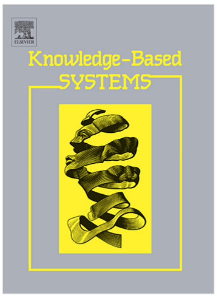


Word-level emotion distribution with two schemas for short text emotion classification

Zongxi Li, Haoran Xie, Gary Cheng, Qing Li

PII: S0950-7051(21)00426-3
DOI: https://doi.org/10.1016/j.knosys.2021.107163
Reference: KNOSYS 107163

To appear in: *Knowledge-Based Systems*

Received date : 20 October 2020
Revised date : 15 March 2021
Accepted date : 17 May 2021

Please cite this article as: Z. Li, H. Xie, G. Cheng et al., Word-level emotion distribution with two schemas for short text emotion classification, *Knowledge-Based Systems* (2021), doi: https://doi.org/10.1016/j.knosys.2021.107163.

Word-level Emotion Distribution with Two Schemas for Short Text Emotion Classification

Zongxi Li[a], Haoran Xie[b,*], Gary Cheng[c], Qing Li[d]

[a]Department of Computer Science, City University of Hong Kong,
Tat Chee Avenue, Kowloon, Hong Kong SAR
[b]Department of Computing and Decision Sciences, Lingnan University,
8 Castle Peak Road, Tuen Mun, New Territories, Hong Kong SAR
[c]Department of Mathematics and Information Technology, The Education University of
Hong Kong, Tai Po, New Territories, Hong Kong SAR
[d]Department of Computing, The Hong Kong Polytechnic University,
Hung Hom, Kowloon, Hong Kong SAR

Abstract

Understanding word-level emotion in terms of both category and intensity has always been considered an essential step in addressing text emotion classification tasks. Existing studies have mainly adopted the categorical lexicons that are tagged by predefined emotion taxonomies to link affective words with discrete emotions. However, in these lexicons, emotion tags are restricted to a specific set of *basic emotions*. Moreover, the emotional intensity is ignored, making these methods less flexible and less informative. This paper proposes a novel method to generate a word-level emotion distribution (WED) vector by incorporating domain knowledge and dimensional lexicon. The proposed method can link a word with more generic and fine-grained emotion taxonomies with quantitatively computed intensities. We propose two schemas to utilize the WED vector implicitly and explicitly to facilitate classification. The implicit approach implements a rule-based conversion strategy to augment the information in the label space. The explicit approach exploits WED as an emotional word embedding to enhance the sentiment feature. We conduct extensive experiments on seven multiclass datasets. The results indicate that both proposed schemas

[*]Corresponding author
Email address: hxie2@gmail.com (Haoran Xie)

produce competitive results compared with the state-of-the-art baselines.

Keywords: Emotion classification, emotional embedding, VAD, emotional lexicon

2010 MSC: 00-01, 99-00

1. Introduction

With the advent of social media and e-commerce, increasing user-generated content has facilitated the development of opinion mining and analysis [1, 2, 3]. Two fundamental tasks have emerged and have received significant attention from academia. One is sentiment analysis, which aims to detect sentiment polarity, *i.e.*, positive, neutral, or negative, from text, audio, and video media. The other is emotion categorization. Compared with sentiment analysis, emotion categorization drills deeper to identify the exact emotions [4, 5]. Recently, **both** sentiment- and emotion-related studies have expedited different research topics and been investigated in tasks, such as dialogue modeling [6], causality analysis [7], speech recognition [8], and face review detection [9].

As a specific natural language processing (NLP) task, text emotion classification aims to correctly detect and identify emotions expressed by short texts. Automatic emotion detection can help analyze user attitudes and opinions from online textual data such as reviews, comments, blogs, and news reports. Automatic emotion detection can be applied to various fields, *e.g.*, customer service, personalized searching, and stock prediction [10, 11]. Researchers have created and utilized different handcrafted lexicons to understand sentiments at the word-level, which has been proven to enhance the classifier's robustness [12, 13]. Furthermore, the effect of linking with discrete emotions has been validated via multi-component analysis [14]. Commonly adopted lexicons are categorical lexicons that annotate each word with one or several tags from the identified collection of emotions, such as NRC (National Research Council of Canada) [15] and SenticNet [16]. We note, however, that this approach suffers from several limitations in its actual use.

	Valence		Arousal		Dominance	
Term	Mean	SD	Mean	SD	Mean	SD
1. Admire	.81	.21	.44	.30	.51	.34
2. Joyful	.76	.22	.48	.26	.35	.31
3. Excited	.62	.25	.75	.20	.38	.29
4. Happy	.81	.21	.51	.26	.46	.38
5. Grateful	.55	.26	.07	.30	-.14	.35
6. Love	.87	.17	.54	.26	-.18	.30
7. Surprise	.40	.30	67	.27	-.13	.38
8. Disgust	-.60	.20	.35	.41	.11	.34
9. Hate	-.56	.16	.59	.29	.13	.34
10. Anger	-.51	.20	.59	.33	.25	.39
11. Fear	-.64	.20	.60	.32	-.43	.34
12. Guilty	-.57	.19	.28	.38	-.34	.28
13. Sadness	-.63	.23	-.27	.34	-.33	.22
14. Bored	-.65	.19	-.62	.24	-.33	.21
15. Shame	-.57	.23	.01	.46	-.34	.30

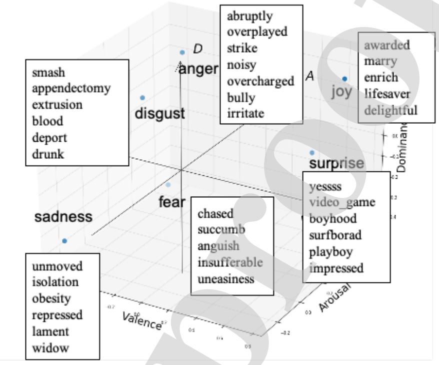


Figure 1: (Left) 15 common emotions from Three-Factor Theory. (Right) Visualization of six basic emotions from Ekman [17] in VAD space with nearest words from NRC-VAD lexicon.

The first limitation is that the emotion tags used in lexicons are restricted to a limited set. The tags that annotate words are defined according to the emotion theory that the lexicon follows. To simplify the annotation process, researchers typically choose theories with universally recognized emotions, such as the *six basic emotions* [17], which contain fear, anger, joy, sadness, disgust, and surprise. However, such taxonomies are overly general and cannot linked to more delicate emotions. Furthermore, emotional tags can vary in different lexicons and are inconsistent with dataset labels in a classification task, leading to compatibility issues in implementation.

The second limitation is the lack of emotional intensity information in the current categorical-lexicon-based approaches. Lexicon knowledge cannot quantitatively measure how strong a word expresses a specific emotion, making the use of categorical lexicons less flexible and less informative.

This paper proposes an effective method to associate the words with more fine-grained emotions by incorporating a dimensional lexicon dictionary and psychological domain knowledge to address the problems of existing categorical lexicons identified above. Russell and Mehrabian [18] proposed the Three-Factor

Theory, which constructs a system with three independent and bipolar dimensions, i.e., Valence (**V**, pleasantness), Arousal (**A**, intensity of emotion), and Dominance (**D**, degree of power exerted). They suggested that such a system is necessary and sufficient to adequately define emotional states. They have identified 151 concepts denoting emotions in terms of Valence-Arousal-Dominance (VAD) with mean and standard deviation, which can be regarded as the *Knowledge of Emotions* (KoE). Moreover, NRC-VAD [19] annotated more than 20,000 English affective words using a similar approach, which can be leveraged as the *Knowledge of Words* (KoW). In Figure 1, we visualize a subset of emotions from the Three-Factor Theory and their nearest words from NRC-VAD in the same space to demonstrate the compatibility of KoE and KoW. Referring to these their works, we assume that each emotion follows a multivariate Gaussian distribution within the VAD space and apply a probabilistic model to determine the relationship between a given pair of words and emotions. Finally, we obtain the *word-level emotion distribution* (**WED**) to establish a connection between words and general emotions with quantified intensities, resulting in a unique affective word representation.

To apply WED to sentence emotion classification, we propose two novel schemas to demonstrate how informative WED is in a real application.

Schema 1 Generate the *sentence-level emotion distribution* (**SED**) label. The emotions of a sentence are subjective and fuzzy, with more than one emotion being expressed simultaneously. However, in the majority of scenarios, sentences are annotated with a one-hot label, which leads to the issue of label ambiguity [20], referring to the uncertainty, incompleteness, and even mistakes among the ground-truth labels. Incomplete label information can limit classifier performance. Researchers have made multiple efforts to address emotion categorization tasks using distribution learning [21, 22]. A smoother emotion label can be learned and utilized to compensate for the missing values in the one-hot label. In particular, Zhang et al. [22] adopted a categorical lexicon to generate an emotion distribution. We exploit a rule-based conversion strategy to generate a SED label using WED. Similar to [22], we train the classifier with both ground-truth labels for single-label prediction and the SED labels for auxiliary distribution learning synchronously. A detailed discussion on generating a SED using the categorical lexicon and WED is provided in Section 5.2.

