Distributed Representations of Emotion Categories in Emotion Space

Xiangyu Wang[1,2] and **Chengqing Zong**[1,2,3*]
[1]National Laboratory of Pattern Recognition, Institute of Automation, CAS
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]CAS Center for Excellence in Brain Science and Intelligence Technology
{xiangyu.wang, cqzong}@nlpr.ia.ac.cn

Abstract

Emotion category is usually divided into different ones by human beings, but it is indeed difficult to clearly distinguish and define the boundaries between different emotion categories. The existing studies working on emotion detection usually focus on how to improve the performance of model prediction, in which emotions are represented with one-hot vectors. However, emotion relations are ignored in one-hot representations. In this article, we first propose a general framework to learn the distributed representations for emotion categories in emotion space from a given emotion classification dataset. Furthermore, based on the soft labels predicted by the pre-trained neural network model, we derive a simple and effective algorithm. Experiments have validated that the proposed representations in emotion space can express emotion relations much better than word vectors in semantic space.

1 Introduction

In the past decades, a lot of tasks have been proposed in the field of text emotion analysis. The most primary one among them is emotion classification task (Alm et al., 2005). Based on emotion classification task, many new tasks have been proposed from different considerations. Lee et al. (2010) proposed the task of emotion cause extraction, which aims at predicting the reason of a given emotion in a document. Based on the emotion cause extraction task, Xia and Ding (2019) introduced the emotion-cause pair extraction task for the purpose of extracting the potential pairs of emotions and corresponding causes in a document. Jiang et al. (2011) proposed a target-dependent emotion recognition task, which aims at predicting the sentiment with the given query. To express the intensity of a specific emotion in text, Mohammad and Bravo-Marquez (2017) proposed the emotion intensity detection task. However, all the above tasks treat emotions as independent ones and represent emotions with one-hot vectors, which definitely ignore the underlying emotion relations.

Based on existing emotion detection tasks, many efforts have been made to achieve better performance (Danisman and Alpkocak, 2008; Xia et al., 2011; Kim, 2014; Xia et al., 2015; Li et al., 2018; Zong et al., 2019) and many datasets have been introduced to train and evaluate the corresponding models (Ghazi et al., 2015; Mohammad et al., 2018; Liu et al., 2019). The vast majority of existing emotion annotation work assumes that the emotions are orthogonal to each other and represent the emotion categories with one-hot vectors (Mohammad, 2012; Gui et al., 2016; Klinger et al., 2018). Actually, the boundaries as well as the relations among emotion categories are not clearly distinguished and defined.

Typical word embedding learning algorithms only use the contexts but ignore the sentiment of texts (Turian et al., 2010; Mikolov et al., 2013). To encode emotional information into word embedding, sentiment embedding and emotion(al) embedding have been proposed (Tang et al., 2014; Yu et al., 2017; Xu et al., 2018). Tang et al. (2015) proposed a learning algorithm dubbed sentiment-specific word embedding (SSWE). Agrawal et al. (2018) proposed a method to learn emotion-enriched word embedding (EWE). However, all the above algorithms represent emotions in semantic space rather than emotion space. As shown in Table 1, each emotion category represented in semantic space reflect a piece of semantic information rather than a specific emotional state. In this work, we regard each emotion category as a specific emotional state in emotion space and represent each emotion category with a point in emotion

[*] Corresponding author.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 2364–2375
August 1–6, 2021. ©2021 Association for Computational Linguistics

Semantic Space	**Emotion Space**
Each word corresponds to a point in semantic space.	Words cannot be represented in emotion space.
Emotional states cannot be represented in semantic space.	Each emotional state corresponds to a point in emotion space.
Each emotion category is encoded with a piece of specific semantic information.	Each emotion category is encoded with a specific emotional state.

Table 1: Differences between semantic space and emotion space.

space. The further experiments show that our representations in emotion space can express emotion relations much better than word vectors in semantic space.

From the perspective of psychology, some studies have discussed the complexity of the human emotional state (Russell, 1980; Griffiths, 2002; Fontaine et al., 2007; Clark, 2010) and the shared psychological features across emotions (Fehr and Russell, 1984; Mauss and Robinson, 2009; Campos et al., 2013). However, psychological researches mainly focus on the human emotional state itself and do not pay attention to emotion relations hidden in the text. As there are lots of emotion detection tasks and corresponding datasets in NLP field, it is very meaningful to investigate what is the relations among emotion categories hidden in corpora. In this paper, we detect the underlying relations among emotion categories labeled in corpora from the perspective of NLP.

Distributed representations of emotion categories in emotion space can also benefit NLP applications. Take depression recognition for example, depression is a serious mood disorder and manifested by a complex emotional state (Blatt, 2004; Beck et al., 2014). Most existing emotion taxonomies or datasets do not contain depression as a specific category. In this article, we generate the latent encoding for each emotion category. Based on the psychological researches (Rottenberg, 2005; Joormann and Stanton, 2016) on relations between depression and existing emotion categories, we can predict the distributed representations of depression in the text even if there are no samples annotated as depression.

