Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

A novel dropout mechanism with label extension schema toward text emotion classification☆,☆☆

Zongxi Li [a], Xianming Li [b], Haoran Xie [c], Fu Lee Wang [a,*], Mingming Leng [c], Qing Li [d], Xiaohui Tao [e]

[a] School of Science and Technology, Hong Kong Metropolitan University, Kowloon, Hong Kong Special Administrative Region of China
[b] Ant Group, Shanghai, PR China
[c] Department of Computing and Decision Sciences, Lingnan University, Tuen Mun, Hong Kong Special Administrative Region of China
[d] Department of Computing, Hong Kong Polytechnic University, Kowloon, Hong Kong Special Administrative Region of China
[e] School of Sciences, University of Southern Queensland, Toowoomba, Australia

ARTICLE INFO

Keywords:
Leaky dropout
Emotion classification
Sentiment analysis
Label extension
Distribution learning

ABSTRACT

Researchers have been aware that emotion is not one-hot encoded in emotion-relevant classification tasks, and multiple emotions can coexist in a given sentence. Recently, several works have focused on leveraging a distribution label or a grayscale label of emotions in the classification model, which can enhance the one-hot label with additional information, such as the intensity of other emotions and the correlation between emotions. Such an approach has been proven effective in alleviating the overfitting problem and improving the model robustness by introducing a distribution learning component in the objective function. However, the effect of distribution learning cannot be fully unfolded as it can reduce the model's discriminative ability within similar emotion categories. For example, "Sad" and "Fear" are both negative emotions. To address such a problem, we proposed a novel emotion extension scheme in the prior work (Li, Chen, Xie, Li, and Tao, 2021). The prior work incorporated fine-grained emotion concepts to build an extended label space, where a mapping function between coarse-grained emotion categories and fine-grained emotion concepts was identified. For example, sentences labeled "Joy" can convey various emotions such as *enjoy*, *free*, and *leisure*. The model can further benefit from the extended space by extracting dependency within fine-grained emotions when yielding predictions in the original label space. The prior work has shown that it is more apt to apply distribution learning in the extended label space than in the original space. A novel sparse connection method, i.e., Leaky Dropout, is proposed in this paper to refine the dependency-extraction step, which further improves the classification performance. In addition to the multiclass emotion classification task, we extensively experimented on sentiment analysis and multilabel emotion prediction tasks to investigate the effectiveness and generality of the label extension schema.

☆ This document is the results of the research project funded by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (UGC/FDS16/E01/19), and Faculty Research Grants (DB22B4, DB22B7, and DB23A3) and Lam Woo Research Fund (LWP20019) of Lingnan University, Hong Kong, and the General Research Fund from the Hong Kong Research Grants Council (Project no. 11204919).

☆☆ This article is an extended journal version based on our conference paper, which has been published at Li, Chen, Xie, Li, and Tao (2021). We have sufficient new contents and contributions for this extended journal version according to the journal regulations. Some notations, equations, algorithms, descriptions, tables and figures may be re-used from this paper for smooth presentations.

* Corresponding author.
E-mail addresses: zoli@hkmu.edu.hk (Z. Li), niming.lxm@antgroup.com (X. Li), hrxie@ieee.org (H. Xie), pwang@hkmu.edu.hk (F.L. Wang), mmleng@ln.edu.hk (M. Leng), qing-prof.li@polyu.edu.hk (Q. Li), xiaohui.tao@usq.edu.au (X. Tao).

https://doi.org/10.1016/j.ipm.2022.103173
Received 2 August 2022; Received in revised form 6 November 2022; Accepted 15 November 2022
Available online 30 November 2022

1. Introduction

Understanding people's opinions and sentiments from online posts and comments can benefit stakeholders, such as social media advertisements and e-commerce (Chen, Xie, Cheng, & Li, 2022; Yang & Wang, 2003, 2007). The advent of natural language processing (NLP) technologies has empowered stakeholders to analyze online-generated content with a volume far exceeding the human reading ability. Deep learning-based emotion detection and sentiment analysis can improve service quality and create new business opportunities. Among various tasks concerning users' emotions, two emotion-relevant text classification tasks have received tremendous attention from academia and industry. The first task is sentiment analysis, which mines user-generated content's sentiment orientation and intensity. The other is emotion classification which aims to identify the specific emotion categories.

When attending to classification tasks, whether for images or texts, it is common to utilize the one-hot label for computing the cross-entropy loss function. Such a methodology works perfectly well in a task where the labels are objective things or categories. If an image is labeled "Cat", there is a cat in the image; if a news article is categorized as "Sport News", then the content will be related to sports events. However, emotion is all about subjectiveness. A different audience can indistinctly perceive the categories and intensities of emotion expressed by the same sentence. The recognition of mixed emotions, or fuzzy emotions, has been examined under both image and multimodal settings (Aly & Tapus, 2015; Liliana, Basaruddin, & Widyanto, 2017; Liliana, Basaruddin, Widyanto, & Oriza, 2019). In text emotion classification, researchers have discerned that mapping texts to labels in the conventional one-hot encoding approach is inadequate as the relation between texts and labels is not adequately revealed. Recently, many works have followed the idea of employing distribution learning in a classification task to address the emotion classification task (Fei, Zhang, Ren, & Ji, 2020; Guo, Han, Han, Huang, & Lu, 2021; Lee, 2022; Li, Li, Xie, Li and Tao, 2021; Qin et al., 2021; Xu, Liu, & Geng, 2020; Zhang et al., 2018; Zhao & Ma, 2019; Zhou, Zhang, Zhou, Zhao, & Geng, 2016, *inter alia*).

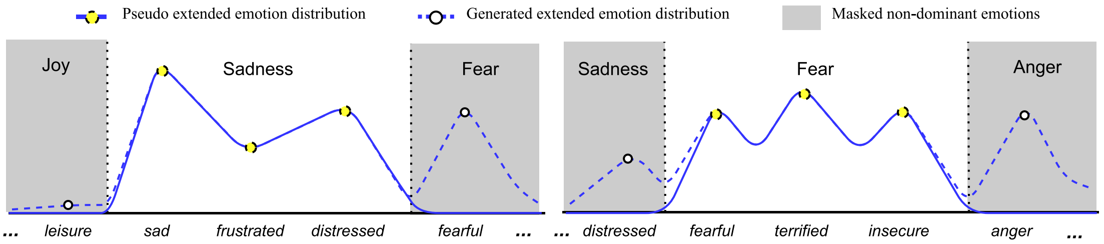
We categorize the existing efforts into two genres. The first genre focuses on distribution prediction instead of single-label prediction. The works under this genre adopt a machine learning approach to learn an emotion distribution explicitly, such as Qin et al. (2021), Xu et al. (2020), Zhou et al. (2016), and Zhu et al. (2017), where a distribution label has already been annotated in the corresponding datasets. The models aim to optimize the divergence between the predicted distribution and distribution label. This kind of work is limited in quantity, as the datasets with distribution labels are scarce in the community due to expensive annotation. The second category (Fei et al., 2020; Guo et al., 2021; Lee, 2022; Zhang et al., 2018) attends to the label ambiguity issue brought by training with one-hot labels. Label ambiguity refers to the challenge caused by the unreliability, deficiency, and even errors in the ground-truth labels (Gao, Xing, Xie, Wu, & Geng, 2017). Incorporating distribution learning into a classification framework has been proven beneficial. Although it is not practical to deal with wrong labels, we can address the ambiguity issue by enhancing the incomplete label information. Existing works generate a distribution label by generating weights for all label classes based on text input. Such a distribution label encompasses additional features that can make up for the missing information in one-hot labels. Furthermore, training with smoothly distributed labels prevents overconfidence in the model and enhances its robustness. The research described in this article follows the second genre.

Despite the recent success, leveraging a generated distribution label in the classification framework causes two potential problems.

1. First, it is contradictory to introduce a distribution label into a single-label prediction task, even if an accurate distribution label is available. Intuitively, we hope the signal of the ground-truth label can be as prominent as possible in training to differentiate instances with the same label from others, but the distribution label does the opposite. Among all the emotion labels, a subset could have comparable weights in the distribution label. For example, given sentences narrating the event "failing an examination", the generated distribution may share similar weights for negative emotions, such as "Shame", "Guilty", and "Fear". On the other hand, positive emotions "Joy" and "Surprise" will have similar weights when a student "passes the examination". Employing such an emotion distribution will confuse the classifier in the multiclass classification task where a single-label prediction is expected. Consequently, the classifier will struggle to differentiate the labels of similar emotions. This is the problem that we refer as *interclass confusion* of applying the distribution learning in the classification task.
2. Second, improper values in the distribution introduce noise in the distribution learning and harm the classification model. Since most datasets do not have ground-truth distribution labels, researchers have to define the meaning of the distribution and produce distributions with unsupervised methods from texts, such as model-based methods (Guo et al., 2021; Lee, 2022) and rule-based methods using lexicons (Li, Li, Xie, Li, Tao, 2021; Li, Xie, Cheng and Li, 2021; Zhang et al., 2018). However, it is intractable for these unsupervised methods to precisely profile the sentence-level representation due to semantic compositionality and sentiment deviation within the sentence. Consequently, the obtained distribution label could be far from a reasonable label with undesirable values. The training stage will take in the noisy signal and deteriorate classification performance.
3. Thirdly, because of the previous two pointers, existing works, intentionally or unintentionally, take the edge off distribution learning in the framework, although they claim to leverage the goodness of distribution learning. The learning of single-label classification plays a dominant role in the overall learning stage, and distribution learning works as an auxiliary component in the loss function.

The abovementioned problems pose genuine restrictions on applying distribution learning in the classification task.

We argue that generating emotion distributions from the original label space accounts for the limitations. Thus, a label extension approach exploiting fine-grained emotions was proposed in our prior work to create an extended label space (Li, Li, Xie, Li, Tao,



(a) Somebody's 22 year old son was killed this morning. (Sadness)

(b) I was attacked by a teenage boy and had my wallet stolen. (Fearful)

Fig. 1. To facilitate the understanding, we present examples of the generated emotion distribution adopted from Li, Xie et al. (2021). By extending the label space with fine-grained concepts, we can generate the extended emotion distribution for sentences by using a rule-based method. In the pseudo distribution, we mask the concepts of non-dominant emotions as 0.

2021), promising to address these challenges. Due to the subjective nature, extending labels for emotion task is more difficult than other tasks like image classification, as emotions cannot be objectively and quantitatively characterized. Therefore, we need domain knowledge to identify the fine-grained emotion categories and their relationship with each emotion label in the dataset. However, no such work has studied this issue, so we have to find a feasible approach based on the available knowledge, such as domain theory, emotion lexicon, and dictionary, which are essential to our task. Viewing the descriptions on the annotation process from the papers of popular emotion classification datasets, such as ISEAR (Scherer & Wallbott, 1994) and SemEval (Mohammad, Bravo-Marquez, Salameh, & Kiritchenko, 2018), the annotating works start from finding an apt emotion theory as the theoretical foundation. The commonly adopted emotion theories are the *six basic emotions* (Ekman, 1992), *wheel of emotions* (Plutchik, 1980), and *hourglass model* (Susanto, Livingstone, Ng, & Cambria, 2020). These theories categorize the emotions universally perceived by people from different cultural backgrounds, such as "Sadness", "Anger", and "Joy", which are general emotion categories. Nevertheless, we notice that each emotion category has broad coverage and can breakdown into subtle concepts. Sentences under the same emotion category can express delicate meaning from a fine-grained perspective. For example, we label the sentences portraying the following situations as "Joy": "playing video games with friends", "petting the dog with families on the weekend", and "getting promoted in the company". However, different fine-grained emotions can be identified: "playing games" makes people feel *happy*,[1] "getting promoted in the company" is an *exciting* moment, and "petting the dog with families on the weekend" is an *enjoyable* thing, in which people feel *leisure*. Regarding "Sadness", people may feel *helpless* when their houses are destroyed by a tornado and feel *anguished* if they suffer from deadly cancers. We exploit three-factor theory (Russell & Mehrabian, 1977), a classic psychology domain knowledge that identifies and describes 151 fine-grained emotion concepts, to create the mapping between coarse-grained emotion "Categories" and fine-grained emotion *concepts*. Concretely, for each emotion category, which has been used as a label in a dataset, we have established a set of fine-grained emotions associated with it. We consolidate the existing lexicon knowledge as the theoretical foundation, such as SenticNet 6[2] (Cambria, Li, Xing, Poria, & Kwok, 2020) and NRC Word-Emotion Association Lexicon[3] (Mohammad & Turney, 2013), with manual deliberation.