Schema 2 Employ WED as the initialization of *emotion word embedding* (**EWE**). To capture the sentiment deviation and aggregation, we prefer a dense representation of text rather than discrete input in a deep architecture. Pretrained word embedding cannot explicitly reflect emotion-specific properties because of the distributional hypothesis. In this schema, we exploit all 151 identified emotions from the Three-Factor Theory as anchor points to produce a 151-dimensional WED vector for each word. Subsequently, the vectors are applied as the initialization of the EWE and are trainable during the optimization process. Instead of random initialization, emotion representation learning does not start from scratch. We can alleviate the influence of limited training data and improve the interpretability of dense representations.

Our main contributions are summarized as follows:

- We incorporate psychological domain knowledge and a commonly adopted dimensional lexicon to generate a WED as the affective representation of words, which is more informative compared to using a categorical lexicon;
- We propose two novel schemas based on WED to improve model performance for sentence-level emotion classification.

2. Related Work

2.1. Emotion Theory and Emotional Lexicon

Ekkekakis [23] suggested that existing emotion theories are generally either categorical or dimensional. Hence, we can classify the commonly used lexicons into *categorical lexicons* and *dimensional lexicons* according to the emotion theory they follow. Categorical emotion theories describe emotions as a finite set of discrete taxonomies to characterize the state of mind [17, 24]. Categorical lexicons annotate each word with one or several tags from an identified collection of emotions, such as NRC [15] and SenticNet [16]. Although categorical lexicons explicitly connect words with emotions, the connection is limited to a finite scope and provides no information regarding the intensity. Conversely, dimensional emotion theories conceptualize emotions with measurable variables [18, 25], and dimensional lexicons, such as NRC-VAD [19], annotate each word with a tuple of three dimensions as stated previously: *valence*, *arousal*, and *dominance*. Dimensional lexicons quantitatively describe words in a sentiment space. However, they cannot directly link a concept to a specific emotion.

2.2. Emotion Classification and Emotional Lexicon

Emotion sensing and categorization have been extensively researched in recent years [26]. Different neural network models have been proposed to address this task and have achieved competitive results on several benchmarks. Yu et al. [27] exploited transfer learning to facilitate emotion classification tasks using a sentiment analysis task. Fei et al. [28] proposed a latent emotion memory network to learn latent emotion distribution without external knowledge for multi-label emotion classification. Tanabe et al. [29] extracted narrative context using a Bert-based encoder and employed a priori KoE categories in textual emotion categorization.

Concerted efforts have been made to use emotion lexicons to improve emotion classification performance [30]. Teng et al. [31] proposed a context-sensitive lexicon-based method based on a weighted-sum model to calculate sentiment aggregation using recurrent neural network (RNN) architecture. Liang et al. [32] designed a universal affective model to establish a word-level lexicon dictionary using a supervised topic model. Dragoni et al. [33] presented a commonsense ontology based on SenticNet for sentiment analysis. Xing et al. [34] addressed opinion mining using domain adaptation of sentiment lexicons. Instead of using zero padding, Li et al. [13] proposed a sentiment padding to force consistent size on the input data sample and improve the proportion of sentiment information in each review. All of the above approaches exploit the categorical lexicon. Similar to ANEW [35] and NRC-VAD [19], multidimensional lexicons, which have been widely adopted in sentence-level polarity value prediction tasks, have not been explored in the classification task.

2.3. Representation Learning in Natural Language Processing

Representation learning maps each term to a fixed-length representation in a continuous space. Based on the distributional properties in linguistics, generic word embedding vectors are trained on a large amount of unannotated textual data covering different latent topics [36, 37]. Labutov and Lipson [38] claimed that the effectiveness of generic word representation is task-dependent, and they produced task-specific embeddings from existing word embeddings for sentiment analysis. Several affective representation learning methods rely on training embeddings from scratch on large tweet datasets [39, 40]. Agrawal et al. [41] projected emotionally similar words into neighboring spaces and emotionally dissimilar ones far apart to obtain emotion-enriched word representations. However, they adopted a categorical lexicon to determine the emotions, and the scope of application was limited to a set of emotions. Babanejad et al. [42] systematically examined the role of preprocessing training corpora to produce word representations for affect analysis tasks. Unlike start-from-scratch approaches and word-embedding fine-tuning approaches, we initialize word vectors by explicitly modeling the word-emotion relationship. Thus, the learned representation is emotion-specific and interpretable.

3. Methodology

3.1. Word-level Emotion Distribution

Given a set of K emotions, $E_1, E_2, \ldots, E_K$, we compute the probability of a word belonging to each emotion as the corresponding emotion intensity. We query 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 for each emotion E_k. Moreover, given the word w included in the NRC-VAD dictionary, we can retrieve the VAD tuple $\mathbf{w} = [V^w, A^w, D^w]$. Considering the three factors as independent variables, we apply a multivariate Gaussian distribution to compute the intensity of emotion E_k, $\mathbf{d}_w^{E_k}$:

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

where $\boldsymbol{\Sigma} = diag(\boldsymbol{\sigma}_{E_k})$. For the K emotions, the WED of word w is $\mathbf{d}_w = [d_w^{E_1}, d_w^{E_2}, \ldots, d_w^{E_K}]$.

Consequently, we can produce a unique affective representation of words by modeling their *belongingness* with emotions. Compared with approaches using categorical lexicons, the proposed method has the following advantages:

1. A generalized word-emotion connection is produced beyond a limited collection of tags;
2. The intensity of the emotions can be explicitly described by the "closeness" with the words in the VAD space;
3. By mapping discrete words into higher-dimensional vectors of real numbers utilizing domain knowledge, the WED is encoded with sentiment-specific information and has excellent future utilization flexibility.

The major challenge of this process is the addressing of out-of-lexicon (OOL) words. We observe that out-of-dictionary words can appear in the following situations: (1) the word itself is not affective; (2) certain adverbs are missing, yet they have an adjective or verb form in the dictionary; and (3) the word is misspelled or is a typo. Herein, we do not consider the non-affective or misspelled words. For the words in the second category, we use WordNet[1] to search relevant words on the same branch or synonym words to enhance the VAD lexicon knowledge coverage.

3.2. Schema I: Implicit Approach

WED is not employed directly in the network at the word-level in the implicit approach. Instead, inspired by Zhang et al. [22], we propose a rule-based

[1] https://wordnet.princeton.edu/

Algorithm 1: Rule-based conversion for SED and pSED generation

```
Input: A sentence with $M$ words, $\{w_1, w_2, \ldots, w_M\}$, and
    (1) Ground-truth label $y$ and dominant emotion intensity $\epsilon$;
    (2) WEDs $\mathbf{d}_w$;
    (3) POS tags $\mathrm{POS}_w$;
    (4) Polarity values $p_w$.
Output:
    Sentence-level emotion distribution (SED);
    Pseudo sentence-level emotion distribution (pSED).
Initialization : Vector of $\mathbf{d}_s$, constant $c$.
for each affective word in the sentence do
    if $\mathrm{POS_w}$ = Adj. or Adv.
        then $c = 1$;
    else if $\mathrm{POS_w}$ = Verb or Noun
        then $c = |p_w|$;
    else $c = 0.1$;
    end if
    if negation word appears before the word
        then $\mathbf{d}_s += c \times \mathbf{d}_w^-$;
    else $\mathbf{d}_s += c \times \mathbf{d}_w$;
    end if
end
SED $\leftarrow \mathbf{d}_s^{\mathrm{SED}}$ = Normalize($\mathbf{d}_s$) ;
$\mathbf{d}_s^{\mathrm{SED}}[y] = \epsilon$ ;
pSED $\leftarrow \mathbf{d}_s^{\mathrm{pSED}}$ = Normalize($\mathbf{d}_s^{\mathrm{SED}}$) ;
return SED, pSED
```

method to produce SED labels for distribution learning, which aims to address the label ambiguity problem that is overlooked by existing emotion classification approaches. As mentioned previously, sentences typically express a mixture of emotions in a fuzzy style, thereby increasing the diffculty of annotation. Therefore, the majority of emotion-relevant datasets are annotated in a one-hot manner by indicating the most prominent emotion while neglecting non-dominant emotions, which makes automatic detection more challenging given the incomplete label information.