The main contributions of this work are summarized as follows:

- A general framework to learn distributed emotion representations from an emotion classification dataset is first proposed. Based on soft labels predicted by the pre-trained neural network model, a simple and effective approach is derived. As far as we know, this is the first work to learn the distributed representations for emotion categories in emotion space rather than semantic space.

- Experiments have been conducted to validate the effectiveness of our emotion representations. The results have shown that our emotion representations in emotion space can express emotion relations much better than word vectors, and is competitive with human results.

- Emotion similarities across datasets have been detected to validate the quality of our emotion representations across corpora. The results have shown the good consistency of our representations in emotion similarities across datasets although they are created for a variety of domains and applications.

2 Related Work

Emotion Taxonomy: The existing studies on emotion taxonomy usually divide emotion space into specific emotion categories. Ekman (1992) classified emotions into six discrete states (*anger*, *disgust*, *fear*, *joy*, *sadness* and *surprise*), which are contained in vast majority of the existing emotion classification datasets. With the discrete emotion questionnaire method, Harmon-Jones et al. (2016) captured eight distinct state emotions in their study: *anger*, *disgust*, *fear*, *anxiety*, *sadness*, *happiness*, *relaxation*, and *desire*. Similarly, Cowen and Keltner (2017) introduced a conceptual framework to analyze reported emotional states and elicited 27 distinct varieties of reported emotional experience. However, above work only gives the basic emotions of human emotional state from a psychological perspective. The quantitative relations among basic emotions remain to be detected. In this work, emotion relations are quantitatively revealed based on our emotion representations.

Emotion Datasets: Strapparava and Mihalcea (2007) introduced first emotion recognition dataset, Affective Text, in the domain of news headlines. After that, many emotion datasets that vary in domain, size and taxonomy have been developed. Wang et al. (2012) automatically created a large emotion-labeled dataset (of about 2.5 million tweets) by harnessing emotion-related hashtags available in the tweets. Abdul-Mageed and Ungar (2017) introduced a fine-grained dataset with up to 24 types of emotion categories with Twitter data. Li et al. (2017) developed a multi-turn dialog dataset, DailyDialog, for detecting the emotions in the field of dialog systems. Öhman et al. (2018) presented a multi-dimensional emotion dataset with annotations in movie subtitles for the purpose of creating a robust multilingual emotion detection tool. Demszky et al. (2020) built a manually dataset with up to 27 fine-grained emotion categories on Reddit comments for emotion prediction. However, all above datasets are annotated with discrete basic emotion categories, which means the emotion categories are represented with one-hot vectors. One-hot representations ignore the underlying relations among emotion categories. In this work, the underlying emotion relations contained in the datasets are revealed with our emotion representations.

Soft Labels: Hinton et al. (2015) observed that it is easier to train classifier using the soft targets output by trained classifier as target values than using manual ground-truth labels. Phuong and Lampert (2019) provided their insights into the working mechanisms of distillation by studying the special case of linear and deep linear classifiers. Szegedy et al. (2016) proposed a label smoothing mechanism for the purpose of encouraging the model to be less confident by smoothing the initial one-hot labels. Imani and White (2018) investigated the reasons for the improvement of the model performance by converting hard targets to soft labels in supervised learning. Zhao et al. (2020) proposed a robust training method for machine reading comprehension by learning soft labels. In this work, soft labels output by the trained neural network model are used to generate distributed representations for emotion categories.

3 Methodology

In this section, we describe how to learn the distributed representations for emotion categories. First, a general framework is proposed. Then, a simple and effective algorithm is derived based on the soft labels from a pre-trained neural network model. After that, we extend our method to multi-label datasets. At last, detailed approaches of the algorithm are listed.

3.1 The General Framework

As shown in Table 2, the four instances from dataset SemEval-2007 task 14 (Strapparava and Mihalcea, 2007) are annotated with both emotion categories and valence values. Although both instance 1 and instance 2 are labeled with *joy* category, their valence values are very different, which means there is a big difference between their emotional states. Actually, emotions in instance 1 seem to be more *excited* while emotions in instance 2 seem to be more *hopeful*. On the other hand, instances 3 and 4 are annotated with the same valence value while they are divided into different categories. Fontaine et al. (2007) also find that emotional state is high-dimensional and valence-arousal-dominance representation model is not sufficient to describe the emotional state.

The above examples show emotional states contained in different documents, even if they are annotated with the same emotion category or valence value, are not exactly the same. In this work, we regard text emotional states as an emotion space. The emotion contained in a specific document corresponds to a specific emotional state, further corresponds to a point in the space. As a result, documents annotated with same emotion category probably correspond to different emotional states and points in the space, which means the emotion category is a random variable rather than a specific vector in the space.