Furthermore, inspired by Li, Li, Xie, Li, Tao (2021), Li, Xie et al. (2021) generated an extended distribution representation for words. The sentence-level distribution is produced with an unsupervised method. In particular, the pseudolabel only covers the distribution of the ground-truth label, and the remaining entries of the pseudolabel are masked as 0. We provide two examples to facilitate understanding, as shown in Fig. 1. We propose that the model first learns the distribution at the extended label space and decides from the original labels based on the extended distribution. To serve this purpose, we employ the generated pseudolabel and the ground-truth label at the penultimate layer and the output layer of the classifier as constraints, respectively. Learning the extended label space is deemed more challenging than learning the original labels, so the model can extract and exploit more informative features. Subsequently, the learned distribution is projected back to the original label space with a fully-connected layer. In this process, the output layer can learn a refined transformation guided by the ground-truth label. If the model learns a less precise distribution in the penultimate layer, it is possible to make corrections in the output layer. Training with such a pseudolabel benefits the classification task from two perspectives:

1. *Interclass confusion* is shifted to innocuous *intraclass confusion*. Learning a distribution for all emotion labels raises the risk of making a wrong decision among similar labels, which will be considered a false prediction in the evaluation metrics. Nevertheless, in the extended label space, there is only the distribution of fine-grained emotion *concepts* from the same emotion "Category". Even if the model yields a wrong prediction within these fine-grained *concepts*, the final output is still correct.

[1] For concise writing and easy understanding to the audience, we use "Title Case" to refer to the emotion category, and use *italics* to refer to fine-grained emotion concepts in the whole manuscript.

[2] https://sentic.net/api/.

[3] http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm.

2. Noises are avoided from nondominant emotions. The masking operation removes the information of nondominant emotion categories from the pseudo distribution. Therefore, no noise from other labels is introduced in the training stage.

Hence, the model can tolerate errors in the generated distribution.

We enhance the label extension schema by proposing an innovative sparse connection mechanism, yielding substantial improvements to our model. A Dropout layer was found to be helpful in refining the prediction when the model projected back to the original label space (Li, Li, Xie, Li, Tao, 2021). We note that the sparse connection is an important component in our proposed extension framework, as the transformation from the extended label space to the original space is a self-correction process for the classification model. Particularly, this self-correction process is where the improvements come from. However, we notice that the conventional Dropout mechanism can reduce the effect of exploiting dependencies in the extended label space as additional information. Moreover, when the dropout rate is relatively large, the essential information of concerned fine-grained concepts can be deactivated, compromising the classification performance. We propose a Leaky Dropout mechanism to reduce the value of a selected neuron instead of abandoning it, which is proven to be highly effective in our experiments. Although Leaky Dropout is a simple mechanism, we have shown that it is helpful to stabilize the training process with extensive experiments.

Our initial focus is on the multiclass classification task. At the same time, we carried out subsequent work to validate the effectiveness of sentiment analysis and multilabel classification tasks. This work's contributions can be concluded as follows:

- We recognize a mapping function between emotion labels widely adopted in datasets and fine-grained emotion concepts from domain knowledge.
- We devise a novel emotion extension schema based on the identified mapping.
- We design an emotion classification framework incorporating distribution learning in the extended label space. In addition to a multiclass dataset, this framework is proven effective for tasks such as sentiment analysis and multilabel classification.
- We propose a new sparse connection method to boost model performance and conduct experiments to investigate how the leaky factor affects the classification model.

Preliminary research of this work was published in Li, Li, Xie, Li, Tao (2021). This manuscript extends the prior work from both methodological and experimental perspectives. The new enhancements include the following: (1) we suggest a novel sparse connection, i.e., Leaky Dropout, to refine the conversion from the extended to the original space; (2) in addition to the multiclass classification task, we also experiment with the sentiment analysis task and multilabel classification task, which shows the generality of the proposed framework in emotion-relevant NLP tasks; and (3) we conduct an extensive investigation and provide more insights into the proposed Leaky Dropout from empirical studies. To be consistent with the prior publication and maintain the integrity of logic, we reused some content from the conference version.

2. Related works

Our research addresses a domain-specific text classification using a deep learning technique and leverages domain knowledge to incorporate additional information in the learning. Therefore, we review the works and studies in the relevant fields in this section.

2.1. Text classification and text mining

Neural networks have been extensively employed as semantic feature extractors for different purposes. Kim (2014) proposed a classic text convolutional neural network (TextCNN), which extracts local and position-invariant features from the word embedding matrix. Numerous works have been conducted to leverage TextCNN as a feature extractor. In addition to CNNs, recurrent neural networks (RNNs) are also mainstream feature extractors for semantic mining. Socher et al. (2013) employed recursive networks to model time-series features from the text. Various works have diversified the recurrent model with different variants, such as bidirectional long short-term memory (BiLSTM) and gated recurrent units (GRUs), with more complex gate mechanisms. A breakthrough occurred in 2017 when Vaswani et al. (2017) from Google presented Transformer by stacking multiple self-attention blocks together. Transformer has an encoder-to-decoder framework and can learn powerful sentence-level representations. Devlin, Chang, Lee, and Toutanova (2019) revisited the language model and proposed a pre-trained language model, Bert, by incorporating a self-attention-based architecture and an enormous text corpus. The resultant language model presented a powerful capacity by benchmarking the state-of-the-art performance in a wide spectrum of downstream tasks. Inspired by Bert, the family of large-scale pre-trained language models has grown with new variants.

The applications of text mining have gained significant attention from academia and industry (Altınel & Ganiz, 2018; Huang, Xie, Rao, Feng, & Wang, 2020; Li et al., 2016; Tubishat, Idris, & Abushariah, 2018). A number of models have yielded competitive results on benchmark datasets, which can be done in two ways: machine learning-based approaches and lexicon-based approaches. Machine learning methods extract and utilize semantic or linguistic patterns from texts for classification. Huang, Rao, Xie, Wong, and Wang (2017) proposed a topic-based machine learning model for cross-domain sentiment classification. Ali et al. (2019) employed an ontology-based topic modeling model to analyze online-generated contents related to traffic enhancing transportation management services. Liang, Xie, Rao, Lau, and Wang (2018) proposed a topic model-based universal affective method for short text classification. Feng, Rao, Xie, Wang, and Li (2020) presented a user group-based topic modeling for short text emotion mining. Li, Li, Xie and Li (2021) proposed merging traditional statistical information into the deep learning framework. The corpus-level statistics are encoded as feature vectors by a variational autoencoder. A well-designed adaptive gate network (AGN) is used to consolidate semantic features with additional information selectively, according to their confidence toward the prediction. Apart from supervised models, some clustering-based (Guan et al., 2022) and contrastive-based (Fu & Liu, 2022) self-supervised models have recently been proposed to address NLP tasks. Lexicon-based methods incorporate handcrafted lexicon knowledge into the decision making process. Some works have proposed homemade lexicons for specific purposes. Li, Chen, Xie, Li, and Tao (2020) encoded the intensity variation of emotions, with the help of domain knowledge, in the sentence as an EmoChannel vector and exploited a simple attention layer to leverage the dependency within fine-grained emotions in the classification framework. Li, Chen, Zhong, Gong, and Han (2022) integrated social cognitive theory and the dedicated Intent-Indicator sentiment lexicon for emotional analysis of online dating services' comments. The other works that fall into this field adopt public emotion and sentiment lexicons, which will be elaborated in the next subsection.

2.2. *Emotion theory and emotional lexicon*

Since we incorporate various emotion theories and lexicons, reviewing relevant knowledge and works is necessary. Ekkekakis and Russell (2013) suggested that the existing emotion theories characterize emotions from either categorical or dimensional perspectives. Categorical theories define emotions as discrete classes, that is, each emotion is an individual category and independent from other categories, and develop a set of terms to "characterize the state of mind", such as the six basic emotions (Ekman, 1992), wheel of emotion (Plutchik, 1980), and hourglass model (Susanto et al., 2020). These theories identify the most prominent and representative emotion concepts humans can universally recognize. A lexicon dictionary annotates each word according to the theory it follows. For example, SenticNet (Cambria et al., 2020) adopts eight emotion tags following hourglass model theory, and the NRC Emotion Lexicon (also known as EmoLex) (Mohammad & Turney, 2013) follows the wheel of emotion (Plutchik, 1980) and adopts eight prototypical emotions. Lexicon-based approaches have been widely used in domain-dependent text classification tasks. Muñoz and Iglesias (2022) proposed a machine learning-based framework to detect psychological stress from texts by combining various lexicon resources. Although such lexicons explicitly associate words with emotions, it is inflexible to employ these lexicons in practice (Li, Xie et al., 2021), as the tags in the lexicon may be inconsistent with the dataset labels, and the binary indication does not provide the intensity of emotion. On the other hand, dimensional theories build a system to profile emotions with different factors, such as three-factor theory (Russell & Mehrabian, 1977) and the circumplex model (Posner, Russell, & Peterson, 2005). Dimensional lexicons, such as NRC-VAD (Mohammad, 2018), provide a tuple of three values. The values denote the information of *valence*, *arousal*, and *dominance* (VAD), respectively. Such information quantitatively portrays the word with three factors in the VAD space. The setbacks come as they can associate a concept with emotion categories. A recent study (Li, Xie et al., 2021) devised a method to generate a general emotional representation of words combining three-factor theory and the NRC-VAD lexicon. We adopt a similar method to quantitatively calculate the intensity of each emotion expressed in a sentence.

2.3. *Emotion distribution learning*

More attention has focused on addressing emotion classification with the distribution learning approach. In particular, Geng (2016) devised the label distribution learning framework primarily for computer vision applications, where the distribution represents the corresponding entities' proportion in an image. Zhou et al. (2016) first identified both emotion categories and their corresponding intensities with a ground-truth emotion distribution as the target in the distribution learning framework. Domain knowledge is incorporated as a constraint to improve learning accuracy. Xu et al. (2020) proposed a new partial multilabel learning strategy to enhance the label information by using multilabel data. Topological information in the feature space and label dependencies are used to reconstruct the label distributions. An important preface of the works mentioned above is that the ground-truth emotion distribution is available in the dataset, so these works can concentrate on building the machine learning framework. However, such datasets are rare in the NLP community. Thus, a compromising research direction is generating pseudo distributions to facilitate single-label classification. Biddle, Joshi, Liu, Paris, and Xu (2020) leveraged sentiment distributions for health mention classification from literal health reports on Twitter. Zhang et al. (2018) was the first work to leverage distribution learning into text emotion classification. They used a categorical lexicon and devised a naïve increment strategy to produce a pseudo distribution. The generated distribution label is exploited under a multitask learning framework. Following a similar philosophy, Li, Xie et al. (2021) focused on the generation of distribution and proposed a rule-based strategy based on the word emotion distribution. The representation is optimized by minimizing the Kullback–Leibler divergence in the latent space. Lee (2022) was also aware that emotion is not one-hot in the sentences; hence, five methods were conceived to generate grayscale emotion labels automatically. A similar conclusion was drawn that introducing a distribution label is helpful to the classification task.