The proposed solution is to produce a dense vector of distribution-style as an analog of the real emotion distribution to compensate for the label ambiguity setback. Given a sentence and the WED of each word[2], one can adopt a straightforward weighted-sum operation to count all WEDs and normalize them to percentages summed to one. Such a method is based on the bag-of-words model, and hence suffers from *semantic compositionality* [43]. The composition effects exhibit intricacies [31] such as negation over intensification (*e.g.*, not very good) and shifting (*e.g.*, not interesting). To alleviate this problem, we propose a rule-based strategy to adapt WEDs dynamically by considering the appearance of negation tokens ("no" and "n't"). In addition, considering that each word contributes differently to the sentence-level emotion because of grammatical functions, *i.e.*, the part-of-speech (POS) and degree of sentiment polarity, the rule-based strategy is also designed to highlight more affective words, such as adjectives and adverbs. We apply a syntax parser from the spaCy[3] package to extract the POS tags of each word to leverage the syntax information. To exploit polarity information, we adopt the polarity value from SenticNet 6[4] [16] for each word, which is a real number between 1 and -1. Here, we exploit the absolute numerical value only as the sentiment intensity.

More specifically, we assign a weight to each affective word in the incre-

[2] Note that, for the implicit approach, we only select a subset of emotions from the Three-Factor Theory to form the WED, which is consistent with the dataset labels.

[3] `https://spacy.io/usage`

[4] `https://sentic.net/api/`

mental process according to its POS tag. A full weight is assigned if a word is descriptive, such as an adjective or adverb. If a word is a noun or verb, the weight will be its polarity value. Moreover, the appearance of a negation word is also considered.

As generating the SED is unsupervised, the emotion with the maximum value in the SED can be different from the ground-truth label. The differences are due to various reasons, such as word sense variation (*e.g.*, cancer is eliminated) and ironic statements (*e.g.*, feel so good to fail an exam). Because our ultimate goal remains single-label prediction, the inconsistency can confuse the classifier during training. Furthermore, the weights of non-dominant emotions must be restricted; otherwise, noise can be introduced and causing instability in the learning process. Therefore, we adopt *pseudo-sentence-level emotion distribution* (**pSED**) by substituting the real value of the ground-truth emotion from the SED with a greater value ϵ, maintaining the ground-truth emotion prominent within the distribution. To train the classifier, pSED is combined with the one-hot label for distribution learning and single-label classification, respectively.

We present the proposed rule-based method for generating both the SED and pSED in Algorithm 1. $\mathbf{d}_w^-$ means we use a visual point $\mathbf{w}' = [-V^w, -A^w, -D^w]$ to compute the probability in Equation 1, which we found is a workable shortcut to address sentiment shifting when negation tokens appear, such as "n't" and "not." The implementation details of the classifier are presented in Section 3.4.

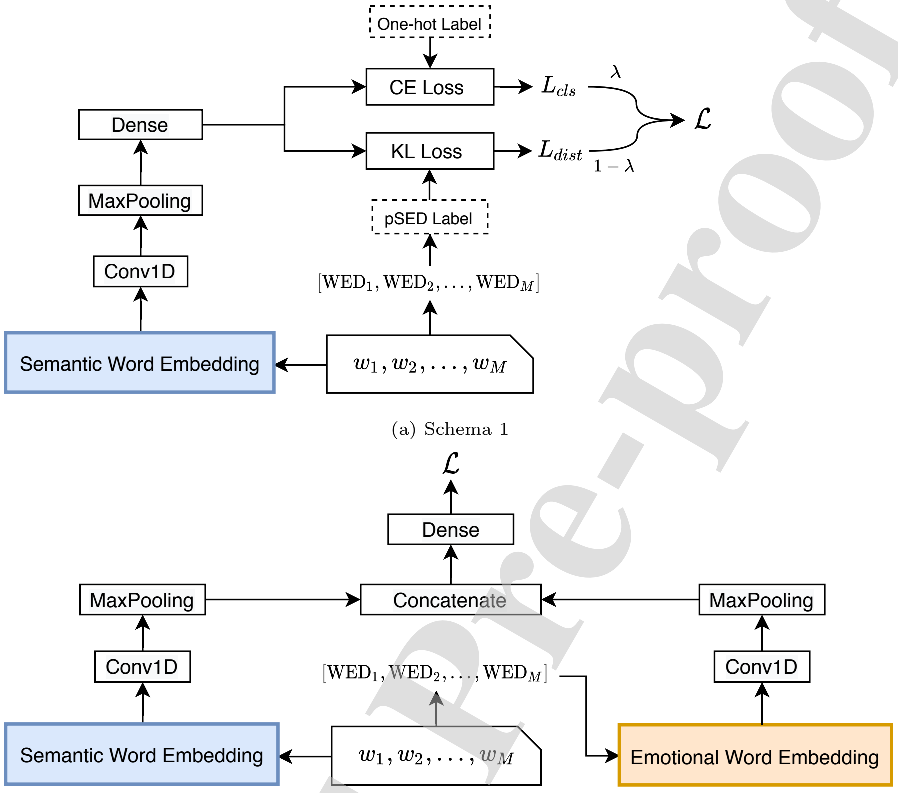
3.3. Schema II: Explicit Approach

The above implicit approach is designed to augment the information in the label space and enhance model performance by employing a more informative distribution label. In this section, we devise an explicit approach to utilizing WED as a feature of the word. We construct the EWE using WED and employ a convolutional neural network (CNN) to model the sentiment aggregation and shifting.

To be consistent with dataset labels, we adopted only a subset of emotions from the Three-Factor Theory in the implicit approach. However, we consider all 151 identified emotions as anchor points for the explicit approach and compute $\mathbf{d}_w$ with each emotion (via Equation 1) to produce a more informative representation in the EWE. Hence, each affective word is mapped to a 151-dimensional dense vector. However, there could be a large number of OOL tokens without a corresponding EWE vector, and consequently, the entire EWE matrix could be undesirably sparse. We randomly initialize vectors for OOL tokens to mitigate this sparsity issue, following a similar method to implementing pretrained word embedding. Specifically, initialization values are sampled from a distribution with a considerably smaller standard deviation, e.g., 0.001, to distinguish affective tokens from OOL tokens. Moreover, all weights in the EWE matrix are trainable during the training process, allowing the decision-making to be more adaptive.

The advantages of using WED as the initialization of the EWE can be summarized as:

1. WED encodes sentiment-specific features. In the majority of pretrained embeddings, emotionally dissimilar words, such as "happy" and "sad", share a higher cosine similarity than that of emotionally similar words, such as "happy" and "joy" owing to the distributional hypothesis [41]. Conversely, emotions within a close region of the VAD space have a high similarity to the WED because their relationships with anchor emotions are comparable.
2. WED has superior interpretability and greater efficiency compared with random initialization. Though emotions are mapped to a high-dimensional space, WED can be explained because the vector values reflect sentiment distances with corresponding emotions. Moreover, WED is a more favorable initialization option given limited training data compared to starting from scratch.



(a) Schema 1

(b) Schema 2

Figure 2: Generic framework of proposed method.

3.4. Generic Framework

In this section, we present the implementation of the schemas described above. As a heuristic exploration of the potential applications of WEDs and SEDs, we avoid using an overly complicated network in the framework. As indicated in Figure 2, the proposed framework employs a CNN to extract the semantic features from word vectors. In our experiments, CNN achieved better results with more stable performance than recurrent networks.

For the implicit schema, the loss function consists of cross-entropy loss L_{cls} for the classification task and Kullback–Leibler loss L_{dist} for distribution learning. We denote the output logits as $\mathbf{a} \in \mathbb{R}^K$, with

$$L_{cls}(\mathbf{a}, y) = -\frac{1}{N}\left[\sum_{i}^{N}\sum_{j}^{K}\mathbf{1}(y_i = j)\ln\frac{e^{a_j^{(i)}}}{\sum_t e^{a_t^{(i)}}}\right], \tag{2}$$

$$L_{dist}(\mathbf{a}, \mathbf{d}^{\mathrm{pSED}}) = -\frac{1}{N}\left[\sum_{i}^{N}\sum_{j}^{K}(\mathbf{d}_j^{\mathrm{pSED}})\ln\frac{e^{a_j^{(i)}}}{\sum_t e^{a_t^{(i)}}}\right]. \tag{3}$$

where $\mathbf{1}(y_i = j) = 1$ if $y_i = j$ and $\mathbf{1}(y_i = j) = 0$, otherwise. To optimize the model, weighted sum $\mathcal{L} = \lambda \cdot L_{cls}(\mathbf{a}, y) + (1 - \lambda) \cdot L_{dist}(\mathbf{a}, \mathbf{d})$ is calculated based on Equations 2 and 3.