For category K, we define x as the sample annotated with category K and $\boldsymbol{V}_K$ as the specific distributed representations of category K. Let $\boldsymbol{\mathcal{V}}(x)$ be the distributed representations of sample x and $p(x)$ be the probability density of sample x. Let Ω be the integral domain of x. We further use $\mathcal{L}(\boldsymbol{V}_K, \boldsymbol{\mathcal{V}}(x))$ as the distance function between $\boldsymbol{V}_K$ and $\boldsymbol{\mathcal{V}}(x)$. In order to obtain a better distributed representation for category K, we must minimize the expectation of $\mathcal{L}$. Thus, we obtain the calculation formula for specific distributed representation of category K as the following:

$$\boldsymbol{V}_K = \arg\min_{\boldsymbol{V}} \int_{\Omega} \mathcal{L}(\boldsymbol{V}, \boldsymbol{\mathcal{V}}(x)) p(x) dx. \quad (1)$$

Index	Instances	Emotion	Valence
1	Goal delight for Sheva	joy	87
2	Making peace from victory over poverty	joy	39
3	New Indonesia Calamity, a Mud Bath, Is Man-Made	anger	-59
4	Waste plant fire forces 5,000 to evacuate	sadness	-59

Table 2: Four instances in dataset AffectiveText.

3.2 A Simple Method

Although we can not directly obtain the strict probability distribution of each emotion category in emotion space, there are many available emotion classification dataset, in which the instances can be regarded as samples of the corresponding annotated emotion categories.

For emotion dataset $\mathcal{D}$ and emotion category K, we use all samples annotated as category K in the dataset to estimate the distribution of category K. Thus, we can rewrite formula 1 as:

$$\boldsymbol{V}_K = \arg\min_{\boldsymbol{V}} \sum_{x \in S_K} \mathcal{L}(\boldsymbol{V}, \boldsymbol{\mathcal{V}}(x)), \tag{2}$$

where S_K is the set of all instances labeled with category K in dataset $\mathcal{D}$.

In this paper, we use squared Euclidean distance as the distance metric between two representations. Therefore, formula 2 can be simplified as follows:

$$\boldsymbol{V}_K = \arg\min_{\boldsymbol{V}} \sum_{x \in S_K} ||\boldsymbol{V} - \boldsymbol{\mathcal{V}}(x)||_2^2. \tag{3}$$

By solving formula 3, we have:

$$\boldsymbol{V}_K = \frac{\sum_{x \in S_K} \boldsymbol{\mathcal{V}}(x)}{N_K}, \tag{4}$$

where N_K is the size of S_K.

Since then we have derived that the distributed representation of emotion category K is exactly the average of the distributed representation of all instances labeled as category K in dataset $\mathcal{D}$.

Now, let's discuss how to obtain the distributed representation for the instances in the dataset. As shown in Figure 1, the output of the neural network model is a soft label regardless of the specific architecture of the model. It has been verified that soft labels output by the trained model tend to have higher entropy and contain more information than manual one-hot labels (Hinton et al., 2015; Phuong and Lampert, 2019). Inspired by previous work on soft labels, we directly take the soft labels output by the trained neural network model as the distributed representation of the input instance. As a result, the dimension of $\boldsymbol{V}_K$ is equal to the number of categories annotated in dataset $\mathcal{D}$.

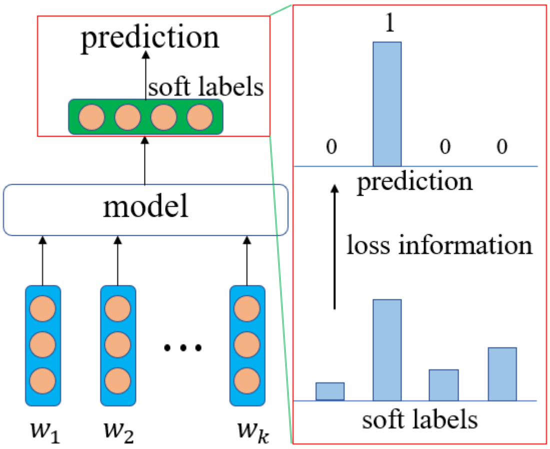


Figure 1: Schematic diagram of emotion classification models based on word vectors.

We define soft labels output by the trained neural network model of the input instance x as $\boldsymbol{f}(x)$. Thus, we derive a simple method to calculate the specific distributed representation for category K:

$$\boldsymbol{V}_K = \frac{\sum_{x \in S_K} \boldsymbol{f}(x)}{N_K}. \tag{5}$$

3.3 How to Deal with Multilabel Data?

In some corpora, instances are annotated with multiple emotion categories (Strapparava and Mihalcea, 2007; Demszky et al., 2020). To deal with multilabel instances, we regard each multilabel instance as multiple single label instances with weights summing to 1, and the weight of each single label data is set to the reciprocal of the number of the annotated labels. For example, suppose document D is labeled with category A and B. We regard D as two half instances, one half is labeled with category A and the other half is labeled with category B.