These works generate emotion distributions based on the original label space, or the coarse-grained emotion categories. Expanding the label space with fine-grained concepts was not studied due to two limitations: (1) there was no way to find the connections between words and emotion concepts, and (2) there was not a general mapping that identifies the relationship between emotion categories and fine-grained emotion concepts. The first limitation was addressed in Li, Xie et al. (2021) by the word emotion distribution model, which will be elaborated in Section 3. Furthermore, our prior work (Li, Li, Xie, Li, Tao, 2021) provided a mapping function and extended the label for distribution generation.

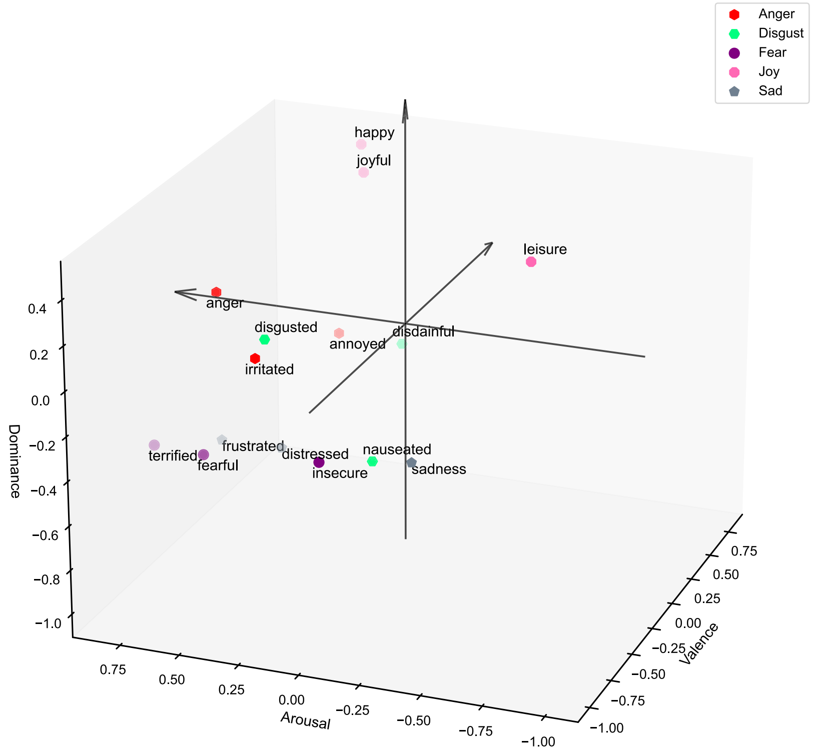


Fig. 2. Visualization of fifteen selected fine-grained emotion concepts (Russell, 1980), grouped by the associated emotion category. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3. Preliminary

To facilitate the proposed method, we need to create the extended label space and generate the emotion distribution. No previous work has ever addressed the extension of the emotion label space, and no such tool or resource is available for us to complete these crucial steps. Nevertheless, our previous research on emotion models makes this task achievable. To deliver the full picture of what we intend to do, this section briefly introduces the adopted domain knowledge and the method for generating the word-level emotion distribution. The details can be found in our previous work (Li, Xie et al., 2021).

3.1. Domain knowledge

Russell and Mehrabian (1977) established an emotion complex with three perpendicular dimensions in their three-factor theory. The dimensions that used to measure emotions are valence, arousal, and dominance. In this theory, 151 fine-grained emotion concepts were identified, and the mean and variance of VAD factors characterized. We adopt the three-factor theory as the *Knowledge of Emotions* (KoE). Fifteen concepts, together with the respective emotion categories that identified by our mapping function are listed in Table 1. To deliver an overall picture of three-factor theory, we visualize these concepts in three-dimensional space, as shown in Fig. 2.

3.2. Lexicon dictionary

We adopt a dimensional lexicon in this work. The NRC-VAD lexicon (Mohammad, 2018) provides a VAD tuple for more than 20,000 English words. Li, Xie et al. (2021) confirmed the consistency between three-factor theory and NRC-VAD by scattering the tuples of emotions and words in the same space. Therefore, we leverage the NRC-VAD lexicon as the *Knowledge of Words* (KoW).

Furthermore, in Table 1, we compare the information of emotion concepts from three-factor theory and the VAD tuples queried from NRC-VAD with the respective *word*, which could be helpful for us to understand the relation between the domain knowledge and the lexicon knowledge. By comparing the numbers, we notice that the two sets of values do not perfectly match. Essentially, these terms have distinguished meanings in three-factor theory and the NRC-VAD lexicon. In three-factor theory, they are characterized by a sphere with mean and variance in the space and represent the general definition of the emotion from a higher level. In contrast, like other words, the tuples from lexicon exclusively represent a semantic term, which are considered dots in the space. More specifically, the concept of *disgusted* tries to cover as many terms as possible that refers to disgusted feeling, such as “blood”, “vomit”, and

Table 1
Fifteen fine-grained emotion concepts (Russell, 1980) and the associated emotion category. Meanwhile, we also present the VAD tuple of the corresponding *word* in NRC-VAD.

Concept	Label	Three-factor theory						NRC-VAD		
		Valence		Arousal		Dominance		Valence	Arousal	Dominance
		Mean	SD	Mean	SD	Mean	SD			
happy	Joy	.81	.21	.51	.26	.46	.38	1.000	0.470	0.544
hoyful		.76	.22	.48	.26	.35	.31	0.980	0.480	0.334
Leisure		.58	.35	−.32	.33	.11	.33	0.262	−0.142	−0.158
anger	Anger	−.51	.20	.59	.33	.25	.39	−0.666	0.730	0.314
irritated		−.58	.16	.40	.37	.01	.40	−0.580	0.632	−0.438
annoyed		−.28	.16	.17	.28	.04	.31	−0.792	0.566	−0.444
fearful	Fear	−.64	.20	.60	.32	−.43	.30	−0.834	0.680	−0.444
terrified		−.62	.20	.82	.25	−.43	.34	−0.820	0.804	−0.226
insecure		−.57	.34	.14	.42	−.42	.29	−0.772	0.076	−0.736
sadness	Sad	−.63	.23	−.27	.34	−.33	.22	−0.896	−0.424	−0.672
frustrated		−.64	.18	.52	.37	−.35	.30	−0.840	0.302	−0.490
distressed		−.61	.17	.28	.46	−.36	.21	−0.714	0.542	−0.352
disgusted	Disgust	−.60	.20	.35	.41	.11	.34	−0.898	0.546	−0.452
nauseated		−.61	.25	−.01	.28	−.36	.33	−0.876	0.320	−0.648
disdainful		−.32	.32	−.11	.27	.05	.33	−0.806	−0.040	−0.474

"extrusion". In contrast, the word "disgusted" is pinpointed to its semantic meaning, i.e., feeling or expressing revulsion or strong disapproval, which is different from "disgusting" referring to arousing revulsion or strong indignation. Therefore, we used the information from three-factor theory, in lieu of a lexicon dictionary, as the knowledge of emotions.

3.3. Word-level emotion distribution

We generate the word-level emotion distribution (WED) (Li, Xie et al., 2021) leveraging the domain theory and lexicon knowledge. Given K emotions, $\{E_1, E_2, \dots, E_K\}$, we calculate a word's soft assignment to each emotion, which can be regarded as the intensity of the emotion. Concretely, given a pair of emotion E_k and word w, we retrieve the VAD mean $\boldsymbol{\mu}_{E_k} = [V^m_{E_k}, A^m_{E_k}, D^m_{E_k}]$ and standard deviation $\boldsymbol{\sigma}_{E_k} = [V^{sd}_{E_k}, A^{sd}_{E_k}, D^{sd}_{E_k}]$ from the three-factor theory and the VAD tuple $\mathbf{w} = [V^w, A^w, D^w]$ if w appears in NRC-VAD. The intensity of the emotion E_k, $\mathbf{d}^{E_k}_w$ is modeled as the probability measurement in a multivariate Gaussian distribution, considering VADs as independent variables:

$$\mathbf{d}^{E_k}_w = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{\mu}_{E_k} - \mathbf{w})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_{E_k} - \mathbf{w})\right)}{\sqrt{(2\pi)^3 |\boldsymbol{\Sigma}|}}, \tag{1}$$

where $\boldsymbol{\Sigma} = diag(\boldsymbol{\sigma}_{E_k})$. For the K emotions, the WED of word w is $\mathbf{d}_w = [d^{E_1}_w, d^{E_2}_w, \dots, d^{E_K}_w]$. To this end, we can generate a distribution representation by measuring the distance with a set of emotions in the three-dimensional Gaussian space.

4. Methodology

The generic flowchart of the framework is depicted in Fig. 3. This section introduces how to create the mapping function to expand the label space and generate the pseudo sentence emotion distribution.

4.1. Mapping fine-grained concepts with the emotion category

The significance of using three-factor theory as KoE has been discussed. However, no available resource can help us build the relationship between fine-grained concepts and emotion categories, which is the major challenge that we have to overcome. Therefore, we have to find a feasible approach based on the available knowledge, such as lexicons and dictionaries. This section will elaborate on how we create the mappings between emotion categories and fine-grained concepts. The first step is preprocessing, which aims to remove the compounded concepts and those linked with mixed emotions. Then, manual deliberation is incorporated. Emotion itself is very subjective and fuzzy, and subjective manual deliberation must be implemented to comprehensively build the mapping. We invite volunteers from different backgrounds to annotate each concept based on their common-sense knowledge. The annotation results from volunteers are consolidated to produce the final mapping.