For the explicit schema, we concatenate the latent representations of both embeddings after the MaxPooling layer. After passing through the fully-connected and softmax layers, the concatenated feature vector is mapped to logits in the label space for label prediction; the loss is calculated using Equation 2.

Data	*Train*	*Test*	*Label*
AT	$1,000$	250	anger, disgust, fear, joy, sadness, surprise
IS	$7,666$	CV	anger, disgust, fear, joy, sadness, shame, guilt
FT	$1,211$	CV	anger, fear, happy, sadness, surprised
TEC	$21,051$	CV	anger, disgust, fear, joy, sadness, surprise
CF	$40,000$	CV	happiness, worry, surprise, love, sadness
SST-1	$11,855$	$2,210$	very positive/negative, positive/negative, neutral
SST-2	$9,613$	$1,821$	positive, negative

Table 1: Summary statistics for datasets. *Train*: Dataset size. *Test*: Size of test set (CV means no standard train/test split). *Label*: emotion/sentiment labels used in the dataset

4. Experiments

4.1. Datasets

We conducted experiments on datasets of different topics (with summary statistics in Table 1).

AffectiveText (AT) contained extracted news headlines from major news websites [44]. The headlines were annotated using crowd-source voting on six emotion labels. Each voting was normalized to percentages summed to one as the ground-truth distribution label.

ISEAR (IS) was collected by inviting personnel with and without psychological backgrounds to describe a personal experience where they scored the experienced according to seven emotions [45].

Fairy Tales (FT) contained 185 children's stories. Each sentence was annotated with one of the five emotion classes. Among all the sentences, 1,211 highly affective sentences agreed upon by the majority of annotators were listed [46].

TEC included emotional tweets selected by prespecified hashtags. Each tweet was labeled by one of the six emotions [47].

CrowdFlower (CF) included 40,000 tweets annotated by crowd-sourcing with 12 emotions. In this work, we adopted a subset of **CF** because the remaining emotions had limited data.

The datasets described above were annotated with fine-grained emotional tags; thus, the first schema could be implemented by producing pseudo labels according to the dataset emotion labels. In addition, we further demonstrated the effectiveness of emotional word embedding (the second schema) on sentiment analysis tasks with polarity labels, *i.e.*, positive and negative.

SST-1[5] is the Stanford Sentiment TreeBank dataset of movie reviews with five fine-grained sentiment labels [48].

SST-2 is the Stanford Sentiment TreeBank dataset with binary sentiment labels [48].

We deployed 10-fold cross-validation on the datasets without a standard train/test split and reported the average values with standard deviation. For datasets with standard train/test splits, we performed ten trials and reported the average values.

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

4.2. Baselines

For the distribution dataset AT, we compared the proposed method with the label distribution learning methods including **AA-KNN, BFGS, CPNN** [49]. AA-KNN directly applies k-nearest neighbors with neural networks to predict the label distribution. BFGS and CPNN are parametric models that use machine-learning techniques. For the single-label datasets, we compared the proposed model with the following emotion categorization baselines:

OCC [50] is a feasible rule-based emotion detection model, where a rule-based pipeline approach is proposed to detect implicit emotions. **DACNN** [51] combines the multi-channel convolutional network and attention mechanism of automatic weight adjustment to improve emotion recognition performance. **ES-TeR** [52] integrates large-scale word co-occurrences and word associations for unsupervised emotion detection. A similarity function is adopted based on random walks on the graphs. **MCCNN** [53] exploits a multi-channel convolutional network to incorporate different emotion and sentiment indicators such as hashtags, emoticons, and emojis. **DERNN** [54] encodes sentence syntactic dependency and document topical information into the document representation for emotion classification. **Word Rep.** [42] examines word vector-based models applied to affective systems and provides a comprehensive analysis of the role of preprocessing techniques in affective analysis based on word vector models. **ASEDS** [55] proposes a framework to learn sentiment representation using Facebook posts and reactions. **TESAN** [56] constructs a neural topic model to learn topical information and generates a topic embedding for a document. A topic-enhanced self-attention module and a fusion gate are used for predicting the emotion label. **WLTM** [57] is a weighted labeled topic model to address the sparsity issue of short-text sentiment mining. **MTCNN** [22] is a multi-task CNN for emotion classification; the distribution learning utilizes pseudo sentence emotion distribution generated using categorical lexicon knowledge.

We also **compared it to** the following widely-adopted text classification baselines:

TextCNN [58] is a popular CNN-based classifier exploiting one-dimensional convolution operation on an embedding matrix and max-over-time pooling on the extracted the feature map. **DPCNN** [59] is a low-complexity word-level deep CNN model in pyramid shape employing a downsampling module and shortcut connections; **Bi-LSTM** [60] is a bi-directional long short-term memory (LSTM) model extracting both forward and reverse sequential features. **C-LSTM** [61] employs a CNN model to extract a sequence of higher-level semantic features and feeds these vectors into an LSTM network to obtain the sentence representation for classification. **Transformer** [62] stacks multiple blocks of self-attention to produce a robust sentence-level representation by learning global dependency. We use the encoder part of the Transformer followed by a classifier.

We also tested the proposed method using a state-of-the-art contextual embedding approach, such as Bert. **Bert** [63] has made significant progress in numerous NLP tasks. It combines Transformer-based architecture and a large corpus to maximize Transformer's ability. We employed the Bert model on the textual input and fine-tuned the pretrained Bert base with classifier. To provide a fair comparison, we compared the Bert model to **Bert+Ours** (imp.) and **Bert+Ours** (exp.) to test if the proposed method could substantially improve the Bert model.

4.3. Evaluation metrics

We evaluated the model performance and significance of improvement using the following metrics: **F1 score** assesses multi-class classification performance comprehensively as it measures both precision and recall as a whole. We report the *Macro*-average results in this paper. **Accuracy** measures how many instances are correctly classified among all instances. **T-test** reveals how significant the improvements are; we report the p-value of the proposed model compared to the baselines for each trail.

4.4. Word embedding and preprocessing

We adopted the publicly available pretrained language model *FastText* [6] [37], which has one million word vectors with the dimensionality of 300. Words not present in the pretrained model are initialized randomly (in this work, we adopted non-subword embedding instead of subword embedding, as using subword embedding produces lower results). The preprocessing of all datasets followed that of Kim [58]. Moreover, we filtered sentences of foreign languages and those with low quality from two tweeter datasets by setting a threshold on the proportion of English words in one sentence; the threshold was 0.4 and no more than 500 sentences were removed.

4.5. Hyperparameter setting

In the proposed framework, the variables that influence model performance are the dominant emotion intensity ϵ in the pSED and weight of the cross-entropy loss λ in the weighted loss term, which both control the proportion of non-dominant emotions considered in the model optimization. To determine the appropriate values of these two variables, we conducted a grid-search on a validation set of each dataset, which was a 20% subset of the training set. In general, $\epsilon = 0.8$ and $\lambda = 0.7$ produced the best results for the validation sets of all datasets.

The hyperparameters of the network components were consistent between the baseline models and the proposed models. More specifically, the CNN-based models had a filter size of $[3, 4, 5]$ with 100 filters each; the RNN-based models had hidden dimensions of 128. For Transformer, we used the Transformer's encoder as the feature extractor, specifically with eight heads and three blocks. The employed Bert model was the Bert-base Uncased model, which included 12 layers, 768 hidden units, 12 heads, and 110 M parameters. All models adopted the Adam optimizer with a batch size of 64 and dropout rate of 0.5.

[6] `https://fasttext.cc/docs/en/english-vectors.html`

4.6. Results

We conducted experiments using two variants of the explicit schema. **Ours (exp. rand)** is a variant where all EWE vectors are randomly initialized and then updated during training. **Ours (exp. static)** is another variant where the EWE consists of WEDs and the randomly initialized vectors remain static. **Ours (combined)** is a combination of the implicit and explicit approaches. The results of the proposed models against other non-Bert baseline methods on the emotion datasets are listed in Tables 2 and 3. The results of the Bert-based models on the emotion datasets are reported in Table 4. The results for the sentiment datasets are presented in Table 5. In general, the proposed method achieved the best results for both distribution and single-label datasets. The improvements were significant compared with the majority of non-Bert baselines. For the Bert-based method, both schemas yielded substantial improvements to the Bert baseline. Remarkably, the proposed explicit approach significantly outperformed the baseline methods on SST-1 and SST-2, which validates the effectiveness of the EWE in sentiment analysis tasks.