Let $\mathcal{Y}(x)$ denote the set of the annotated labels of sample x and $|\mathcal{Y}(x)|$ denote the size of set $\mathcal{Y}(x)$. Take above document D as an example, then $\mathcal{Y}(D)$ is equal to $\{A, B\}$ and $|\mathcal{Y}(D)|$ is equal to 2 as

Positive(P):	admiration, amusement, approval, caring, desire, excitement, gratitude, joy, love, optimism, pride, relief
Negative(N):	anger, annoyance, disappointment, disapproval, disgust, embarrassment, fear, grief, nervousness, remorse, sadness
Ambiguous(A):	confusion, curiosity, realization, surprise

Table 3: Artificial classification results of 27 emotion categories by the creators of GoEmotions.

there are two labels contained in $\mathcal{Y}(D)$. Therefore, we obtain the calculation formula of specific distributed representation for category K:

$$\boldsymbol{V}_K = \frac{\sum_{x \in S_K} w_K(x) \boldsymbol{f}(x)}{\sum_{x \in S_K} w_K(x)}, \quad (6)$$

where $w_K(x)$ is equal to $1/|\mathcal{Y}(x)|$, which is the weight of instance x in category K,

3.4 Algorithm

In this part, we describe the algorithm of learning the **D**istributed **R**epresentations for **E**motion **C**ategories (DREC). First, go through every instance in the dataset, and calculate the total weight and weighted sum of soft labels output by the trained model for each category. Then, the weighted sum is divided by the total weight to obtain the final distributed representation for each emotion category. The detailed approaches are stated in Algorithm 1.

Algorithm 1 DREC

```
Input: $\mathcal{D} = \{(\mathcal{T}^{(n)}, \mathcal{Y}^{(n)})_{n=1}^{N}\}$   // dataset
Output: $\boldsymbol{V} = \{\boldsymbol{V}_1, \boldsymbol{V}_2, ..., \boldsymbol{V}_C\}$
  // distributed representations for emotions
$\boldsymbol{f} \leftarrow \mathcal{D}$   // train a neural network model
$\boldsymbol{V} \leftarrow \{\mathbf{0}, \mathbf{0}, ..., \mathbf{0}\}$
$\{W_1, W_2, ..., W_C\} \leftarrow \{0, 0, ..., 0\}$ // weight
for $n = 1$ to $N$ do
    for each $j \in \mathcal{Y}^{(n)}$ do
        $\boldsymbol{SL} \leftarrow \boldsymbol{f}(\mathcal{T}^{(n)})$ // soft labels
        $\boldsymbol{V}_j \leftarrow \boldsymbol{V}_j + \boldsymbol{SL}/|\mathcal{Y}^{(n)}|$
        $W_j \leftarrow W_j + 1/|\mathcal{Y}^{(n)}|$
    end for
end for
for $i = 1$ to $C$ do
    $\boldsymbol{V}_i \leftarrow \boldsymbol{V}_i / W_i$
end for
```

4 Experiments

In order to validate the intrinsic quality of our emotion representations, we conducted three experiments in this section. First of all, *arrangement* experiment is conducted to show the emotion distribution. Then, relations between different emotion taxonomies are detected in *mapping* experiment. At last, the emotion representations extracted from various corpora are compared to show the consistency of our approach across corpora.

4.1 Datasets

There are four datasets we use to detect emotion relations. The detailed information of each dataset is described as follows:

GoEmotions: GoEmotions is annotated of 58k English Reddit comments extracted from popular English subreddits (Demszky et al., 2020), multilabeled for 27 emotion categories, which is proposed by Cowen and Keltner (2017). GoEmotions is created for the purpose of building a large dataset with a large number of positive, negative, and ambiguous emotion categories. The detailed emotion categories are shown in Table 3.

AffectiveText: AffectiveText consists of 1250 instances on the domain of news headlines (Strapparava and Mihalcea, 2007). The dataset is multilabel annotated. There are six emotion categories (*anger*, *disgust*, *fear*, *joy*, *sadness* and *surprise*) and valence contained in the dataset.

ISEAR: ISEAR is created from questionnaires by Scherer and Wallbott (1994). Each instance is annotated with only one label. There are seven emotion categories contained in ISEAR: *anger*, *disgust*, *fear*, *guilt*, *joy*, *sadness*, and *shame*.

Affect in Tweets: “Affect in Tweets” is created from tweets (Mohammad et al., 2018). There are ten emotions contained in “Affect in Tweets”: *anger*, *anticipation*, *disgust*, *fear*, *joy*, *love*, *optimism*, *pessimism*, *sadness*, *surprise*, and *trust*.

GoEmotions is used to conduct the first two experiments (*arrangement* and *mapping*), and the above four datasets are used to validate our representations across corpora in last experiment.