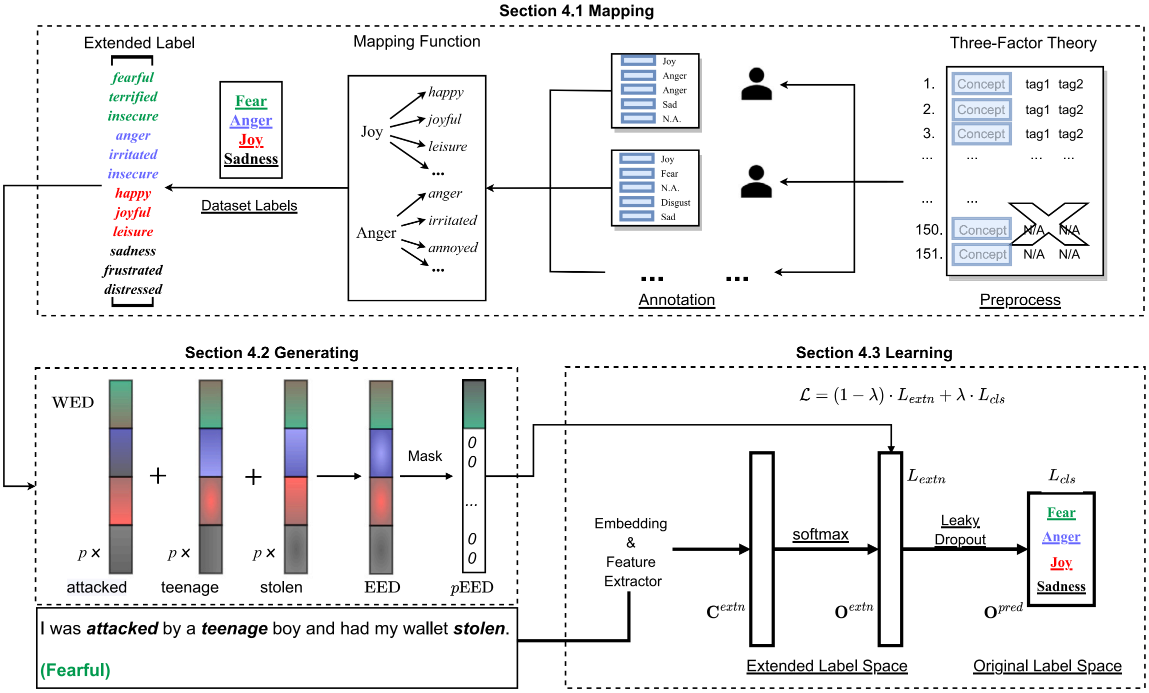


Fig. 3. Generic framework. Compared with Li, Xie et al. (2021), a Leaky Dropout is applied at the penultimate layer. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.1.1. *Preprocessing of fine-grained concepts*

A total of 151 emotion concepts were identified in the KoE. However, not all of these concepts are suitable in our task. The selected concepts are expected to be representative of an emotion category and should be widely recognized in the contemporary language environment. In the preprocessing stage, we aim to narrow the manual annotation scope by filtering out compounded emotion concepts and trivial ones. Eleven compounded concepts, such as *snobbish_and_lonely*, *hostile_but_controlled*, and *angry_but_detached* are removed as they are too delicate to precisely characterize. Besides, language is constantly changing over time, and the way and frequency that we use some words also shift. Three-factor theory was proposed in 1977, so some of the concepts may no longer be tangible in the era we live. We turn to recent emotion lexicons, such as SenticNet 6 and EmoLex, to filter out potentially outdated concepts. In the end, we retained 132 out of 151 concepts for further deliberation.

4.1.2. *Creating the mapping function*

We intend to create a mapping function between fine-grained emotion concepts and emotion category:

$$\mathbf{Map}(concept) \rightarrow \text{Category}. \tag{2}$$

This step invokes the emotion lexicons, i.e., SenticNet 6 and EmoLex, to facilitate the manual deliberation of mapping. However, these lexicons are not regarded as the gold standard, as lexicon knowledge pinpoints a specific word, rather than referring to the concept, similar to what we have argued in Section 3.2.

Another major setback of lexicon-based method is the inconsistency between dataset labels and lexicon annotations. For example, neither SenticNet 6 nor EmoLex is fully compatible with the datasets used in this work. Special treatment is needed for the emotion categories that are rarely used in lexicons, such as “Guilty” and “Shame”. Moreover, the accuracy of large-scale annotations in both SenticNet 6 and EmoLex cannot be guaranteed. SenticNet was automatically constructed by using deep learning methods, and EmoLex was built by crowdsourcing on Mechanical Turk. Therefore, it is necessary to include manual deliberation in the mapping construction.

Manual deliberation is incorporated to guarantee that the mapping can be widely acknowledged and recognized. Five volunteers (or annotators), two males and three females, from different linguistic environments and cultural backgrounds, were invited to complete the manual annotation process. They speak different languages: two native speakers of Mandarin, two native speakers of Cantonese, and one native speaker of French and English. The volunteers received English-medium education on diverse subjects in top universities of Hong Kong. They all have a master’s degree or above. For each concept, the annotators were provided with the tags retrieved from both SenticNet 6 and EmoLex for their reference. A general guideline was provided to the volunteers to facilitate their annotation:

Table 2
Our identified mapping function between emotion categories and emotion concepts.

Coarse-grained emotion label	Dataset(s)	Fine-grained emotion concepts
"Anger"	ISEAR + TEC + SemEval	*enraged*, *irritated*, *anger*, *insolent*, *annoyed*
"Disgust"	ISEAR + TEC + SemEval	*scornful*, *disdainful*, *nauseated*, *disgusted*, *arrogant*
"Fear"	ISEAR + TEC + SemEval	*awed*, *confused*, *terrified*, *domineering*, *insecure*, *fearful*, *aggressive*
"Joy"	ISEAR + TEC + SemEval	*vigorous*, *joyful*, *enjoyment*, *fascinated*, *lucky*, *leisurely*, *reserved*, *dignified*
"Sadness"[a]	ISEAR + TEC + SemEval	*helpless*, *burdened*, *anguished*, *upset*, *sad*, *defeated*
"Guilt"	ISEAR	*sinful*, *regretful*, *guilty*, *selfish*[b]
"Shame"	ISEAR	*shamed*, *humiliated*, *embarrassed*[b]
"Surprise"	TEC + SemEval	*wonder*[b], *surprised*, *astonished*, *curious*
"Anticipation"	SemEval	*activated*, *hopeful*, *anxious*[b]
"Love"	SemEval	*loved*, *in_love*[b], *affectionate*[b]
"Optimism"	SemEval	*strong*, *bold*, *free*[b], *grateful*, *appreciative*
"Pessimism"[a]	SemEval	*depressed*, *distressed*, *frustrated*
"Trust"	SemEval	*secure*, *responsible*, *friendly*

[a] The fine-grained concepts under "Sadness" and "Pessimism" can be combined when a dataset only has the emotion label "Sadness".

[b] The mappings added by the final review. These mappings can be agreed upon by at least one annotator but cannot be agreed upon by all.

1. They are allowed to exploit any additional resources, such as other available lexicons, authoritative dictionaries, and related examples, to understand the semantic meaning and common usage in daily life. From the literal perspective, the words of emotion concepts have very similar meanings. Extra information can be helpful to distinguish concepts with delicate differences.
2. The most appropriate emotion category from the emotion tags provided by the lexicons is determined. It is common that one word has several tags from the lexicons. For example, the word "depressed" is associated with "Anger", "Fear", and "Sadness" in EmoLex. However, when referring to a kind of emotion, the most relevant emotion should be "Sadness".
3. We asked the annotators to reference the provided lexicon tags but not stick with them. The annotators are educated on the difference between affective "words" and emotion *concepts*. They can overturn the given tags based on their personal understanding if necessary and name the one that they think is more suitable. To avoid excessive subjectivity, they need to justify the changes with evidence.
4. We find it particularly difficult to find relevant fine-grained concepts for a number of emotion categories. "Shame" and "Guilty" are not used in lexicon annotation. "Surprise" and "Disgust" cause some debates among annotators on what concepts should be included and what should be excluded, although both are used as tags in the lexicons. Special attention is suggested for aforementioned emotion categories.

On average, the annotators successfully found the associated emotion category for 69.4 out of 132 concepts. The annotators show some disagreements on several concepts, while we hope the final mapping function can be agreed upon all. We discarded the concepts whose associated category cannot be unanimously agreed upon by the annotators. For example, the concept *lonely* was excluded, as it received a three to two votes on "Fear" and "Sadness". Nevertheless, emotion categories such as "Surprise", "Shame", and "Guilty", have relatively limited associated concepts if the all-agree criterion is enforced, which will make the mapping severely unbalanced. Therefore, the results were screened by the four investigators again to finalize the mapping with careful adjustments. The adjustments are adding *wonder* to "Surprise", *in_love* and *affectionate* to "Love", and *free* to "Optimism". In addition, the concepts in "Guilt" and "Shame" are tuned to be balanced. In the end, the final mapping was managed to achieve an overall agreement of 76.3% between all annotators. We present our identified mapping function in Table 2.

We want to emphasize that the work itself is not targeting at rigorously defining a hierarchical emotion system from a professional psychological perspective, given our limited domain expertise. Instead, we find a feasible way in the field of sentic computing to create the mapping and apply the WED technique for distribution learning based on available resources and manual deliberation. The proposed methodology may not be the best one, and we are humble and open to any suggestion. Moreover, the domain knowledge and lexicons are open-sourced information, which can be easily accessed from the corresponding stakeholder. Thus, one can easily reproduce the mapping and perform the selection according to their knowledge, and even redefine a new mapping function for their necessity.

4.1.3. Fine-grained emotion selection

Based on the identified mapping, we can now extend the label space for a dataset. In general, we can tell from Table 2 that, although the mapping has been consolidated, the overall mapping is not balanced per category. However, we intend to extend each emotion category with the same number of fine-grained concepts so that the dimension of each label can be equally expanded. In this work, for instance, we select three concepts per emotion category.

Regarding the question of how to select the concepts, at this moment, we think there is no gold standard for which one to choose and which one not to choose because all the associations are designed to be universally accepted. The selection is mainly based on people's understanding or experience with the emotion categories and concepts. A workable solution is to select those that are both representative and relatively scattered in the VAD space. As shown in Table 1, the selected concepts under each category show differences in each dimension. The purpose is to ensure that the distribution of each emotion category can encode diverse information. If all the selected concepts were pinpointed in a small area in the VAD space, their intensities in WED computed by Eq. (1) and values in the final distribution will be similar, which means we merely repeat the labels without enriching the information. As the inconsistent annotations were handled strictly, we find at most three concepts for some emotion categories, which restricts the application of machine learning-based selections. We believe it is a promising research direction to include concepts linked with several emotion categories. By doing so, the label space can be further extended, and we will have flexibility to incorporate automatic or semiautomatic methods in the selection process.

4.2. Generating the sentence-level emotion distribution

Given a sentence, we generate a pseudo distribution according to the extensions confirmed in the previous step. Based on the WED method, we produce a WED vector in the extended label space. Next, the rule-based method in Li, Li, Xie, Li, Tao (2021) is used to produce the pseudo extended emotion distribution (pEED), which is a simplified version from Li, Xie et al. (2021). Interested audiences can find the algorithm in both published papers. The proposed rule-based method increments the emotion distribution of each affective word to obtain a sentence-level distribution label, i.e., the generated extended emotion distribution (EED), as shown in the generating part in Fig. 3. The polarity value from SenticNet is used as a weight for each word.

In particular, we use a masking function to remove the effects of non-dominant emotions in the learning stage. Given the ground-truth emotion category $\mathbf{y}$, for emotion concepts in which $\mathbf{Map}(concept) \neq \mathbf{y}$, their values in the EED are masked to 0. Moreover, the masked distribution is normalized to form the distribution label pEED. The pseudo distribution label will be used in the classification framework.

4.3. Learning the classifier with the extended distribution

This work devises a framework to incorporate distribution learning into the classification framework. Distribution learning has been exploited in some existing works at the output layer, where the distribution learning loss works as an auxiliary term in the objective function. In contrast, we adopt a pipeline framework and apply distribution learning and classification learning at different layers.