4.6.1. Results of non-Bert models

Ours (imp.) outperforms both TextCNN and MTCNN because the generated SED can supplement the beneficial information during training. However, the improvements on tweet datasets such as TEC and CF are less impressive from the dataset aspect. The informal language usage of tweets causes difficulty in the semantic feature extraction and SED generation. Therefore, a greater weight for distribution learning can benefit a dataset with a formal language (such as IS and FT). Similarly, on TEC and CF, the model depends more on classification tasks owing to the bias in the generated pSED. We provide more discussion in Section 5.1. Nevertheless, it is noticeable that Ours (imp.) yields a higher magnitude of improvement compared to MTCNN on CF than that on TEC. Apart from the fact that the generated SED is more desirable, the reasons also include the problematic issue of using a categorical lexicon. More discussion is provided in Section 5.2. We observed that the improvement of accuracy

Model	Accu. (%)				
	AT	IS	FT	TEC	CF
AA-KNN [49]	24.10	–	–	–	–
BFGS [49]	24.22	–	–	–	–
CPNN [49]	25.18	–	–	–	–
WLTM [57]	–	36.50	–	–	–
OCC [50]	–	51.5	–	–	–
DACNN [51]	–	–	–	62.73	–
MCCNN [53]	–	–	–	56.5	–
TESAN [56]	**52.28**	61.14	–	–	–
TextCNN	41.73	64.85§ ±1.06	78.79§ ±3.12	61.94‡ ±0.97	46.39‡ ±0.95
DPCNN	34.66	64.24§ ±1.16	81.19† ±3.29	60.36§ ±1.35	45.90‡ ±0.73
Bi-LSTM	32.67	64.88§ ±0.87	74.93§ ±3.86	62.46 ±1.10	46.27† ±0.69
C-LSTM	32.67	64.56§ ±1.30	74.53§ ±4.08	59.97§ ±1.01	44.13§ ±0.44
Transformer	–	65.37† ±1.20	72.99† ±4.34	62.25† ±0.77	46.69‡ ±1.54
MTCNN [22]	51.60	66.03† ±0.58	81.12† ±2.92	62.32† ±0.82	46.04‡ ±0.90
Ours (imp.)	**52.10**	66.76 ±0.54	83.53 ±2.71	62.57 ±0.95	**47.41** ±0.83
Ours (exp. rand)	40.26	66.21 ±1.23	81.93 ±2.92	62.33 ±0.97	46.91 ±1.08
Ours (exp. static)	48.25	66.52 ±0.51	82.20 ±3.16	62.78 ±0.96	47.06 ±1.24
Ours (exp.)	52.01	**66.98** ±0.69	**84.13** ±3.39	**63.51** ±1.06	47.32 ±1.17
Ours (combined)	51.83	66.93 ±0.57	83.46 ±3.47	63.48 ±0.97	47.34 ±1.21

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

Table 2: Results of Accuracy on emotion datasets (Upper: results from referenced papers; Middle: reproduced baselines; Bottom: Proposed methods). Baseline results with specified p-value indicate that proposed methods demonstrated significant improvement to the baselines.

on AT dataset is not impressive. We believe that this is because the dataset contains limited data samples and each data sample is a short news heading, which leads to severe data sparsity issues that cause difficulty in performance improvement. Conversely, we noticed that the proposed method produced more performance gain in the F1 score, which means that the proposed method can facilitate learning a relative balanced model on each class.

Ours (exp. rand) can produce competitive results with additional randomly initialized embedding for the explicit approaches. However, the margins to models with WED as the initialization are considerable because the randomly initialized vectors cannot provide any auxiliary information to the model. Moreover,

Model	F1 (%)				
	AT	IS	FT	TEC	CF
AA-KNN [49]	18.33	–	–	–	–
BFGS [49]	20.69	–	–	–	–
CPNN [49]	16.41	–	–	–	–
ESTeR [52]	–	–	–	39.8	42.2
MCCNN [53]	–	–	–	**55.6***	–
ASEDS [55]	–	42.2*	46.0*	47.3*	–
DERNN [54]	–	60.44	–	–	–
Word Rep. [42]	–	59.48	–	–	–
TextCNN	35.11	64.13§ ±0.94	74.30§ ±2.60	53.10§ ±1.01	38.76§ ±1.17
DPCNN	31.55	64.04§ ±1.19	75.01§ ±3.54	50.28§ ±1.35	40.60§ ±1.00
Bi-LSTM	20.69	63.89§ ±1.09	70.29§ ±4.24	54.49 ±1.61	42.47† ±0.92
C-LSTM	23.09	63.37§ ±1.09	70.01§ ±3.25	51.03‡ ±1.45	40.87‡ ±0.51
Transformer	–	65.02† ±1.08	66.32† ±4.81	53.52† ±1.53	41.47‡ ±1.22
MTCNN [22]	41.41	65.93‡ ±0.79	76.20§ ±2.66	54.19 ±1.54	40.46† ±1.38
Ours (imp.)	**43.93**	66.45 ±0.71	78.56 ±2.03	54.27 ±1.41	42.62 ±1.86
Ours (exp. rand)	34.45	65.87 ±0.62	77.86 ±2.98	52.62 ±1.80	41.17 ±1.19
Ours (exp. static)	40.52	65.72 ±0.87	78.61 ±2.73	**54.56** ±1.31	42.91 ±1.14
Ours (exp.)	41.68	**66.72** ±0.58	**80.09** ±3.59	54.40 ±1.42	45.95 ±0.97
Ours (combined)	42.32	66.48 ±0.66	79.46 ±3.28	54.41 ±1.73	**42.96** ±0.97

$^\dagger p < .05$, $^\ddagger p < .01$, $^\S p < .001$, * micro F1 score

Table 3: Results of Macro F1 on emotion datasets (Upper: results from referenced papers; Middle: reproduced baselines; Bottom: Proposed methods). Baseline results with specified *p*-value indicate that proposed methods demonstrated significant improvement to the baselines.

the models with WED as the initialization demonstrated substantial improvements to the implicit approach on all single-label datasets (except for accuracy on CF), indicating that CNN over EWE can better capture sentiment aggregation than the rule-based method. In addition, fine-tuning the EWE vectors for each benchmark (except F1 on TEC) gives further improvements over the static method.

We also examined the model performance with both schemas acting together. Although acceptable results are observed, the combined method could not significantly outperform individual schema. This observation indicates that the effects resulting from the schemas are not additive.

Model	Accu. (%)			
	IS	FT	TEC	CF
Bert	70.11‡ ±1.56	80.95† ±3.62	64.07† ±0.70	47.98† ±2.57
Bert w/				
Ours (imp.)	70.93 ±1.14	81.57 ±2.15	**64.50** ±0.95	48.49 ±1.48
Ours (exp.)	**71.09** ±1.04	**82.39** ±2.09	64.47 ±1.31	**48.52** ±1.17

Model	F1 (%)			
	IS	FT	TEC	CF
Bert	69.94§ ±1.42	74.68† ±3.34	55.08‡ ±1.42	44.12§ ±1.37
Bert w/				
Ours (imp.)	70.82 ±1.04	75.96 ±2.15	56.04 ±1.77	44.36 ±1.45
Ours (exp.)	**71.26** ±1.04	**77.38** ±2.30	**56.05** ±1.56	**44.89** ±1.53

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

Table 4: Results of Bert-based models on emotion datasets. Baseline results with specified p-value indicate that proposed methods demonstrated significant improvement to the baselines.

4.6.2. Results of Bert-based methods

Compared with the non-Bert baselines, Bert-based methods produce significant improvements on the majority of datasets because of their superior contextual semantic extraction ability. Bert-based models fail to achieve the best results on FT dataset, where the parameters in the pretrained Bert model are difficult to fine-tuned well because of the limited data samples.

Moreover, both proposed methods demonstrate improvements to the Bert classifier on all datasets, and the T-test indicates that the improvements to Bert are significant. For example, Bert+Ours (imp.) increases the accuracy and F1 score on IS by 0.82% and 0.88%, and Bert+Ours (exp.) increases the accuracy by 1.44% and the F1 score by 2.7% on FT. The results validate that both schemas using WED can improve the Bert-based models.