4.2 Model Settings

Any model that outputs are soft labels can be employed to learn the distributed representations for emotion categories. In our experiments, TextCNN (Kim, 2014), BiLSTM (Schuster and Paliwal, 1997) and BERT (Devlin et al., 2019) are used as the training models. For comparison, experiments on word embedding learning algorithms are conducted to show emotion relations in semantic space. For a specific emotion category, we use its word embedding as its representations in semantic space. 100-dimensional GloVe (Pennington et al., 2014) is the word vectors used in TextCNN and BiLSTM. The detailed model settings are listed as follows:

TextCNN: The height of convolutional kernel size is divided into three groups (3,4,5) and the width is 100, which is equal to the dimension of the word vectors. There are 32 channels in each group. Batch size and learning rate are set to 16 and 0.001.

BiLSTM: There is only one layer in this model. Batch size and learning rate are set to 16 and 0.001 separately, which are the same as for TextCNN. There are 32 neurons in the hidden layer in each direction.

BERT: BERT-based model is used in this experiment. A fully connected layer is added on top of the pre-trained model. Batch size and learning rate are separately set to 8 and 2e-5 for fine-tuning.

4.3 Arrangement

As shown in Table 3, the emotion categories are divided into three groups corresponding to the positive, negative, and ambiguous emotions, which are divided by the creators of GoEmotions[1] (Demszky et al., 2020).

We conduct the experiments 10 times with same model and different initial parameters, and the average representations are employed to show the following results. After final emotion representations obtained, to better understand the arrangement of emotion categories in emotion space, we reduce the dimension of the emotion representations to two with singular value decomposition (Wall et al., 2003). The two-dimensional average vectors are displayed as shown in Figure 2. Three color-shape pairs, red-circle, gray-square and black-triangle, correspond to positive, negative and ambiguous emotions respectively. Figure 2 (a)-(c) correspond to the results of word representations in semantic space. Figure 2 (d)-(f) show the results of TextCNN, BiLSTM and BERT in emotion space.

As shown in Figure 2 (a)-(c), the results of three word embedding algorithms (GloVe, SSWE and EWE) are displayed. We can find that the word vectors of emotion terms are displayed relatively random in semantic space and there are no clear linear boundaries among positive, negative and ambiguous emotions.

As shown in Figure 2 (d)-(f), it can be found that in emotion space, regardless of the constructed model, there are obvious boundaries among positive, negative and ambiguous emotions. The two blue dashed lines separate each type of emotion category from the others, which means that different types of emotion categories are linearly separable from each other in emotion space. The ambiguous emotions are just located between positive and negative emotions in Figure 2 (d)-(f), which shows our representations in emotion space can better describe the relative relation between ambiguous emotions and the others. In addition, the arrangement of emotions in Figure 2 (d) and (e) are very similar, which means TextCNN and BiLSTM have similar emotion relation extraction capabilities.

From this experiment, we can conclude that similar emotions are more likely to get together in emotion space than in semantic space, which further demonstrates that our representations can express emotion relations much better than word vectors.

4.4 Mapping

Demszky et al. (2020) manually mapped these 27 emotion categories to Ekman's basic emotions (Ekman, 1992).[2] In this experiment, we automatically generate these mapping relations based on the proposed distribution representations of emotion categories.

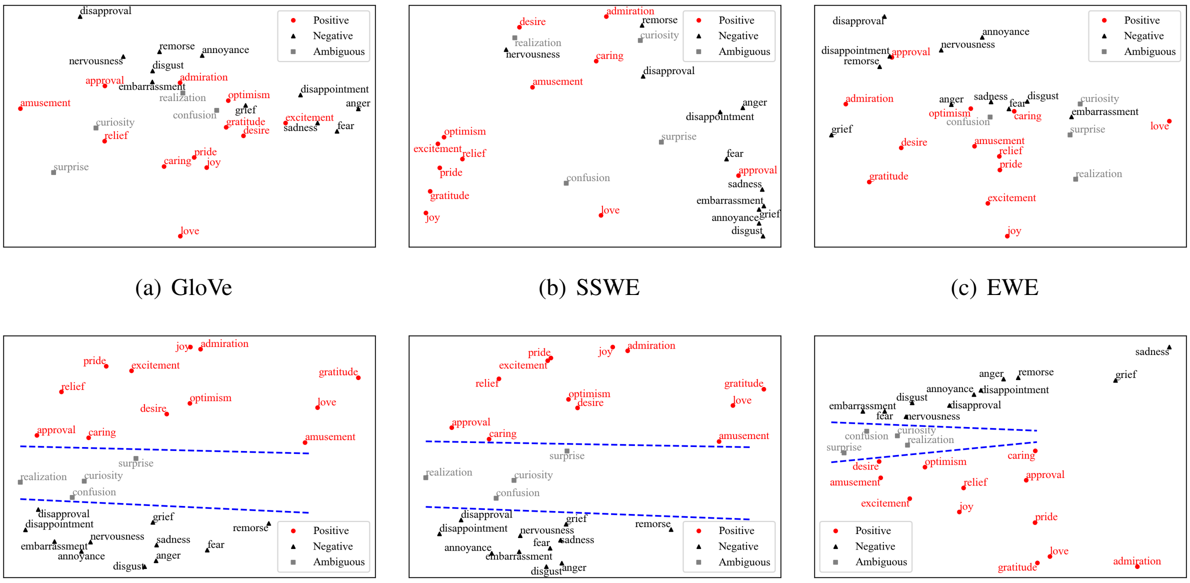
In this experiment, we take Ekman's basic emotions as target emotions and the remaining 21 categories as source emotions. For each source emotion, we select the most similar one from the target emotions as its mapping result. The calculation formula is listed as follows:

$$e = \arg\max_{e_t} \mathrm{sim}(e_s, e_t), \tag{7}$$

where e_t is the emotion category in target emotions, e_s is the emotion category in source emotions and