4.3.1. Semantic feature extraction

We employ a feature extractor layer, such as CNNs, BiLSTM, and BERT, to extract semantic features from textual input. Given a sentence s with m tokens, non-BERT implementations project discrete word tokens into a continuous embedding space via the randomly initialized word embedding model and obtain a dense vector $\mathbf{x}_i \in \mathbb{R}^k$ for each word, where k is the embedding dimension. Given a sentence with m words, all embedding vectors are concatenated as the model input: $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m]$. The sentences are padded to maintain a uniform length. We first apply a CNN layer (with a pooling layer) or a recurrent layer to extract feature vector $\mathbf{C}$:

$$\begin{aligned} &\mathbf{C} = \mathrm{TextCNN}(\mathbf{x}), \text{or} \\ &\mathbf{C} = \mathrm{BiLSTM}(\mathbf{x}). \end{aligned} \tag{3}$$

For the BERT model, we extract the *[CLS]* token from the output layer of the pre-trained BERT model as the sentence representation:

$$\mathbf{C} = \mathrm{Bert}(s). \tag{4}$$

4.3.2. Learning the extended emotion distribution

The feature vector $\mathbf{C}$ is projected to the extended label space, where distribution learning is employed. A softmax layer is employed to transfer the resultant vector into a probability distribution of the extended label $\mathbf{O}^{extn}$:

$$\begin{aligned} &\mathbf{C}^{extn} = \mathrm{FullyConnect}(\mathbf{C}) \\ &\mathbf{O}^{extn} = Softmax(\mathbf{C}^{extn}). \end{aligned} \tag{5}$$

We apply the pEED at this layer and calculate the Kullback–Leibler loss L_{extn} for distribution learning,

$$L_{extn}(\mathbf{O}^{extn}, \mathrm{pEED}) = -\frac{1}{N}\left[\sum_{i}^{N}\sum_{j}^{K} \mathrm{pEED}_j \cdot \ln(\mathbf{O}_j^{extn})\right], \tag{6}$$

where N denotes the batch size.

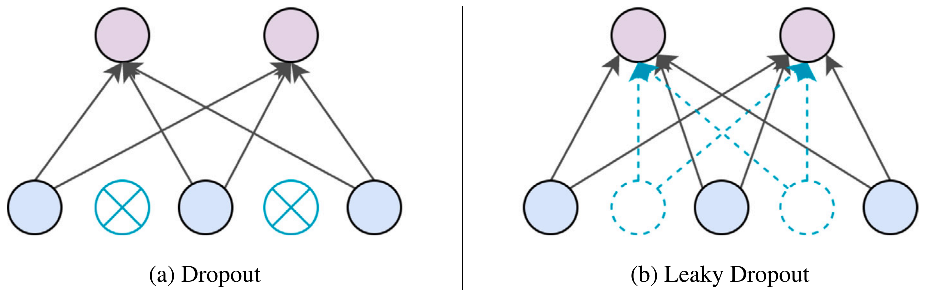


Fig. 4. Difference between conventional Dropout mechanism and the proposed Leaky Dropout.

4.3.3. A sparse connection for prediction

Subsequently, another layer projects the learned distribution into the label space,

$$\begin{aligned}\mathbf{O}^{pred} &= \text{FullyConnect}(\mathbf{O}^{extn})\\ \mathbf{O}^{pred} &= \text{Softmax}(\mathbf{O}^{pred}).\end{aligned} \tag{7}$$

Cross-entropy loss L_{cls} is computed for guiding the classification learning:

$$L_{cls}(\mathbf{O}^{pred}, y) = -\frac{1}{N}\left[\sum_{i}^{N}\sum_{j}^{K*}\mathbf{1}(y_i = j)\cdot\ln(\mathbf{O}_j^{pred})\right], \tag{8}$$

where K^* is the number of labels, and $\mathbf{1}(y_i = j) = 1$ if $y_i = j$; otherwise, $\mathbf{1}(y_i = j) = 0$.

When projecting the extended emotion distribution space back to the original label space, we expect the classifier be able to extract the dependency from the fine-grained concepts. Although the non-dominant emotion concepts are masked in the pEED, the penultimate layer still produces a dense vector. Nevertheless, a special overfitting situation occurs, where the classifier only learns the positional mapping between the entries of the extended distribution with the index of the label. Such a situation does not harm our model, as we have found that using an extended distribution label to train the classifier only and predicting the output from the learned distribution can bring improvements to baseline models. However, the classifier cannot fully exploit the dependency information in the extended space in this way. Therefore, we propose a sparse connection approach, namely, *Leaky Dropout*, to avoid such a situation.

The Dropout mechanism (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014) builds a sparse connection between layers to alleviate the overfitting issue in a deeper network. The sparse connection is performed by randomly deactivating neurons in a layer, as shown in Fig. 4(a), so that the information of these selected neurons will not feed to the next layer. However, deactivating selected neurons causes unexpected effects in our model. More specifically, a smaller dropout rate cannot prevent overfitting between the last two layers (the model shows limited improvements compared with training with the extended label only), while a greater rate can block valid information flow. As a main contribution, we propose a soft Dropout mechanism, namely Leaky Dropout, in this paper. The term "leaky" means that the selected neurons will not be fully deactivated; instead, only the magnitude of their values will be suppressed, and the weights can be updated by gradient backpropagation. The difference between conventional Dropout and the propose Leaky Dropout is presented in Fig. 4(b).

Given the activated layer $\mathbf{x}$ and dropout rate β, we partially suppress the vector by computing elementwise production of $\mathbf{x}$ and a masking vector $\mathbf{m}$:

$$\mathbf{x}' = \mathbf{x} \odot \mathbf{m}, \tag{9}$$

and $\mathbf{m}$ is calculated as:

$$\begin{aligned}\mathbf{m}_i &= \begin{cases}\frac{1}{1-\beta}, & z_i = 0\\ \gamma, & z_i = 1\end{cases}\\ \gamma &= \frac{1-\beta}{c^2},\end{aligned} \tag{10}$$

where $z_i \sim Bernoulli(\beta)$ are indicators (i.e., preserve if $z = 0$ or partially drop if $z = 1$), and the suppression factor c controls the magnitude of suppression. For example, a neuron selected by the Bernoulli distribution with value x will return $\frac{0.8\times x}{10000}$ if $\beta = 0.2$ and $c = 100$. The quadratic denominator helps to reduce the parameter searching space. Following Srivastava et al. (2014), the kept cells are magnified to maintain the layer's expectation *largely* unchanged. An exceedance is expected in the overall expectation as the selected neurons are not fully dropped. However, due to the quadratic denominator, the exceedance is very small and can be neglected. Moreover, the nominator $1 - \beta$ is a harmonic factor that caps the expectation exceedance $\frac{\beta(1-\beta)}{c^2}$ at $\frac{0.25}{c^2}$ when $\beta = 0.5$.

We apply the Leaky Dropout in the last two layers by changing Eq. (7) to

$$\begin{aligned}\mathbf{O}^{pred} &= \text{FullyConnect}(\mathbf{LeakyDropout}(\mathbf{O}^{extn}))\\ \mathbf{O}^{pred} &= \text{Softmax}(\mathbf{O}^{pred}).\end{aligned} \tag{11}$$

Table 3
General statistics of the adopted datasets. "CV" means that the dataset does not have a standard training/testing split, and 10-fold cross validation is applied.

Dataset	Average length	Train	Dev	Test	Emo. #
ISEAR	21.2	7666	–	CV	7
TEC	15.3	21,051	–	CV	6
SemEval	16.0	63,838	886	3259	11
SST-1	18.1	11,855	–	2210	5
SST-2	18.9	9613	–	1821	3

4.3.4. Model training

The loss function $\mathcal{L}$ is a weighted sum of the Kullback–Leibler loss (the distribution learning loss) and cross-entropy loss (the classification loss) based on Eqs. (6) and (8),

$$\mathcal{L} = \lambda \cdot L_{cls} + (1 - \lambda) \cdot L_{extn}. \tag{12}$$

The hyperparameter λ is a weighting term to balance the proportion of distribution learning loss and classification learning loss in the objective function. Intuitively, if $\lambda = 1$, the model is trained with a one-hot label, and there is no constraint in the extended label space. Adam optimizer is employed to train the model.

5. Experiment

5.1. Datasets

We show the effectiveness of the proposed methodology in different tasks. The description of adopted datasets are reported in Table 3.

5.1.1. Multiclass datasets

For multiclass classification task, we adopts ISEAR and TEC. **ISEAR** collects the personal experience when seven emotions were perceived (Scherer & Wallbott, 1994). **TEC** includes six emotions to label the tweets (Mohammad, 2012).

5.1.2. Multilabel datasets

We have also experimented on a multilabel dataset. **SemEval2018** annotates the affectual state from a tweet using automatic systems (Mohammad et al., 2018). The task is to classify a tweet as a 'neural or no emotion' type or an emotional type with the eleven emotions.

5.1.3. Sentiment analysis task

We investigate the effectiveness in sentiment analysis, where binary labels representing sentiment orientations, positive and negative, are used. **SST-1**[4] and **SST-2** contain five sentiment label and binary labels, respectively (Socher et al., 2013). Particularly, we select the same quantity of fine-grained concepts for both positive and negative labels and construct pEED for each sentence. Note that, for this work, we will not specify what emotions are of very positive/negative and what are of positive/negative. The neutral class is also removed in the extended space. For positive sentiment, *joyful*, *fascinated*, *affectionate*, *free*, and *leisurely* are included; for negative sentiment, *anger*, *fearful*, *depressed*, *sad*, and *guilty* are exploited.

5.2. Baselines

The baseline models are selected according to the nature of the task. For multiclass classification datasets, i.e., ISEAR and TEC, we compare the proposed model with the following baseline models:

MTCNN (Zhang et al., 2018) is the first work to incorporate a generated distribution label for distribution learning. A multitask framework is proposed for emotion classification. **DERNN** (Wang, Wang, Xiang, & Xu, 2019) uses an RNN-based framework to exploit topical information and syntactic dependency for emotion label prediction. **TESAN** (Wang & Wang, 2020) designs a topic model and produces a topic embedding for a document, which is used to predict the emotion label. The **DACNN** proposed by Yang and Chen (2020) employs an attention mechanism by adjusting the weights for features extracted from a multichannel CNN to improve emotion classification performance. **WLTM** (Pang et al., 2019) uses a topic model to alleviate the data sparsity issue. **ESTeR** proposed by Gollapalli, Rozenshtein, and Ng (2020) exploits word co-occurrences and associations from a large-scale corpus for unsupervised emotion classification. **EmoChannel** (Li et al., 2020) looks into fine-grained emotions and models their intensity variations in the sentence as a novel feature. The dependency within fine-grained emotions is extracted by an attention mechanism for improving the model performance. **WED** proposed by Li, Xie et al. (2021) presents a method to produce word-level emotion distribution as an affective representation. Two schemas are suggested to utilize WED for emotion classification.

[4] http://nlp.stanford.edu/sentiment/.

Table 4
Parameter settings used in this research.

Model	Parameters
CNN module	*Filter size* $= [3, 4, 5]$, 100 filters per size
RNN module	*Hidden dimension* $= 128$
Transformer	*Multi-head* $= 8$ and 3 *blocks*

For multilabel dataset, SemEval2018, we compare TextCNN, Bi-LSTM, and C-LSTM under the binary relevance (BR) framework and joint binary (JB) framework proposed by He and Xia (2018) to examine the feasibility of our proposed method in multilabel classification.

5.3. Evaluation metrics

We adopt the **accuracy** and **F1 score** to evaluate the single-label prediction performance. We conduct a **t-test** against the reproduced baselines. For datasets without a training/testing split, such as ISEAR and TEC, we conduct tenfold cross-validation and report the average result with a t-test based on five rounds. As to SST-1/2 where standard train/test is provided, the reported results are based on ten rounds of experiment.

For multilabel classification, we employ **hamming loss** (HL), **ranking loss** (RL), **micro F1** (miF1), **macro F1** (maF1) and **average precision** (AP) to measure the performance.

5.4. Experiment configuration and parameter settings

For non-BERT implementations, we set the word embedding size to 300 following the conventional setting in NLP tasks. The embedding vectors are initialized from scratch and set to trainable during training to remove the performance margin caused by different pre-trained language models.

