Wei Zhou, and Songlin Hu. 2021. Label-specific dual graph neural network for multi-label text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3855–3864, Online. Association for Computational Linguistics.

Mikołaj Małkiński and Jacek Mańdziuk. 2022. Multi-label contrastive learning for abstract visual reasoning. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–13.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465.

Hala Mulki, Chedi Bechikh Ali, Hatem Haddad, and Ismail Babaoğlu. 2018a. Tw-StAR at SemEval-2018 task 1: Preprocessing impact on multi-label emotion classification. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 167–171, New Orleans, Louisiana. Association for Computational Linguistics.

Hala Mulki, Chedi Bechikh Ali, Hatem Haddad, and Ismail Babaoğlu. 2018b. Tw-StAR at SemEval-2018 task 1: Preprocessing impact on multi-label emotion classification. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 167–171, New Orleans, Louisiana. Association for Computational Linguistics.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. SemEval-2013 task 2: Sentiment analysis in Twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA. Association for Computational Linguistics.

Jinseok Nam, Eneldo Loza Mencía, Hyunwoo J Kim, and Johannes Fürnkranz. 2017. Maximizing subset accuracy with recurrent neural networks in multi-label classification. In *Advances in Neural Information Processing Systems*, volume 30.

Yujia Qin, Yankai Lin, Ryuichi Takanobu, Zhiyuan Liu, Peng Li, Heng Ji, Minlie Huang, Maosong Sun, and Jie Zhou. 2021. ERICA: Improving entity and relation understanding for pre-trained language models via contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3350–3363, Online. Association for Computational Linguistics.

Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2021. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 82–91.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Boaz Shmueli, Soumya Ray, and Lun-Wei Ku. 2021. Happy dance, slow clap: Using reaction GIFs to predict induced affect on Twitter. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 395–401, Online. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Xi'ao Su, Ran Wang, and Xinyu Dai. 2022. Contrastive learning-enhanced nearest neighbor mechanism for multi-label text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 672–679, Dublin, Ireland. Association for Computational Linguistics.

Peter Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Lianxi Wang, Xiaotian Lin, and Nankai Lin. 2021. Research on pseudo-label technology for multi-label news classification. In *Document Analysis and Recognition – ICDAR 2021*, pages 683–698, Cham. Springer International Publishing.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yaoqiang Xiao, Yi Li, Jin Yuan, Songrui Guo, Yi Xiao, and Zhiyong Li. 2021. History-based attention in seq2seq model for multi-label text classification. *Knowledge-Based Systems*, 224:107094.

Peng Xu, Zihan Liu, Genta Indra Winata, Zhaojiang Lin, and Pascale Fung. 2020. Emograph: Capturing emotion correlations using graph networks. *CoRR*, abs/2008.09378.

Yu Xu, Dong Zhou, and Séamus Lawless. 2016. Inferring your expertise from twitter: Integrating sentiment and topic relatedness. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 121–128.

Pengcheng Yang, Fuli Luo, Shuming Ma, Junyang Lin, and Xu Sun. 2019. A deep reinforced sequence-to-set model for multi-label classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5252–5258, Florence, Italy. Association for Computational Linguistics.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. SGM: Sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3915–3926, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Chih-Kuan Yeh, Wei-Chieh Wu, Wei-Jen Ko, and Yu-Chiang Frank Wang. 2017. Learning deep latent space for multi-label classification. In *Thirty-first AAAI conference on artificial intelligence*.

Selim F. Yilmaz, E. Batuhan Kaynak, Aykut Koç, Hamdi Dibeklioğlu, and Suleyman Serdar Kozat. 2021. Multi-label sentiment analysis on 100 languages with dynamic weighting for label imbalance. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–13.

Jianfei Yu, Luís Marujo, Jing Jiang, Pradeep Karuturi, and William Brendel. 2018. Improving multi-label emotion classification via sentiment classification with dual attention transfer network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1097–1102, Brussels, Belgium. Association for Computational Linguistics.

Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah. 2022a. Use all the labels: A hierarchical multi-label contrastive learning framework. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16639–16648.

Ximing Zhang, Qian-Wen Zhang, Zhao Yan, Ruifang Liu, and Yunbo Cao. 2021. Enhancing label correlation feedback in multi-label text classification via multi-task learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1190–1200, Online. Association for Computational Linguistics.

Yangjun Zhang, Pengjie Ren, Wentao Deng, Zhumin Chen, and Maarten Rijke. 2022b. Improving multi-label malevolence detection in dialogues through multi-faceted label correlation enhancement. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3543–3555, Dublin, Ireland. Association for Computational Linguistics.

Deyu Zhou, Yang Yang, and Yulan He. 2018. Relevant emotion ranking from text constrained with emotion relationships. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 561–571, New Orleans, Louisiana. Association for Computational Linguistics.

Yangyang Zhou, Xin Kang, and Fuji Ren. 2022. Employing contrastive strategies for multi-label textual emotion recognition. In *Intelligent Information Processing XI*, pages 299–310, Cham. Springer International Publishing.

Yu Zhu and Ou Wu. 2022. Elementary discourse units with sparse attention for multi-label emotion classification. *Knowledge-Based Systems*, 240:108114.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
section 6

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
abstract and section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Not applicable. Left blank.

C ☑ Did you run computational experiments?

section 4

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Our paper focuses more on the performance of the model.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
section 4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
section 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
section 4

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.