[1] https://github.com/google-research/google-research/tree/master/goemotions/data/sentiment_mapping.json

[2] https://github.com/google-research/google-research/tree/master/goemotions/data/ekman_mapping.json



(a) GloVe

(b) SSWE

(c) EWE

(d) TextCNN

(e) BiLSTM

(f) BERT

Figure 2: Visualization of emotion vectors in different spaces. (a)-(c) In semantic space, there are no linear boundaries among positive, negative and ambiguous emotions. (GloVe: global vectors for word representation (Pennington et al., 2014); SSWE: sentiment-specific word embedding (Tang et al., 2015); EWE: emotion-enriched word embedding (Agrawal et al., 2018).) (d)-(f) In emotion space, each type of emotions is linear separated with the others by blue lines.

e is the mapping result of e_s. sim is the similarity function and the cosine similarity is selected here.

The emotion representations are calculated 10 times with same model and different initial parameters and the average results are employed to conduct this experiment. Table 4 shows the mapping results with different models. We also calculate the results of word vectors for comparison. Manual results are chosen as the gold answers. GloVe correctly maps 3 out of 21 emotions, which is comparable to a random result. By encoding emotional information into word representations, SSWE (Tang et al., 2015) maps 10 emotions correctly and EWE (Agrawal et al., 2018) maps 7 emotions correctly. The results indicate that although sentiment embedding (SSWE) and emotion embedding (EWE) map more emotions correctly than typical word embedding (GloVe), SSWE and EWE still mismatch more than half of the source emotions as they are constructed under semantic space.

In emotion space, our emotion representations correctly map 18 out of 21 emotions, which is much better than the result in semantic space. The scores undoubtedly show that our emotion representations can describe emotion relations much better than word vectors. Besides, detailed mapping results for each emotion can be seen in Table 4. Results of TextCNN and BiLSTM are exactly the same, which is consistent with their similar arrangement in emotion space in first experiment. BERT maps *disapproval* to *disgust* while the others map it to *anger*. The most confusing emotions are *caring* and *embarrassment*, human maps them to *joy* and *sadness* respectively, while our representations in emotion space map them to *sadness* and *disgust*.

The inconsistency of the two emotions (*embarrassment* and *caring*) in emotion space and in human results shows the complexity of emotion relations. Existing psychological study (Scherer, 2005) shows that *embarrassment* is close to both *sadness* and *disgust*, which means *sadness* and *disgust* can both be regarded as the mapping result for *embarrassment*. As for *caring*, it has been discussed (Scherer et al., 2013) that *caring* is a positive emotion in nature but accompanied by the occurrence of negative events.

The mapping results of the three models are roughly the same as human-provided mapping results, which shows our emotion representations are effective. However, when a certain emotion has high similarities to multiple emotions (such as *embarrassment* to *disgust* and *sadness*), there may exist some differences between different mapping results. In other words, there are no absolutely cor-

Source Emotions	Human	Semantic Space			Emotion Space		
		GloVe	SSWE	EWE	TextCNN	BiLSTM	BERT
admiration	joy	disgust	anger	anger	joy	joy	joy
amusement	joy	anger	joy	disgust	joy	joy	joy
annoyance	anger	anger	disgust	anger	anger	anger	anger
approval	joy	fear	disgust	fear	surprise	surprise	joy
caring	joy	anger	anger	anger	sadness	sadness	sadness
confusion	surprise	anger	joy	anger	surprise	surprise	surprise
curiosity	surprise	fear	surprise	surprise	surprise	surprise	surprise
desire	joy	fear	joy	joy	joy	joy	joy
disappointment	sadness	fear	fear	anger	sadness	sadness	sadness
disapproval	anger	disgust	anger	disgust	anger	anger	disgust
embarrassment	sadness	disgust	sadness	fear	disgust	disgust	disgust
excitement	joy	anger	joy	joy	joy	joy	joy
gratitude	joy	joy	joy	joy	joy	joy	joy
grief	sadness	anger	disgust	sadness	sadness	sadness	sadness
love	joy	joy	surprise	surprise	joy	joy	joy
nervousness	fear	anger	joy	sadness	fear	fear	fear
optimism	joy	anger	joy	anger	joy	joy	joy
pride	joy	anger	joy	anger	joy	joy	joy
realization	surprise	sadness	joy	joy	surprise	surprise	surprise
relief	joy	anger	joy	anger	joy	joy	joy
remorse	sadness	disgust	anger	sadness	sadness	sadness	sadness
Score	—	3	10	7	**18**	**18**	**18**

Table 4: The results of mapping Cowen taxonomy to Ekman taxonomy. Human results are chosen as the gold answers and wrong results are marked in red.

rect mapping results for all emotions, which further indicates the relations among emotions are indeed complex.