4.6.3. Results on Sentiment Analysis Task

In the sentiment analysis task, fusing latent emotional features extracted from the EWE with latent semantic features extracted from textual input pro-

Model	Accu. (%)		F1 (%)	
	SST-1	SST-2	SST-1	SST-2
TextCNN	$45.39^{\S}$	$84.07^{\S}$	$41.63^{\S}$	$84.06^{\S}$
DPCNN	$46.43^{\S}$	$84.84^{\S}$	$43.86^{\S}$	$84.84^{\S}$
Bi-LSTM	$44.64^{\S}$	$83.80^{\S}$	$42.58^{\S}$	$83.78^{\S}$
C-LSTM	$46.11^{\S}$	$82.61^{\S}$	$44.32^{\S}$	$82.54^{\S}$
Transformer	$43.41^{\S}$	$83.68^{\S}$	$41.56^{\S}$	$83.67^{\S}$
Ours (exp. rand)	46.78	85.10	43.12	85.57
Ours (exp. static)	45.39	85.68	42.03	85.67
Ours (exp.)	**48.00**	**86.36**	**45.14**	**86.36**
Bert	$53.27^{\ddagger}$	$91.25^{\ddagger}$	$51.00^{\ddagger}$	$91.25^{\ddagger}$
Bert w/ Ours (exp.)	**54.37**	**93.47**	**53.38**	**93.46**

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

Table 5: Results of Accuracy and F1 on SST-1 and SST-2. Baseline results with specified p-value indicate that proposed methods demonstrated significant improvement to the baselines.

duces significant improvements when compared to both Bert-based and non-Bert-based baselines. Compared with TextCNN, the Ours (exp.) increases the accuracy and F1 by 2.61% and 3.15% on SST1, and 2.29% and 2.30% on SST-2, respectively. Compared to Bert, Bert+Ours (exp.) increases the accuracy and F1 by 1.10% and 2.38% on SST1, and 2.22% and 2.21% on SST-2, respectively. This observation substantiates the effectiveness of WED as an emotional embedding in deep architecture.

5. Discussion

5.1. *Proportion of Non-dominant Emotion in Schema 1*

This section discusses the effect of different values of dominant emotion ϵ in pSED and weights of loss, λ. These two parameters control the extent to which non-dominant emotions in pSED are used during the training.

Using a smoothed pSED label helps provide auxiliary information to the one-hot label and prevents the network from becoming overfit, which is similar to the label smoothing technique. However, as described in Section 3.2, the

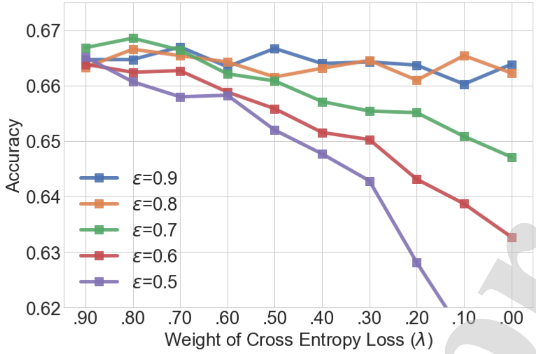


Figure 3: Visualization of experiment results on ISEAR dataset using different values of cross-entropy loss weight (λ) and dominant emotion value in pSED (ϵ).

proposed rule-based conversion is an unsupervised process that can be influenced by the language quality and sentence complexity. Therefore, the weights in the generated pSED can be undesirable and consequently introduce noise to 480 the classifier. To examine how such auxiliary information influences the model performance, we experimented with different values of dominant emotion intensity (ϵ) and cross-entropy loss weight (λ) on the IS dataset. The results are displayed in Figure 3. From the results, we can observe the following:

1. When the cross-entropy dominates the loss term (i.e., $\lambda = 0.9$), all models 485 can achieve remarkable results and the variance is marginal;
2. When the label emotion dominates the pSED (i.e., $\epsilon \geq 0.8$, the models fluctuate marginally as λ changes;
3. When the label emotion is less dominant in the pSED (i.e., $\epsilon \leq 0.7$), the model performance quickly deteriorates as λ becomes small.

490 We can conclude that the information on non-dominant emotions in pSED is beneficial to the classifier by providing auxiliary knowledge and alleviating overfitting. However, such information also introduces noise; hence, its proportion in the loss term must be restricted.

		I was misunderstood deliberately by a closest friend .						I was n't misunderstood deliberately by a closest friend .					
Emotions		#1	#2	#3	#4	#5	#6	#1	#2	#3	#4	#5	#6
SED (cat. lexicon)		.4	.2	.0	.2	.2	.0	.4	.2	.0	.2	.2	.0
misunderstood	.83	.20	.27	.29	.00	.38	.03	.01	.01	.00	.23	.01	.08
deliberate	1.	.08	.07	.03	.14	.03	.15	.06	.08	.06	.01	.23	.03
close	1.	.06	.10	.08	.01	.37	.01	.02	.01	.01	.32	.00	.18
SED (dim. lexicon)		.15	.21	.21	.01	.38	.04	.02	.04	.02	.61	.05	.27

Figure 4: Demonstration of two methods on sentences with and without negation words. Emotions #1 to #6 are Anger, Disgust, Fear, Joy, Sadness and Surprise. Values after affective words are constant c in Algorithm 1.

5.2. Dimensional Lexicon vs. Categorical Lexicon in Schema 1

Compared to the categorical-lexicon-based conversion strategy [22], the proposed method using WED has the following advantages.

1. WED can distinguish the intensities of the different emotions from the word-level, which makes using additional information, such as POS tag and polarity value, more effective;
2. A categorical-lexicon-based strategy cannot address the sentiment shifting caused by negation words. We applied both methods to similar sentences with and without negation words to demonstrate this point. Figure 4 illustrates that both the WED and SED generated by the proposed method are more reasonable compared to those generated using a categorical lexicon.
3. The SED generated by a categorical lexicon can lead to cavities (entries with 0s) in the generated distribution if a label of the dataset does not appear in any lexicon. We find that the cavity could bias the classifier during training.

Next, we will discuss the quality of the SED generated by the proposed rule-based method. One can consider the emotion with the greatest SED value as the predicted label. Remarkably, despite the bag-of-words nature and unsupervised process, the rule-based method achieved 38.64% accuracy on perfect match (ground-truth emotion), 48.42% accuracy on fuzzy match (similar to the

Sentence	**Label**	**Pred.**	
		Ours	Baseline
[Case #1] I was attacked by a teenage boy and had my wallet stolen.	Fear	Fear	Sad
[Case #2] I saw a friend of mine, whom I had not seen for a long time and I had lost his address and telephone number.	Joy	Joy	Sad
[Case #3] My boyfriend told me that it would be difficult for him to marry me.	Anger	Sad	Joy
[Case #4] When I failed an exam I thought I would pass.	Anger	Shame	Sad

Table 6: Case studies. Cases 1 and 2 are situations where sentences can be correctly classified by proposed method, yet are incorrectly classified by baseline models. Cases 3 and 4 are situations where proposed method fails to categorize correctly.

ground-truth emotion), and 60.52% accuracy on polarity classification (*i.e.*, positive or negative) on the IS dataset. Conversely, the result of an incremental process without rules is only 30%, and the result will be less than 27% if negations were not considered. The results indicate that the proposed rule-based method can capture over half of the instances' sentiment orientations.

5.3. Case Study on Schema 1

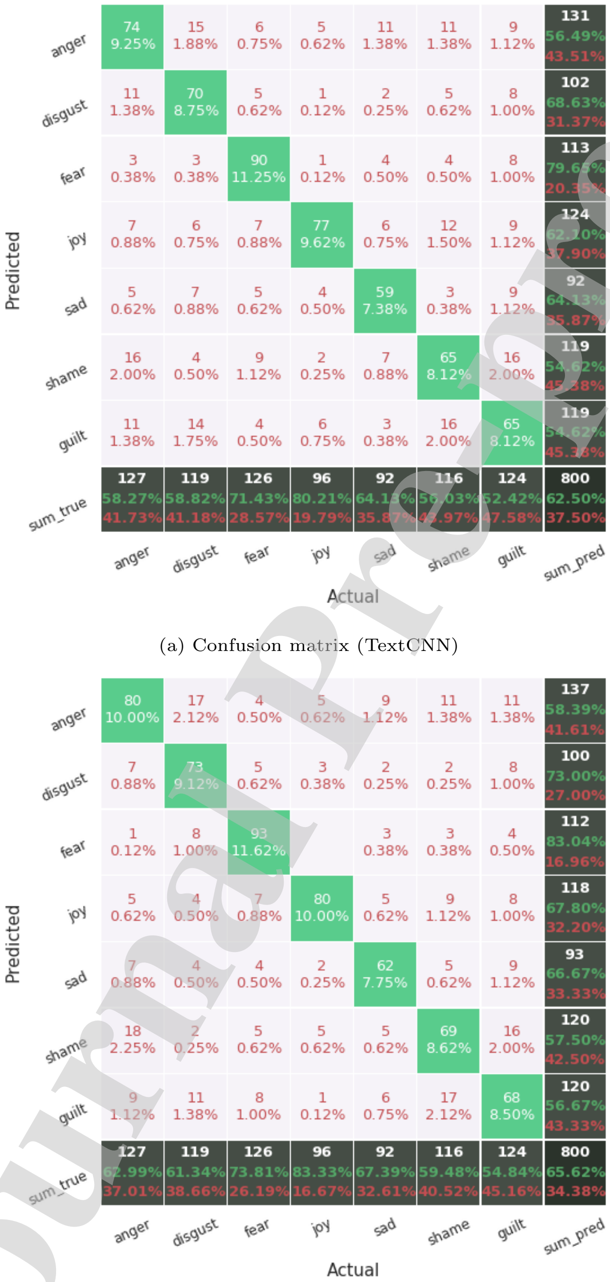
In this section, we provide a deeper insight into the effects of considering the intensities of non-dominant emotions. We present four cases of different situations in Table 6. The first two cases, i.e., Cases #1 and #2, are the sentences that the proposed method can correctly classify, yet the baseline model, a TextCNN model, fails to classify. We can observe that the proposed method is more effective in extracting the prominent emotion from the text than TextCNN, especially when a negation word appears. We can conclude that the reason behind this is that the intensities of the non-dominant emotions can help alleviate the overfitting problem during training and make the model more robust.