Moreover, this work focuses on the feasibility and effectiveness of the label extension schema and the sparse connection mechanism. Thus, we implement the feature extractors with empirical configurations and keep them fixed for ablation studies. The hyperparameter settings are specified in Table 4.

We employ BERT Uncased for BERT implementation. The BERT model is frozen during the training to reduce the training time. The Adam optimizer is employed to train the classifier. The rate for conventional Dropout connecting feature extractor layers is set to 0.5 for all models. As to the dropout applied to the extended label space, according to the cross-validation on ISEAR, the conventional Dropout rate is set as 0.2, and the Leaky Dropout rate is set to 0.2 with the leaky parameter c set to 200.

The number of fine-grained concepts for each category in the extended label space is *three*. We can only identify three associated fine-trained concepts for some labels. For example, the ISEAR dataset has seven emotion labels, so the extended label space contains 21 fine-grained concepts.

5.5. Experimental results

We report the results of our model against baseline methods in Table 5 (multiclass classification), Table 6 (sentiment analysis), and Table 7 (multilabel classification). In all tasks, consistent improvements are observed in our model compared with the baseline methods. The t-test results indicated significant improvements. Particularly, the proposed method sees improvements of 2.5% in accuracy and 2.82% in the F1 score on ISEAR, and 2.52% in accuracy and 2.0% in the F1 score on TEC, compared with TextCNN. Compared with BiLSTM, the proposed framework exploiting BiLSTM yields 2.24% and 2.43% improvements in accuracy and F1 score on ISEAR, and 1.16% and 1.64% improvements in accuracy and F1 score on TEC.

Regarding the BERT-based method, our method produces $1 \sim 1.4\%$ improvements over the BERT baseline on ISEAR and TEC. Furthermore, our model outperforms other methods on sentiment analysis datasets, i.e., SST-1 and SST-2, and the multilabel classification dataset, SemEval2018, which validates our approach's generality across different emotion-relevant tasks.

Ablation studies are conducted to help us understand the function of each model component.[5] The following ablation models are tested on the ISEAR and TEC datasets.

- **Ablation (a):** We train the classifier with the extended distribution label only, and the model predicts from the learned extended distribution. In this ablation model, we remove cross-entropy loss in Eq. (12) (set $\lambda = 0$) and make the penultimate layer the output layer. The classifier searches for the maximum value in the extended label space and directly returns its corresponding emotion category as the output.
- **Ablation (b):** The model does not have a sparse connection, where a dense layer connects the penultimate layer and the output layer, following Eq. (5).
- **Ablation (c):** The model utilizes a conventional Dropout layer as the sparse connection, and the dropout rate is set as 0.2, which is inconsistent with the Leaky Dropout configuration.

[5] This work mainly focuses on the parameters related to Leaky Dropout. The influence of the parameters of WED on model performance was studied in Li, Xie et al. (2021).

Table 5
Results on the multiclass classification task.

Model		ISEAR		TEC	
		Accu.	F1	Accu.	F1
(retrieved results from references)					
DERNN	('19)	–	60.44	–	–
WLTM	('19)	36.50	–	–	–
TESAN	('20)	61.14	–	–	–
DACNN	('20)	–	–	62.73	–
ESTeR	('20)	–	–	–	39.8
(reproduced results)					
CLSTM	('15)	59.88	59.47	59.84	51.03
Transformer	('18)	61.07	60.20	62.17	53.74
MTCNN	('18)	61.15	60.72	62.32	53.82
EmoChannel	('20)	61.87	61.21	62.16	54.13
WED	('21)	62.19	61.24	62.58	54.18
BiLSTM	('13)	59.68	59.42	61.46	54.19
Label extension schema w/BiLSTM					
Predict from extended label space					
(a) w/learned distribution		60.16	61.18	61.62	54.61
Predict from original label space					
(b) w/dense connection		59.83	58.82	61.75	55.02
(c) w/dropout (rate = 0.2)		59.96	59.16	61.92	55.10
(Ours) w/leaky dropout (rate = 0.2)		61.92	61.85	62.62	**55.83**‡
TextCNN	('14)	60.53	60.03	61.00	53.10
Label extension schema w/TextCNN					
Predict from extended label space					
(a) w/learned distribution		61.60	61.52	61.32	53.10
Predict from original label space					
(b) w/dense connection		61.83	61.74	61.35	53.01
(c) w/dropout (rate = 0.2)		62.53	62.23	62.75	53.92
(Ours) w/leaky dropout (rate = 0.2)		**63.03**§	**62.85**‡	**63.52**†	55.10
Bert	('19)	65.33	65.35	64.37	55.58
Label extension schema w/Bert					
Predict from extended label space					
(a) w/learned distribution		65.60	65.58	64.73	55.83
Predict from original label space					
(b) w/dense connection		65.87	65.64	64.81	55.88
(c) w/dropout (rate = 0.2)		66.38	65.93	65.13	56.19
(Ours) w/leaky dropout (rate = 0.2)		**66.73**§	**66.39**‡	**65.56**†	**56.42**†

$^{\dagger}p < .05$, $^{\ddagger}p < .01$, $^{\S}p < .001$.

Table 6
Results on the sentiment analysis task.

Model		SST-1		SST-2	
		Accu.	F1	Accu.	F1
TextCNN	('14)	45.13	41.52	83.82	83.80
BiLSTM	('13)	44.68	42.30	83.57	83.53
CLSTM	('15)	45.83	43.15	82.45	82.33
Transformer	('18)	43.11	40.03	82.68	82.66
Ours		**46.74**‡	**44.12**†	**85.10**§	**84.27**‡
Bert	('19)	53.20	51.00	91.20	91.20
Bert + Ours		**54.52**‡	**53.12**§	**92.48**‡	**92.51**‡

$^{\dagger}p < .05$, $^{\ddagger}p < .01$, $^{\S}p < .001$.

6. Discussion

6.1. Discussion of the ablation study

In Table 5, Ablation (a) outperforms the model trained with one-hot labels, suggesting that the extended distribution is better than a one-hot label. It is an intriguing observation, as intuitively, it is more challenging for the model to learn given a three-folded label space. We give credit to the benefit of distribution learning. The smoothed distribution label accompanied by the Kullback–Leibler loss provides essential knowledge to the classifier. Second, the refinement from extended label space to emotion labels without a Dropout layer, i.e., Ablation (b), performs less impressively compared with Ablation (c). The overfitting issue may occur earlier

Table 7
Results on the multilabel classification task.

Model		AP ↑	MaF1 ↑	MiF1 ↑	HL ↓	RL ↓
BR	TextCNN	46.79	48.37	58.86	17.7	18.2
	BiLSTM	45.90	47.83	56.02	19.7	19.6
	CLSTM	47.37	48.68	59.30	16.4	17.3
Ours (BR)		**50.47**‡	**50.95**‡	**62.10**‡	15.8	16.1
JB	TextCNN	49.22	49.76	61.24	16.1	17.3
	BiLSTM	48.38	47.15	59.88	17.1	17.5
	CLSTM	50.12	50.85	62.10	16.1	16.5
Ours (JB)		**51.92**‡	**53.25**‡	**64.28**‡	16.2	15.7

↑ means "the higher the better"; ↓ means "the lower the better"
$^{\dagger}p < .05$, $^{\ddagger}p < .01$, $^{\S}p < .001$.

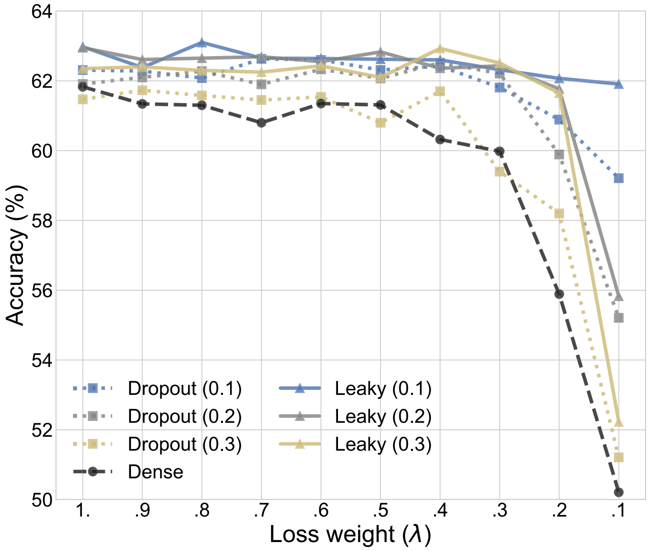


Fig. 5. Results of using different dropout rates (β) with different loss weights (λ). The effect of different dropout mechanisms is discussed in Section 6.2, and the influence of loss weight is discussed in Section 6.3. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

with a dense connection, preventing a better performance, as we discussed in Section 4.3.3. Third, Ablation (c) with a conventional Dropout connection produces satisfactory results, which validates the necessity of a sparse connection, as argued by Li, Li, Xie, Li, Tao (2021). Moreover, we notice that the proposed model can lead to substantial performance improvement against Ablation (c), revealing that a leaky connection is more suitable than neuron deactivation in the emotion label extension schema.

6.2. Sparse connection and Leaky Dropout

Extensive experiments are conducted to investigate how the Dropout mechanism functions in the overall performance. In Fig. 5, we plot the results with different connection mechanisms on the ISEAR dataset, which are dense connection (plotted in black dashed line), Dropout (in dotted line), and Leaky Dropout (in solid line) with various configurations. The contribution of using a sparse connection is distinctly evident since models with Dropout or Leaky Dropout produce significant improvements compared with the model with a dense connection. Notably, with the same dropout rate, the model with Leaky Dropout consistently outperforms the model with conventional Dropout, which validates the effectiveness of Leaky Dropout. Furthermore, we notice that the performance of both traditional Dropout and Leaky Dropout deteriorates as the dropout rate increases. When the dropout rate is larger than $1/K^*$, where K^* is the number of emotion labels, fine-grained emotions associated with the label emotion can be possibly masked, which causes fatal information loss in the learning and thus compromises performance. For example, given seven emotion labels in the ISEAR dataset, we expand each label with three concepts and obtain a 21-dimensional extended space. If the dropout rate is greater than 0.143, or $\frac{3}{21}$, it is possible that all the neurons denoting concepts under the same category can be deactivated at the same time. In such a case, the valid information about the distribution will be fully discarded, and the training can collapse. Nevertheless, the magnitude of performance deduction of using Leaky Dropout is less significant than that of using Dropout, proving that the proposed Leaky Dropout is more robust than conventional Dropout.

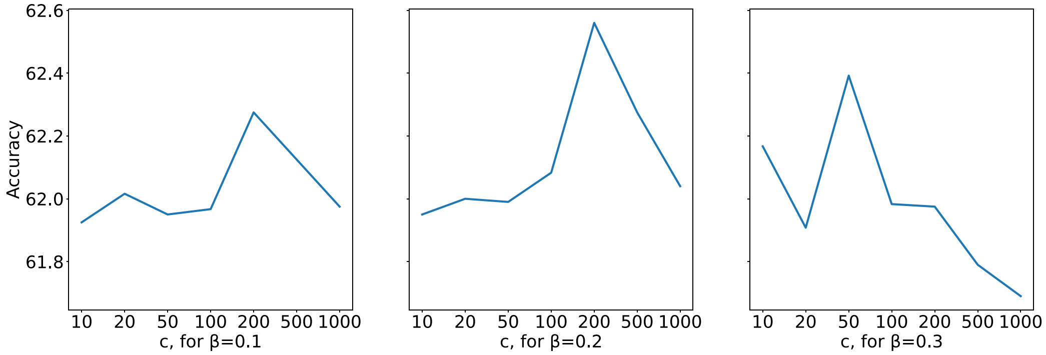


Fig. 6. Relationship between model performance and suppression parameter c with different dropout rates β.

In addition to the dropout rate, we also investigate the effect of the suppression factor c on the overall performance. In Fig. 6, we visualize how the predication accuracy changes with different Leaky Dropout configurations. Generally, the best performance is achieved when $\beta = 0.2$ and $c = 200$. For $\beta = 0.1$ and $\beta = 0.2$, the performance increases as c increases until 200. When $c \geq 200$, the performance decreases. For $\beta = 0.3$, the performance deteriorates severely with a relatively large suppression rate, e.g., 100. We conclude from the observations that there is an optimal combination between c and β. The model is more sensitive to the change of c when β is larger.

6.3. Model performance with loss weight

To adapt the proportion of Kullback–Leibler loss and cross-entropy loss in the objective function, we set a hyperparameter λ as the loss weight. Essentially, the model will focus more on classification learning if λ is large, and *vice versa*. We train the models with different loss weights under different Dropout settings and visualize the results in Fig. 5. Since the goal of the task is still single-label prediction, it is natural that the model with a relatively large weight for cross-entropy loss, say $\lambda = 0.9$, outperforms the model with a small weight, say $\lambda = 0.1$. A similar observation is also found in Li, Xie et al. (2021) and Zhang et al. (2018). However, the model reacts differently than these two works when λ is not that prominent. A good result can still be achieved when λ has an intermediate value, from 0.3 to 0.8. Sometimes the model with a smaller λ is observed to outperform that with a larger λ. It is also intuitive that the classification performance deteriorates when λ approaches 0 as the model is trained to fit the distribution, with limited guidance for label prediction. However, the model with Leaky Dropout can still yield a very competitive result when $\lambda = 0.1$. In the proposed pipeline framework, the model learns the distribution first and predicts from a different layer based on the distribution so that the constraints at each layer will not affect each other. If the model learns the distribution well, it is easy for the model to learn the refinement with weak supervision in classification learning. Nevertheless, in Li, Xie et al. (2021) and Zhang et al. (2018), distribution learning and classification learning are employed at the same layer. Such a setting is harmful, as learning the distribution and predicting the one-hot label are two tasks that conflict in nature. Consequently, the distribution learning loss becomes a better label smoothing method, and the power of distribution learning never unfolds.

We compare our method with Li, Xie et al. (2021) and Zhang et al. (2018) to provide some in-depth discussions on employing a distribution learning module for classification. As mentioned in Section 1, training with an emotion distribution in the original label space can cause interclass confusion and noise introduction. Consequently, distribution and classification learning must be decoupled to stabilize the performance. Therefore, the existing methods (Guo et al., 2021; Li, Xie et al., 2021; Zhang et al., 2018) assign a dominant value for the cross-entropy loss and set a dominant weight for the ground-truth label in the distribution. In contrast, our proposed label extension approach transfers interclass confusion to intraclass confusion and eliminates noise from non-dominant emotions. Moreover, the progressive pipeline design also highlights the role of distribution learning in the overall framework. Hence, two losses can be coupled to enjoy mutual benefits.

6.4. In-depth discussion of the effect of a sparse connection

Extensive investigation has been made to explain why a sparse connection is helpful to the model training. We intend to examine the pattern of weights in the fully-connected layer, which projects extended label space back to the original space with different dropout rates. We train the classifier with different dropout configurations on the ISEAR dataset to achieve this. The model checkpoints are saved at each training epoch. The weights in the matrix connecting the penultimate and output layers are retrieved and visualized as heatmaps. The visualizations at 5 to 10th and 15th epochs are depicted in Fig. 7, where the columns from left to right are the weights of models with a dense connection, Leaky Dropout ($\beta = 0.1$), Leaky Dropout ($\beta = 0.2$), and Leaky Dropout ($\beta = 0.3$), respectively. Note that it is not an attention layer, and the weight matrix is not a square. Each row of the matrix represents a weighted vector applied in the extended distribution to calculate the probability of each emotion label via inner product. A higher weight (more blue in color) means a higher contribution of the neuron in the extended layer toward a specific label. From the

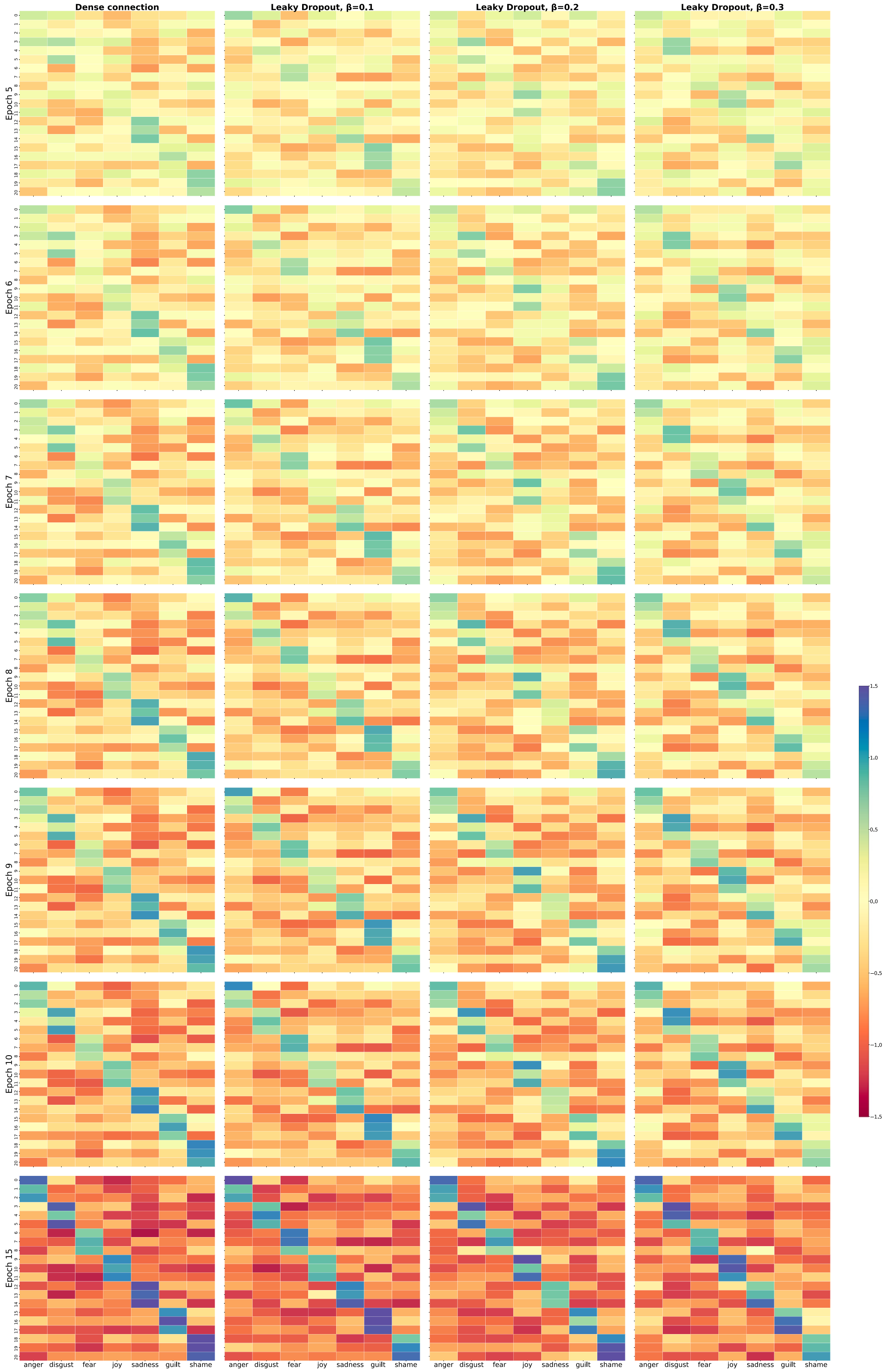
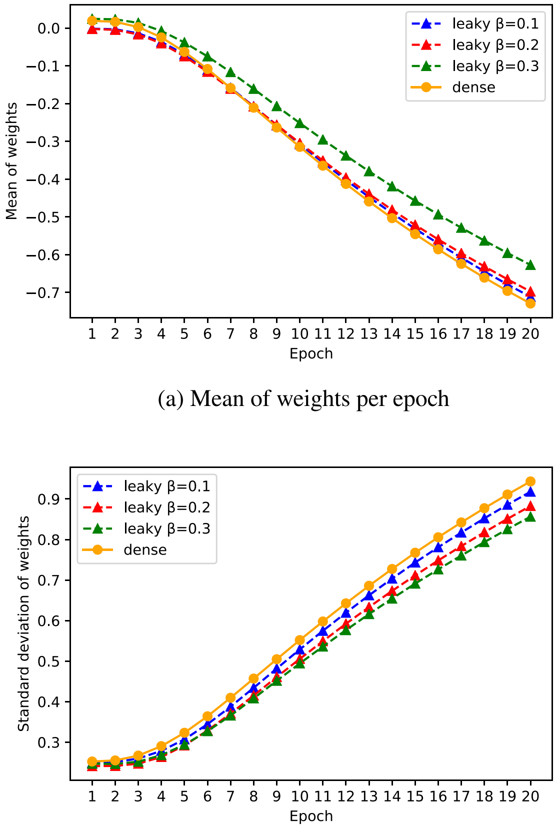


Fig. 7. Visualization of the weights in the output layer. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) Mean of weights per epoch

(b) Variance of weights per epoch

Fig. 8. Mean and variance of weights during the training process. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

visualizations, we observe that as training continues, the connection between fine-grained emotion concepts and the corresponding emotion category identified in the mapping is more evident, showing a distinct diagonal in dark blue at the 15th epoch. However, all these models achieve the best performance at approximately the 8th epoch, suggesting that the dependencies within all fine-grained concepts benefit the classifier.

Intriguingly, we have a visual *feeling* that the diagonal of weight matrix without dropout is better than that with Leaky Dropout at each epoch. We hypothesize that the model with a dense connection tends to recognize the mapping between extended label space and original space earlier than those with a sparse connection and highlight the corresponding mapping in the subsequent training, which signals earlier overfitting. From the scalability perspective, inspired by the normalization technique in deep learning that scalable weights can stabilize the training process, we quantitatively measure such an effect by comparing the mean and variance of the weights, as shown in Fig. 8. The Leaky Dropout is helpful to keep the weights scalable, as the variance of weights is consistently lower than the model without a sparse connection, which benefits the overall performance.

7. Conclusion and future work

In this work, we have established a mapping relationship between emotion categories (or emotion labels) and fine-grained concepts by incorporating domain knowledge and manual deliberation. A novel label extension schema has been proposed to extend the label space of a given dataset based on the identified mapping. We adopted a rule-based method to generate sentence emotion distribution using a general affective representation method, i.e., the Word Emotion Distribution. Additionally, we suggest a classification framework to incorporate distribution learning in the extended label space. We have conducted experiments on various tasks to demonstrate that the proposed method is feasible and effective. Our proposed method can give the distribution learning module much a higher weight than the existing methods, as the proposed framework can handle the interclass confusion and noise introduction issues of incorporating distribution learning in a classification task.

Moreover, we identified the problem of using a dense connection to project the extended label space back to the original label space. We proposed a novel space connection, i.e., Leaky Dropout, which represses the neuron values instead of fully deactivating neurons. Applying the Leaky Dropout in the proposed pipeline framework produces substantial improvement, as a better refinement is yielded when making predictions based on the extended distribution.

Our future work will focus on refining the mapping function. We excluded the fine-grained concepts expressing multiple emotions in this work for constructing an explicit mapping. However, concepts associated with mixed emotions may provide more information to the model. We will devise a better extension method to leverage these concepts, which can increase the degree of freedom in the distribution label. This work only examined the weights in the fully-connected layer, and more informative dependency may be neglected. The self-attention mechanism can be employed on the extended emotion distribution to extract the latent dependency within fine-grained emotion concepts. Besides the subjective text emotion classification, we will try to generalize the label extension methods to other tasks. The extension strategy will be defined according to the subjectiveness and fuzziness of the labels.

CRediT authorship contribution statement

Zongxi Li: Conceptualization, Writing – original draft, Methodology, Software. **Xianming Li:** Validation, Visualization. **Haoran Xie:** Data curation, Writing – original draft, Funding acquisition. **Fu Lee Wang:** Conceptualization, Project administration, Writing – review & editing, Funding acquisition. **Mingming Leng:** Resources, Supervision. **Qing Li:** Writing – review & editing, Supervision. **Xiaohui Tao:** Writing – review & editing, Supervision, Resources.

Data availability

The datasets are publicly available.

References

Ali, F., Kwak, D., Khan, P., El-Sappagh, S., Ali, A., Ullah, S., et al. (2019). Transportation sentiment analysis using word embedding and ontology-based topic modeling. *Knowledge-Based Systems*, *174*, 27–42.

Altınel, B., & Ganiz, M. C. (2018). Semantic text classification: A survey of past and recent advances. *Information Processing & Management*, *54*(6), 1129–1153.

Aly, A., & Tapus, A. (2015). An online fuzzy-based approach for human emotions detection: an overview on the human cognitive model of understanding and generating multimodal actions. *Intelligent Assistive Robots*, 185–212.

Biddle, R., Joshi, A., Liu, S., Paris, C., & Xu, G. (2020). Leveraging sentiment distributions to distinguish figurative from literal health reports on Twitter. In *Proceedings of the web conference 2020* (pp. 1217–1227).

Cambria, E., Li, Y., Xing, F. Z., Poria, S., & Kwok, K. (2020). SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. In *Proceedings of the 29th ACM CIKM* (pp. 105—114). ACM.

Chen, X., Xie, H., Cheng, G., & Li, Z. (2022). A decade of sentic computing: topic modeling and bibliometric analysis. *Cognitive Computation*, *14*(1), 24–47.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the naacl: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). ACL.

Ekkekakis, P., & Russell, J. A. (2013). *The measurement of affect, mood, and emotion: A guide for health-behavioral research*. Cambridge University Press.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3–4), 169–200.

Fei, H., Zhang, Y., Ren, Y., & Ji, D. (2020). Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI conference on artificial intelligence, vol. 34* (05), (pp. 7692–7699).

Feng, J., Rao, Y., Xie, H., Wang, F. L., & Li, Q. (2020). User group based emotion detection and topic discovery over short text. *World Wide Web*, *23*(3), 1553–1587.

Fu, Y., & Liu, Y. (2022). Contrastive transformer based domain adaptation for multi-source cross-domain sentiment classification. *Knowledge-Based Systems*, *245*, Article 108649.

Gao, B.-B., Xing, C., Xie, C.-W., Wu, J., & Geng, X. (2017). Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, *26*(6), 2825–2838.

Geng, X. (2016). Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, *28*(7), 1734–1748.

Gollapalli, S. D., Rozenshtein, P., & Ng, S.-K. (2020). ESTeR: Combining word co-occurrences and word associations for unsupervised emotion detection. In *Findings of the ACL: EMNLP 2020* (pp. 1043–1056). ACL.

Guan, R., Zhang, H., Liang, Y., Giunchiglia, F., Huang, L., & Feng, X. (2022). Deep feature-based text clustering and its explanation. *IEEE Transactions on Knowledge and Data Engineering*, *34*(8), 3669–3680. http://dx.doi.org/10.1109/TKDE.2020.3028943.

Guo, B., Han, S., Han, X., Huang, H., & Lu, T. (2021). Label confusion learning to enhance text classification models. In *Proceedings of the AAAI conference on artificial intelligence, vol. 35* (14), (pp. 12929–12936).

He, H., & Xia, R. (2018). Joint binary neural network for multi-label learning with applications to emotion classification. In *Natural language processing and chinese computing* (pp. 250–259). Cham: Springer International Publishing.

Huang, X., Rao, Y., Xie, H., Wong, T.-L., & Wang, F. L. (2017). Cross-domain sentiment classification via topic-related TrAdaBoost. In *Proceedings of the thirty-first AAAI conference on artificial intelligence* (pp. 4939–4940). AAAI Press.

Huang, M., Xie, H., Rao, Y., Feng, J., & Wang, F. L. (2020). Sentiment strength detection with a context-dependent lexicon-based convolutional neural network. *Information Sciences*, *520*, 389–399.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 conference on EMNLP* (pp. 1746–1751). ACL.

Lee, J. (2022). The emotion is not one-hot encoding: Learning with grayscale label for emotion recognition in conversation. arXiv preprint arXiv:2206.07359.

Li, Z., Chen, X., Xie, H., Li, Q., & Tao, X. (2020). Exploring emotional construction towards multi-class emotion classification. In *2020 IEEE/WIC/ACM International conference on web intelligence (WI)* (pp. 1–8).

Li, H., Chen, Q., Zhong, Z., Gong, R., & Han, G. (2022). E-word of mouth sentiment analysis for user behavior studies. *Information Processing & Management*, *59*(1), Article 102784.

Li, X., Li, Z., Xie, H., & Li, Q. (2021). Merging statistical feature via adaptive gate for improved text classification. In *Proceedings of the AAAI conference on artificial intelligence, vol. 35* (15), (pp. 13288–13296).

Li, Z., Li, X., Xie, H., Li, Q., & Tao, X. (2021). A label extension schema for improved text emotion classification. In *IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology* (pp. 32–39).

Li, Z., Xie, H., Cheng, G., & Li, Q. (2021). Word-level emotion distribution with two schemas for short text emotion classification. *Knowledge-Based Systems*, Article 107163.

Li, X., Xie, H., Rao, Y., Chen, Y., Liu, X., Huang, H., et al. (2016). Weighted multi-label classification model for sentiment analysis of online news. In *2016 International conference on big data and smart computing (bigcomp)* (pp. 215–222). http://dx.doi.org/10.1109/BIGCOMP.2016.7425916.

Liang, W., Xie, H., Rao, Y., Lau, R. Y., & Wang, F. L. (2018). Universal affective model for Readers' emotion classification over short texts. *Expert Systems with Applications*, *114*, 322–333.

Liliana, D. Y., Basaruddin, C., & Widyanto, M. R. (2017). Mix emotion recognition from facial expression using SVM-CRF sequence classifier. In *Proceedings of the international conference on algorithms, computing and systems* (pp. 27–31).

Liliana, D. Y., Basaruddin, T., Widyanto, M. R., & Oriza, I. I. D. (2019). Fuzzy emotion: a natural approach to automatic facial expression recognition from psychological perspective using fuzzy system. *Cognitive Processing*, *20*(4), 391–403.

Mohammad, S. M. (2012). 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* (pp. 246–255). ACL.

Mohammad, S. M. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th annual meeting of the ACL (volume 1: long papers)* (pp. 174–184). ACL.

Mohammad, S., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018). SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation* (pp. 1–17). New Orleans, Louisiana: ACL.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, *29*(3), 436–465.

Muñoz, S., & Iglesias, C. A. (2022). A text classification approach to detect psychological stress combining a lexicon-based feature framework with distributional representations. *Information Processing & Management*, *59*(5), Article 103011.

Pang, J., Rao, Y., Xie, H., Wang, X., Wang, F. L., Wong, T. L., et al. (2019). Fast supervised topic models for short text emotion detection. *IEEE Transactions on Cybernetics*, 1–14.

Plutchik, R. (1980). Chapter 1 - A general psychoevolutionary theory of emotion. In *Theories of emotion* (pp. 3–33). Academic Press.

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, *17*(3), 715–734.

Qin, X., Chen, Y., Rao, Y., Xie, H., Wong, M. L., & Wang, F. L. (2021). A constrained optimization approach for cross-domain emotion distribution learning. *Knowledge-Based Systems*, *227*, Article 107160.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161.

Russell, J. A., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, *11*(3), 273–294.

Scherer, K. R., & Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, *66*(2), 310–328.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., et al. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on EMNLP* (pp. 1631–1642).

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.

Susanto, Y., Livingstone, A. G., Ng, B. C., & Cambria, E. (2020). The hourglass model revisited. *IEEE Intelligent Systems*, *35*(5), 96–102.

Tubishat, M., Idris, N., & Abushariah, M. A. (2018). Implicit aspect extraction in sentiment analysis: Review, taxonomy, opportunities, and open challenges. *Information Processing & Management*, *54*(4), 545–563.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Proceedings of advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, vol. 30* (pp. 5998–6008). Curran Associates, Inc..

Wang, C., & Wang, B. (2020). An end-to-end topic-enhanced self-attention network for social emotion classification. In *Proceedings of the web conference 2020* (pp. 2210–2219). ACM.

Wang, C., Wang, B., Xiang, W., & Xu, M. (2019). Encoding syntactic dependency and topical information for social emotion classification. In *Proceedings of the 42nd international ACM SIGIR conference* (pp. 881—884). ACM.

Xu, N., Liu, Y.-P., & Geng, X. (2020). Partial multi-label learning with label distribution. In *Proceedings of the AAAI conference on artificial intelligence, vol. 34* (04), (pp. 6510–6517).

Yang, C. T., & Chen, Y. L. (2020). DACNN: Dynamic weighted attention with multi-channel convolutional neural network for emotion recognition. In *2020 21st IEEE international conference on mobile data management (MDM)* (pp. 316–321).

Yang, C. C., & Wang, F. L. (2003). Fractal summarization: summarization based on fractal theory. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval* (pp. 391–392).

Yang, C. C., & Wang, F. L. (2007). An information delivery system with automatic summarization for mobile commerce. *Decision Support Systems*, *43*(1), 46–61.

Zhang, Y., Fu, J., She, D., Zhang, Y., Wang, S., & Yang, J. (2018). Text emotion distribution learning via multi-task convolutional neural network. In *Proceedings of the twenty-seventh international joint conference on artificial intelligence, IJCAI-18* (pp. 4595–4601). International Joint Conferences on Artificial Intelligence Organization.

Zhao, Z., & Ma, X. (2019). Text emotion distribution learning from small sample: A meta-learning approach. In *Proceedings of the 2019 conference on EMNLP-IJCNLP* (pp. 3957–3967). Hong Kong, China: ACL.

Zhou, D., Zhang, X., Zhou, Y., Zhao, Q., & Geng, X. (2016). Emotion distribution learning from texts. In *Proceedings of the 2016 conference on EMNLP* (pp. 638–647). ACL.

Zhu, E., Rao, Y., Xie, H., Liu, Y., Yin, J., & Wang, F. L. (2017). Cluster-level emotion pattern matching for cross-domain social emotion classification. In *Proceedings of the 2017 ACM on conference on information and knowledge management* (pp. 2435–2438). New York, NY, USA: Association for Computing Machinery.