4.5 Emotion Relations across Corpora

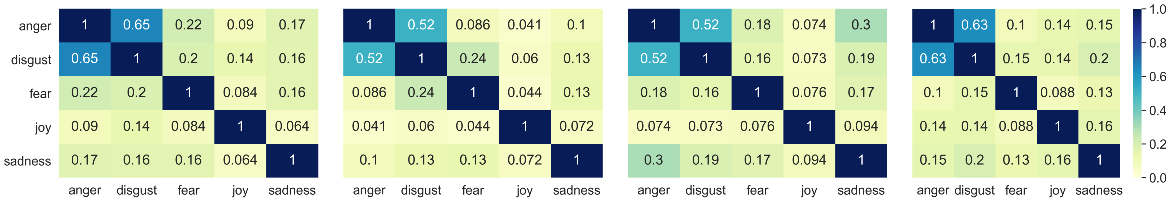
Due to the deviations in different corpora (such as data source bias and annotation bias), there may exist some differences in emotion relations between different corpora. In this part, we analyze the difference in emotion relations across corpora. BERT is chosen as the training model here to eliminate the potential impact caused by models. For each dataset, the experiments are repeated 10 times with same model and different initial parameters, and the average results are reported here.

There are five emotion categories (*anger*, *disgust*, *fear*, *joy* and *sadness*) shared in the four datasets. The shared five emotions are basic emotion categories in many emotion taxonomy theories (Ekman, 1992; Harmon-Jones et al., 2016; Cowen and Keltner, 2017). As a result, the cosine similarities among these emotion categories as shown in Figure 3 are not high. For each dataset, all cosine similarities are not greater than 0.3 except the similarity between *anger* and *disgust*.

On the other hand, the datasets are created based on different annotation standards from different domains. Thus, for specific emotion pair, the similarities across datasets may be quite different. However, the relative magnitude of similarities is consistent across datasets. For each dataset, there is a moderate similarity between *anger* and *disgust* (ranging from 0.52 to 0.65) while the similarities among remaining emotion pairs are relatively small (ranging from 0.04 to 0.30).

In order to quantitatively measure the consistency of emotion relations in different datasets, Pearson correlation coefficients between cosine similarities across datasets are calculated as shown in Table 5. The Pearson correlation coefficients among datasets are pretty high (ranging from 0.867 to 0.949), which indicates the underlying emotion relations are quite similar across datasets even if they are created in different domains.

In this experiment, we detect emotion relations across corpora. The results reveal that there is a



(a) AffectiveText (b) GoEmotions (c) ISEAR (d) Affect in Tweets

Figure 3: Cosine similarities among emotions in different datasets.

	A	G	I	T
A	1.000	0.949	0.917	0.936
G	0.949	1.000	0.873	0.926
I	0.917	0.873	1.000	0.867
T	0.936	0.926	0.867	1.000

Table 5: Pearson correlation coefficients between cosine similarities. (A: AffectiveText; G: GoEmotions; I: ISEAR; T: Affect in Tweets.)

good consistency of our emotion representations across datasets even if they are created on the basis of different annotation standards from different domains.

5 Conclusion and Future Work

In this paper, we argued that the emotion categories are not orthogonal to each other and the relations among emotion categories are very complex. We proposed a general framework to learn the distributed representation for each emotion category in emotion space from a given emotion dataset. Then, a simple and effective algorithm was also derived based on the soft labels predicted by the pre-trained neural network model. We conducted three experiments to validate the effectiveness of our emotion representations and the experimental results demonstrated that our representations in emotion space can express emotion relations much better than representations from word embeddings.

There are three avenues of future work we would like to explore. First, the distributed representations for emotion categories are derived from a specific emotion classification dataset. It would be interesting to build a universal emotion representation that is irrelevant to a specific corpus. Second, the computation of our emotion representations relies on the soft labels predicted by the neural network model, and we would like to investigate a more general method in the future. Finally, we would like to explore more NLP applications of our emotion representations, such as improving the performance of emotion classification models and studying emotion spaces across languages.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 718–728.

Ameeta Agrawal, Aijun An, and Manos Papagelis. 2018. Learning emotion-enriched word representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 950–961.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 579–586.

Aaron T Beck, Brad A Alford, MD Aaron T Beck, and Ph D Brad A Alford. 2014. *Depression*. University of Pennsylvania Press.

Sidney J Blatt. 2004. *Experiences of depression: Theoretical, clinical, and research perspectives*. American Psychological Association.