We can also observe that there are special situations that the proposed method cannot address, i.e., Cases #3 and #4. Although incorrect predictions were produced, the predicted emotions were reasonable and can be found in the sentences. For example, in Case #3, "sadness" can be felt when a lover refuses to marry and is a more suitable prediction than "joy", which was produced by TextCNN. In Case #4, multiple negative emotions can be identified from the sentence, and the sentence is labeled by one of them. In this case, the augmented label information can make the classification more difficult given similar candidate emotions; however, the predicted label is more acceptable than that output by training without augmented labels.

We display the confusion matrix of TextCNN and the proposed model on the IS dataset in Figure 5 to provide a detailed comparison. From a global perspective, we can observe that the proposed method offers improvements for each class, which validates the effectiveness of the proposed method on the classification task.

5.4. Emotion representation in Schema 2

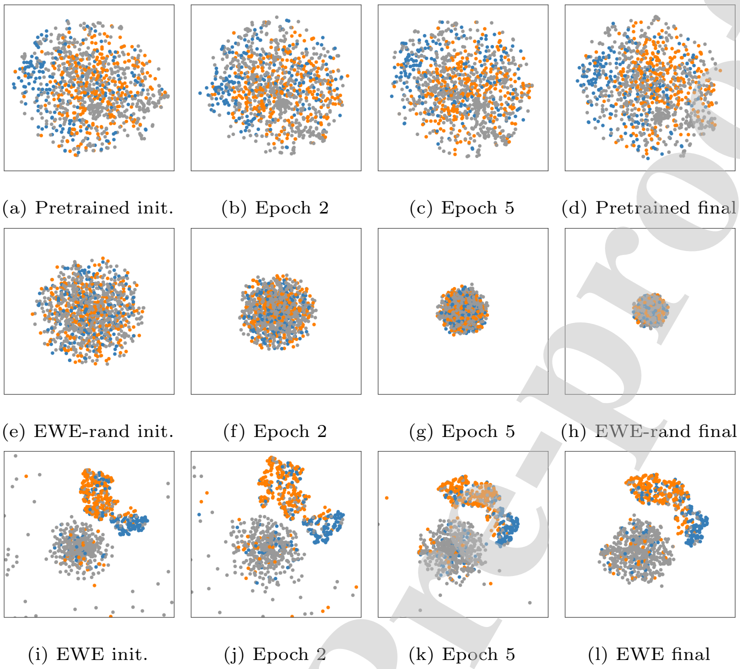
To analyze the learned emotion embedding, we used t-SNE to visualize the word representations at the beginning and the end of training in Figure 6. Visualized embeddings include pretrained FastText word embedding, EWE with random initialization, and EWE with WED initialization. Where affective words were labeled with more than one emotion in a lexicon, it was difficult to choose one as the label. Thus, we colored the dots according to the corresponding word's sentiment orientation in the NRC lexicon, with orange for positive, blue for negative, and gray for OOL (with initialized vectors). Both pretrained word embedding and EWE-rand demonstrated no apparent clustering during the training. In contrast, the dots from the EWE concentrate on forming ap-



(a) Confusion matrix (TextCNN)

(b) Confusion matrix (Ours)

Figure 5: Confusion matrix of (a) TextCNN and (b) proposed method on ISEAR dataset (one split).


(a) Pretrained init. (b) Epoch 2 (c) Epoch 5 (d) Pretrained final

(e) EWE-rand init. (f) Epoch 2 (g) Epoch 5 (h) EWE-rand final

(i) EWE init. (j) Epoch 2 (k) Epoch 5 (l) EWE final

Figure 6: t-SNE visualization of (a–d) word embedding vectors, (e–h) random initialized embedding vectors, and (i–l) proposed EWE vectors during training. Orange for positive polarity words, blue for negative polarity words, and gray for OOL words.

parent clusters. More concretely, we can identify three clusters based on theirs primary color: the gray cluster representing words without sentiment orientation, the orange cluster representing positive polarity words, and the blue cluster representing negative polarity, which means the polarity feature is encoded in the EWE and can explain why the proposed method can outperform the others on sentiment analysis tasks.

5.5. Computational complexity

For the implicit approach, the incremental process of generating a SED requires counting every affective word in a sentence and iterating through the sentence to identify negation tokens ("not" and "n't"). Thus, Schema 1 is of quadratic time, $O(n^2)$, where n is the number of words in a sentence. Nevertheless, because we focused on short-text classification tasks, the length of a sentence was relatively short. Moreover, we are only required to consider the affective words, which are limited in number. Therefore, this rule-based conversion method is manageable and feasible for applications.

For the explicit approach, we only compute the embedding vector for the affective words. Thus, Schema 2 is of linear time, $O(n)$, where n is the vocabulary size of the affective words.

6. Conclusion and Future Work

In this paper, we proposed an effective method to generate emotion-specific representation, the WED, for a word by modeling its relationship with emotions in the VAD space. We designed two novel schemas to implement WED. Extensive experiments on different tasks with CNN-based, RNN-based, Transformer-based, and Bert-based frameworks demonstrated the flexibility and effectiveness of WED in actual utilization. The WED generated by the proposed method was a more informative emotional representation than the existing categorical lexicon and could be applied to several relevant tasks. For future work, we will focus on (1) exploring algorithms to model more accurate term-emotion relationships; and (2) extracting more emotion-enriched features and enhancing model robustness on challenging datasets.

Acknowledgment

The work in this research has been supported by the Faculty Research Fund (102041) and the Lam Woo Research Fund (LWI20011) at Lingnan University, Hong Kong, the One-off Special Fund from Central and Faculty Fund in Support of Research from 2019/20 to 2021/22 (MIT02/19-20), the Research Cluster Fund (RG 78/2019-2020R), and the Interdisciplinary Research Scheme of the Dean's Research Fund 2019-20 (FLASS/DRF/IDS-2) of The Education University of Hong Kong.

References

[1] F. Xing, L. Malandri, Y. Zhang, E. Cambria, Financial sentiment analysis: An investigation into common mistakes and silver bullets, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, 2020, pp. 978–987.

[2] A. Khatua, A. Khatua, E. Cambria, Predicting political sentiments of voters from twitter in multi-party contexts, Applied Soft Computing 97 (2020) 106743. `doi:https://doi.org/10.1016/j.asoc.2020.106743`.

[3] M. S. Akhtar, A. Ekbal, E. Cambria, How intense are you? predicting intensities of emotions and sentiments using stacked ensemble [application notes], IEEE Computational Intelligence Magazine 15 (1) (2020) 64–75. `doi:10.1109/MCI.2019.2954667`.

[4] Z. Wang, S.-B. Ho, E. Cambria, A review of emotion sensing: Categorization models and algorithms, Multimedia Tools and Applications 79 (2020) 35553—-35582.

[5] M. E. Basiri, S. Nemati, M. Abdar, E. Cambria, U. R. Acharya, Abcdm: An attention-based bidirectional cnn-rnn deep model for sentiment analysis, Future Generation Computer Systems 115 (2021) 279–294. `doi:https://doi.org/10.1016/j.future.2020.08.005`.

[6] D. Peng, M. Zhou, C. Liu, J. Ai, Human–machine dialogue modelling with the fusion of word- and sentence-level emotions, Knowledge-Based Systems 192 (2020) 105319. `doi:https://doi.org/10.1016/j.knosys.2019.105319`.

[7] X. Li, S. Feng, D. Wang, Y. Zhang, Context-aware emotion cause analysis with multi-attention-based neural network, Knowledge-Based Systems 174 (2019) 205 – 218.

[8] T. Tuncer, S. Dogan, U. R. Acharya, Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques, Knowledge-Based Systems 211 (2021) 106547.