Belinda Campos, Michelle N Shiota, Dacher Keltner, Gian C Gonzaga, and Jennifer L Goetz. 2013. What is shared, what is different? core relational themes and expressive displays of eight positive emotions. *Cognition & emotion*, 27(1):37–52.

Jason A Clark. 2010. Relations of homology between higher cognitive emotions and basic emotions. *Biology & Philosophy*, 25(1):75–94.

Alan S Cowen and Dacher Keltner. 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900–E7909.

Taner Danisman and Adil Alpkocak. 2008. Feeler: Emotion classification of text using vector space model. In *AISB 2008 Convention Communication, Interaction and Social Intelligence*, volume 1, page 53.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Beverley Fehr and James A Russell. 1984. Concept of emotion viewed from a prototype perspective. *Journal of experimental psychology: General*, 113(3):464.

Johnny RJ Fontaine, Klaus R Scherer, Etienne B Roesch, and Phoebe C Ellsworth. 2007. The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. 2015. Detecting emotion stimuli in emotion-bearing sentences. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 152–165. Springer.

Paul E Griffiths. 2002. Basic emotions, complex emotions, machiavellian emotions.

Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. 2016. Event-driven emotion cause extraction with corpus construction. In *EMNLP*, pages 1639–1649. World Scientific.

Cindy Harmon-Jones, Brock Bastian, and Eddie Harmon-Jones. 2016. The discrete emotions questionnaire: A new tool for measuring state self-reported emotions. *PloS one*, 11(8):e0159915.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Ehsan Imani and Martha White. 2018. Improving regression performance with distributional losses. In *International Conference on Machine Learning*, pages 2157–2166. PMLR.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 151–160.

Jutta Joormann and Colin H Stanton. 2016. Examining emotion regulation in depression: A review and future directions. *Behaviour research and therapy*, 86:35–49.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Roman Klinger et al. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119.

Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. 2010. A text-driven rule-based system for emotion cause detection. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 45–53.

Junjie Li, Haitong Yang, and Chengqing Zong. 2018. Document-level multi-aspect sentiment classification by jointly modeling users, aspects, and overall ratings. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 925–936.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Chen Liu, Muhammad Osama, and Anderson De Andrade. 2019. Dens: A dataset for multi-class emotion analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6294–6299.

Iris B Mauss and Michael D Robinson. 2009. Measures of emotion: A review. *Cognition and emotion*, 23(2):209–237.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Saif Mohammad. 2012. # emotional tweets. In ** SEM 2012: 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 (SemEval 2012)*, pages 246–255.

Saif Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 65–77.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Emily Öhman, Kaisla Kajava, Jörg Tiedemann, and Timo Honkela. 2018. Creating a dataset for multilingual fine-grained emotion-detection using gamification-based annotation. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 24–30.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Mary Phuong and Christoph Lampert. 2019. Towards understanding knowledge distillation. In *International Conference on Machine Learning*, pages 5142–5151.

Jonathan Rottenberg. 2005. Mood and emotion in major depression. *Current Directions in Psychological Science*, 14(3):167–170.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Klaus R Scherer. 2005. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729.

Klaus R Scherer, Vera Shuman, Johnny Fontaine, and Cristina Soriano Salinas. 2013. The grid meets the wheel: Assessing emotional feeling via self-report. *Components of emotional meaning: A sourcebook*.

Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Duyu Tang, Furu Wei, Bing Qin, Nan Yang, Ting Liu, and Ming Zhou. 2015. Sentiment embeddings with applications to sentiment analysis. *IEEE transactions on knowledge and data Engineering*, 28(2):496–509.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1555–1565.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394.

Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha. 2003. Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, pages 91–109. Springer.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2012. Harnessing twitter" big data" for automatic emotion identification. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 587–592. IEEE.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012.

Rui Xia, Feng Xu, Chengqing Zong, Qianmu Li, Yong Qi, and Tao Li. 2015. Dual sentiment analysis: Considering two sides of one review. *IEEE transactions on knowledge and data engineering*, 27(8):2120–2133.

Rui Xia, Chengqing Zong, and Shoushan Li. 2011. Ensemble of feature sets and classification algorithms for sentiment classification. *Information sciences*, 181(6):1138–1152.

Peng Xu, Andrea Madotto, Chien-Sheng Wu, Ji Ho Park, and Pascale Fung. 2018. Emo2vec: Learning generalized emotion representation by multi-task training. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 292–298.

Liang-Chih Yu, Jin Wang, K Robert Lai, and Xuejie Zhang. 2017. Refining word embeddings for sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 534–539.

Zhenyu Zhao, Shuangzhi Wu, Muyun Yang, Kehai Chen, and Tiejun Zhao. 2020. Robust machine reading comprehension by learning soft labels. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2754–2759.

Chengqing Zong, Rui Xia, and Jiajun Zhang. 2019. *Text Data Mining* (in Chinese). Tsinghua University Press, Beijing.