[9] P. Hajek, A. Barushka, M. Munk, Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining, Neural Computing and Applications 32 (2020) 17259–17274.

[10] X. Li, H. Xie, R. Y. K. Lau, T. Wong, F. L. Wang, Stock prediction via sentimental transfer learning, IEEE Access 6 (2018) 73110–73118.

[11] A. Picasso, S. Merello, Y. Ma, L. Oneto, E. Cambria, Technical analysis and sentiment embeddings for market trend prediction, Expert Systems with Applications 135 (2019) 60 – 70.

[12] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Comput. Linguist. 37 (2) (2011) 267—307.

[13] W. Li, L. Zhu, Y. Shi, K. Guo, E. Cambria, User reviews: Sentiment analysis using lexicon integrated two-channel cnn–lstm family models, Applied Soft Computing 94 (2020) 106435.

[14] G. Mohammadi, P. Vuilleumier, A multi-componential approach to emotion recognition and the effect of personality, IEEE Transactions on Affective Computing (2020) 1–1doi:10.1109/TAFFC.2020.3028109.

[15] S. M. Mohammad, P. D. Turney, Nrc emotion lexicon, National Research Council, Canada (2013) 1–234.

[16] E. Cambria, Y. Li, F. Z. Xing, S. Poria, K. Kwok, Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Association for Computing Machinery, 2020, pp. 105—114.

[17] P. Ekman, An argument for basic emotions, Cognition & emotion 6 (3-4) (1992) 169–200.

[18] J. A. Russell, A. Mehrabian, Evidence for a three-factor theory of emotions, Journal of research in Personality 11 (3) (1977) 273–294.

[19] S. M. Mohammad, Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2018, pp. 174–184.

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

[21] D. Zhou, X. Zhang, Y. Zhou, Q. Zhao, X. Geng, Emotion distribution learning from texts, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2016, pp. 638–647.

[22] Y. Zhang, J. Fu, D. She, Y. Zhang, S. Wang, J. Yang, Text emotion distribution learning via multi-task convolutional neural network, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, International Joint Conferences on Artificial Intelligence Organization, 2018, pp. 4595–4601.

[23] P. Ekkekakis, J. A. Russell, The Measurement of Affect, Mood, and Emotion: A Guide for Health-Behavioral Research, Cambridge University Press, 2013. doi:10.1017/CBO9780511820724.

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

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

[26] D. Tang, Z. Zhang, Y. He, C. Lin, D. Zhou, Hidden topic–emotion transition model for multi-level social emotion detection, Knowledge-Based Systems 164 (2019) 426 – 435.

[27] J. Yu, L. Marujo, J. Jiang, P. Karuturi, W. Brendel, Improving multi-label emotion classification via sentiment classification with dual attention transfer network, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2018, pp. 1097–1102.

[28] H. Fei, Y. Zhang, Y. Ren, D. Ji, Latent emotion memory for multi-label emotion classification, Proceedings of the AAAI Conference on Artificial Intelligence 34 (05) (2020) 7692–7699.

[29] H. Tanabe, T. Ogawa, T. Kobayashi, Y. Hayashi, Exploiting narrative context and a priori knowledge of categories in textual emotion classification, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, 2020, pp. 5535–5540.

[30] E. Cambria, D. Hazarika, S. Poria, A. Hussain, R. B. V. Subramanyam, Benchmarking multimodal sentiment analysis, in: A. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing, Springer International Publishing, 2018, pp. 166–179.

[31] Z. Teng, D.-T. Vo, Y. Zhang, Context-sensitive lexicon features for neural sentiment analysis, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2016, pp. 1629–1638.

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

[33] M. Dragoni, S. Poria, E. Cambria, Ontosenticnet: A commonsense ontology for sentiment analysis, IEEE Intelligent Systems 33 (3) (2018) 77–85.

[34] F. Z. Xing, F. Pallucchini, E. Cambria, Cognitive-inspired domain adaptation of sentiment lexicons, Information Processing & Management 56 (3) (2019) 554–564.

[35] M. M. Bradley, P. J. Lang, Affective norms for english words (anew): Instruction manual and affective ratings, Tech. rep., Citeseer (1999).

[36] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, Journal of Machine Learning Research 12 (76) (2011) 2493–2537.
URL http://jmlr.org/papers/v12/collobert11a.html

[37] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, A. Joulin, Advances in pre-training distributed word representations, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), 2018, pp. 52–55.
URL https://www.aclweb.org/anthology/L18-1008

[38] I. Labutov, H. Lipson, Re-embedding words, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, 2013, pp. 489–493.

[39] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, B. Qin, Sentiment embeddings with applications to sentiment analysis, IEEE Transactions on Knowledge and Data Engineering 28 (2) (2016) 496–509.

[40] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, S. Lehmann, Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2017, pp. 1615–1625.

[41] A. Agrawal, A. An, M. Papagelis, Learning emotion-enriched word representations, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, 2018, pp. 950–961.

[42] N. Babanejad, A. Agrawal, A. An, M. Papagelis, A comprehensive analysis of preprocessing for word representation learning in affective tasks, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 5799–5810.

[43] K. Moilanen, S. Pulman, Sentiment composition, in: Proceedings of the Recent Advances in Natural Language Processing International Conference, Association for Computational Linguistics, 2007, pp. 378–382.

[44] C. Strapparava, R. Mihalcea, SemEval-2007 task 14: Affective text, in: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Association for Computational Linguistics, 2007, pp. 70–74.

[45] K. R. Scherer, H. G. Wallbott, Evidence for universality and cultural variation of differential emotion response patterning, Journal of personality and social psychology 66 (2) (1994) 310–328.

[46] C. O. Alm, R. Sproat, Emotional sequencing and development in fairy tales, in: J. Tao, T. Tan, R. W. Picard (Eds.), Affective Computing and Intelligent Interaction, Springer Berlin Heidelberg, 2005, pp. 668–674. URL `https://doi.org/10.1007/11573548_86`

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

[48] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, C. Potts, Recursive deep models for semantic compositionality over a sentiment treebank, in: Proceedings of the 2013 conference on empirical methods in natural language processing, 2013, pp. 1631–1642.

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

[50] O. Udochukwu, Y. He, A rule-based approach to implicit emotion detection in text, in: C. Biemann, S. Handschuh, A. Freitas, F. Meziane, E. Métais (Eds.), Natural Language Processing and Information Systems, Springer International Publishing, 2015, pp. 197–203.

[51] C. T. Yang, Y. L. Chen, Dacnn: Dynamic weighted attention with multichannel convolutional neural network for emotion recognition, in: 2020 21st IEEE International Conference on Mobile Data Management (MDM), 2020, pp. 316–321.

[52] S. D. Gollapalli, P. Rozenshtein, S.-K. Ng, ESTeR: Combining word cooccurrences and word associations for unsupervised emotion detection, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, 2020, pp. 1043–1056.

[53] J. Islam, R. E. Mercer, L. Xiao, Multi-channel convolutional neural network for Twitter emotion and sentiment recognition, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 1355–1365.

[54] C. Wang, B. Wang, W. Xiang, M. Xu, Encoding syntactic dependency and topical information for social emotion classification, in: Proceedings of the 42nd International ACM SIGIR Conference, SIGIR'19, Association for Computing Machinery, 2019, pp. 881—-884.

[55] B. Raad, B. Philipp, H. Patrick, M. Christoph, Aseds: Towards automatic social emotion detection system using facebook reactions, in: 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), IEEE Computer Society, 2018, pp. 860–866.

[56] C. Wang, B. Wang, An end-to-end topic-enhanced self-attention network for social emotion classification, in: Proceedings of The Web Conference 2020, WWW '20, Association for Computing Machinery, 2020, p. 2210–2219.

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

[58] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2014, pp. 1746–1751.

[59] R. Johnson, T. Zhang, Deep pyramid convolutional neural networks for text categorization, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2017, pp. 562–570.

[60] A. Graves, N. Jaitly, A. Mohamed, Hybrid speech recognition with deep bidirectional lstm, in: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, 2013, pp. 273–278. doi:10.1109/ASRU.2013.6707742.

[61] C. Zhou, C. Sun, Z. Liu, F. C. M. Lau, A c-lstm neural network for text classification, ArXiv.
URL https://arxiv.org/abs/1511.08630

[62] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, 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, Curran Associates, Inc., 2017, pp. 5998–6008.

[63] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186.

CRediT author statement

Zongxi Li: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization. **Haoran Xie:** Conceptualization, Methodology, Validation, Formal Analysis, Writing – Review & Editing, Visualization. **Gary Cheng:** Software, Validation, Formal Analysis, Data Curation. **Qing Li:** Visualization, Supervision, Project Administration, Funding Acquisition.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: