Improving Textual Emotion Recognition Based on Intra- and Inter-Class Variation

Hassan Alhuzali and Sophia Ananiadou

Abstract—Textual Emotion Recognition (TER) is an important task in Natural Language Processing (NLP), due to its high impact in real-world applications. Prior research has tackled the automatic classification of emotion expressions in text by maximising the probability of the correct emotion class using cross-entropy loss. However, this approach does not account for intra- and inter-class variations within and between emotion classes. To overcome this problem, we introduce a variant of triplet centre loss as an auxiliary task to emotion classification. This allows TER models to learn compact and discriminative features. Furthermore, we introduce a method for evaluating the impact of intra- and inter-class variations on each emotion class. Experiments performed on three data sets demonstrate the effectiveness of our method when applied to each emotion class in comparison to previous approaches. Finally, we present analyses that illustrate the benefits of our method in terms of improving the prediction scores as well as producing discriminative features.

Index Terms—Textual emotion recognition, emotion classification, learning intra- and inter-class variation, variant triplet centre loss.

1 INTRODUCTION

The growing interest in Textual Emotion Recognition (TER) has been motivated by the proliferation of social media and online data, which have made it possible for people to communicate and share opinions on a variety of topics. Interest in TER has also given rise to new NLP methods focusing on TER identification and classification [1], [2], [3], [4], [5], [6], [7], [8], [9], [10]. Research into TER has contributed to a wide range of real-world applications, e.g., health and well-being [11], [12], [13], [14], author profiling [15], [16], human-machine interaction [17], [18], [19], education [20], [21], financial technology [22], [23], [24] and consumer analysis [25], [26].

The majority of previous research has focused on emotion classification as a single-label prediction problem by selecting the most dominant class for a given emotion expression. This approach makes use of cross-entropy loss, which attempts to maximise the probability of the correct class. However, it does not account for cases in which certain emotions (e.g., anger, disgust or sadness) may be confused with each other. Consider S1 in Table I which contains a strong expression of “joy”, even though it is generally more negative oriented. This can lead TER models to choose the “joy” over “sadness” emotion. S2 is annotated with “disgust”, while at the same time it could be possibly labelled with “anger”, due to the missing of explicit emotion-based keywords for the “disgust” emotion, as well as their similarities in linguistic expressions between the two emotions. This linguistic overlap between different emotion classes can cause TER models to mislabel emotions and affect their performance in selecting the correct label. Mohammad and Bravo-Marquez [27] observe that negative emotions are highly associated with each other.

Based on these observations, we hypothesise that taking into account variations both within and between different classes of emotion can better support TER models in learning discriminative features and improve their prediction capability. In this paper, we refer to examples sharing the same emotion class as “intra-class”, while examples belonging to different emotion classes are referred to as “inter-class”. Our contributions are summarised as follows:

I. We propose a novel loss function aimed at incorporating intra- and inter-class variations into TER. More specifically, we introduce a variant of triplet centre loss (VTCL) as an auxiliary task to emotion classification loss (i.e., cross-entropy loss). The objective of VTCL is to minimise the distance of the examples from the centre within the same emotion class (intra-class), while maximising their distances from the centres of other emotions classes (inter-class).

II. We present a new evaluation method to quantify the impact of intra- and inter-class variations on each emotion class.

III. We demonstrate that taking into account intra- and inter-class variations can improve model performance compared to previous approaches, even without the use of external resources. Empirical evaluation and analysis demonstrate that both intra- and inter-class variations can help the model to achieve high prediction scores as well as rendering it a better discriminator against highly associated emotions.

The paper is organised as follows: section II reviews

### Table 1: Example Tweets from IEST data set [3]. GT represents the ground-truth labels.

<table>
<thead>
<tr>
<th>#</th>
<th>Sentence</th>
<th>GT</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>I love you so much and i am [trigger_word] because you do not know that i exist.</td>
<td>sadness</td>
</tr>
<tr>
<td>S2</td>
<td>I get so [trigger_word] when parents smoke right next to their little kids.</td>
<td>disgust</td>
</tr>
</tbody>
</table>
related work, while section 3 provides an overview of triplet centre loss and describes how our method improves upon it. Section 4 discusses experimental details. We evaluate our method and compare it to related methods in section 5. Finally, we analyse the results in section 6 and provide conclusions in section 7.

2 RELATED WORK

Existing methods for emotion recognition are mostly cast as a single-label classification problem, in which a single emotion class is assigned to each sample. Earlier studies focused on lexicon-based approaches, which make use of a set of emotion seed words and their corresponding labels to identify emotions in text, e.g., NRC [5] and EmoSenticNet [28]. Other methods treat emotion recognition as a supervised learning task, in which a learner (e.g. a linear classifier) is trained on the features of labelled data to assign a single label to each sample. For example, Wang et al. [8] applied two machine learning algorithms to a large Twitter data set collected via distant-supervision by using a list of hashtags to exploit the effectiveness of the size of training data on emotion classification.

More recent studies on TER have focused on learning emotion features/representations via deep learning. Sequential models (e.g., Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU)) [31, 32, 33, 34, 35] and spatial models (e.g., Convolutional neural networks (CNN)) [36, 37] have been used for TER. Abdul-Majeed and Ungar [34] proposed an emotion classification model developed using GRU, whereas Felbo et al. [35] constructed a bi-directional LSTM with a self-attention mechanism using emoji data, and then adapted it to emotion classification. Saravia et al. [37] built contextualised affect representations used as features for training various neural networks (e.g., LSTM, GRU and CNN). Although these methods experiment with different neural networks for TER, their approaches to emotion classification overlook intra- and inter-class variations within and between emotion classes.

In contrast, our approach to TER accounts for both intra- and inter-class variations in TER. It builds upon the work of He et al. [38], who leveraged both intra- and inter-class variations for object recognition. Our work differs from [38] in the following ways: i) We leverage intra- and inter-class information to recognise emotion expressions in text, rather than objects. To the best of our knowledge, this is the first attempt to apply this approach, in conjunction with triplet centre loss, to text. ii) We employ an alternative method to compute inter-class distance, so as to disentangle the positive emotion label (i.e., ground truth) from the negative ones. iii) We empirically quantify the influence of intra- and inter-class variations directly for each emotion class. We finally combine VTCL with the cross-entropy loss and train them jointly.

3 METHODOLOGY

3.1 Triplet Centre Loss

Triplet centre loss (TCL) is a combination of triplet loss (TL) [39] and centre loss (CL) [40]. TL defines a triplet as an anchor sample, a positive sample and a negative sample; the first two samples belong to the same class, while the last one belongs to a different class. The objective of TL is to minimise the distance between an anchor sample and a positive sample, while increasing the distance to a negative sample by at least a margin $m$. However, the number of triplets can grow cubically as the number of samples increases, which requires a long training period. In addition, the performance of TL is highly dependent on the choice of triplet mining technique, which is also computationally expensive. The above-mentioned reasons make TL models hard to train.

An alternative choice to TL is CL, which learns the centre for the samples of each class, with the objective of pulling them as close as possible to their respective centre. Although CL is easier to implement, it runs the risk of degrading all features and centres to zero [40]. To address this problem, CL is trained in conjunction with cross-entropy loss, since the latter can act as a guide to learn better centres. Nevertheless, CL does not guarantee that the centres of different classes are pushed sufficiently far from each other. This is because CL only focuses on minimising intra-class distance, but it does not directly address the issue of maximising inter-class distance.

In response to the above, He et al. [38] proposed TCL, which follows the same method as TL, while simultaneously avoiding its complexity. TCL only requires access to a sample (i.e., its corresponding centre and its nearest negative centre). In this respect, TCL leverages the benefits of both TL and CL, in that it pulls samples as close as possible to their corresponding centre, while pushing the same samples as far away as possible from their nearest negative centre.

3.2 Variant Triplet Centre Loss

Our proposed method is an enhancement of He et al. [38] triplet centre loss, which we call Variant TCL (VTCL). In VTCL, we assume that the features of emotion expressions from one class could be shared by expressions from other emotion classes. This makes our approach distinct from TCL for two reasons. Firstly, since TCL only considers the nearest negative centre, the difference between intra- and inter-class distances for multiple (possibly very similar) emotion classes cannot be established. Secondly, TCL randomly initialises the parametric centres, making the process of selecting the nearest negative centre hard to achieve. This is particularly problematic for a task like TER, in which multiple classes could be used as negative centres, due to the close association between certain emotion classes (e.g., anger, disgust and sadness).

To address the above challenges, we map each emotion class to one corresponding centre and treat all but the one positive class centre as negative centres. This simplifies our method by obviating the need to determine the closest negative centre. In other words, examples belonging to the same class should be as close as possible to each other (intra-class), while the same examples should be as far away as possible from other emotion classes (inter-class). This ensures that the intra-class distance plus the margin are always smaller than the inter-class distance. Our experiments in section 6.4 show the impact of choosing different numbers of negative centres. We compute VTCL as follows:

$$L_{VTCL} = \max \left(\text{intra} + m - \text{inter}, 0\right),$$ (1)
where intra- and inter-class distances are computed by using the Squared Euclidean Distance as shown in equations (2) and (3), respectively. \( m \) is a marginal difference between the intra- and inter-class distances.

\[
\text{intra} = \frac{1}{2} \sum_{i=1}^{B} \left\| f_i - c_{y_i} \right\|_2^2, \tag{2}
\]

\[
\text{inter} = \frac{1}{2} \sum_{i=1}^{B} \sum_{j \neq y_i} \left\| f_i - c_j \right\|_2^2, \tag{3}
\]

where \( B \) is the training batch size, \( C \) corresponds to the number of emotion classes, \( f_i \in \mathbb{R}^d \) is the \( i^{th} \) input representation, \( c_{y_i} \in \mathbb{R}^d \) is the centre of class \( y_i \) and \( c_j \in \mathbb{R}^d \) is the centre of other emotions, with \( d \) defining the dimensional size.

### 3.3 Training Objective

As VTCL initialises the parametric centres randomly and updates them based on the mini-batches, it is difficult to achieve accurate class centres. To mitigate this problem, we train VTCL jointly with the cross-entropy loss function (CEL). VTCL applies metric learning to the learned feature representation directly, while CEL focuses on mapping examples to their emotion classes, helping to achieve discriminative as well as compact features, respectively. The overall training objective can be defined as follows:

\[
L_{\text{JOINT}} = L_{\text{CEL}} + \lambda L_{\text{VTCL}}, \tag{4}
\]

where the first term refers to the cross-entropy loss, which is computed as in shown equation (5), while the second term corresponds to VTCL. \( \lambda \in [0, 1] \) denotes the value used to control the trade-off between \( L_{\text{CEL}} \) and \( L_{\text{VTCL}} \).

\[
L_{\text{CEL}} = - \sum_{i=1}^{M} \sum_{j=1}^{C} I\{y_i = j\} \log \frac{e^{a_j^{(i)}}}{\sum_{j=1}^{C} e^{a_j^{(i)}}}, \tag{5}
\]

where the indicator function \( I\{condition\} = 1 \) if the condition is satisfied, or 0 otherwise. \( a_j^{(i)} \) represents the activation values of emotion classes in the last fully-connected layer for an example.

### 4 Experiments

In this work, we run our method on two widely used networks for TED: the first network is based on a CNN architecture proposed by [41] for text classification, while the other network is based on BERT [45]. Figure 1 illustrates the proposed method, which takes advantage of the same feature representation obtained via either BERT or CNN.

#### 4.1 Implementation Details

The CNN network’s weights were initialised from Word2Vec [44] embedding with a size of 300 dimensions and it included filter windows of (3, 4, 5) with 100 feature maps each, a batch size of 64 and a dropout rate of 0.5. We used the standard normal distribution to initialise the centres and we set the margin \( (m) \) double the number of negative centres. Adam was selected for optimisation [45] with a learning rate of 1e-3 for the network, as well as for the centres. All experiments were performed with a fixed initialisation seed using PyTorch [46] and an Nvidia GeForce GTX 1080 with 11 GB memory. Table 2 summarises the hyper-parameters used in this work, including those related to BERT.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>CNN</th>
<th>BERT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Window sizes</td>
<td>{(3, 4, 5)}</td>
<td>-</td>
</tr>
<tr>
<td>Feature maps</td>
<td>100</td>
<td>-</td>
</tr>
<tr>
<td>Feature dimension</td>
<td>300</td>
<td>768</td>
</tr>
<tr>
<td>Batch size</td>
<td>64</td>
<td>32</td>
</tr>
<tr>
<td>Dropout</td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>Learning rate</td>
<td>1e-3</td>
<td>2e-5</td>
</tr>
<tr>
<td>Margin ((m))</td>
<td>2 \times</td>
<td>NC ]</td>
</tr>
<tr>
<td>Optimiser</td>
<td>Adam</td>
<td>-</td>
</tr>
<tr>
<td>Early stop patience</td>
<td>10</td>
<td>-</td>
</tr>
</tbody>
</table>

#### 4.2 Data Sets and Task Settings

We evaluated our method on three widely used single-label data sets and conducted our experiments in a stratified 10-fold cross-validation setup, ensuring that all folds contain an approximately equal sample of emotion classes. We now turn to discussing each data set separately.

**IEST** [47] stands for “Implicit Emotion shared-task” and contains 191,731 tweets collected through distant supervision [3]. Each tweet was acquired from an expression of a triggered emotion keyword plus either “that, because or when”. Consider the example, “Boys who like Starbucks make me [trigger-word] because we can go on cute coffee dates”, where the task is to predict the emotion of the triggered word in the tweet as “joy”. Due to resource constraints, we selected a small, random subset of this data, consisting of 5,000 tweets for each emotion.

**ISEAR** [3] stands for “International Survey on Emotion Antecedents and Reactions” and is one of the first created emotion corpora [47]. This corpus consists of 7,665 sentences.

1. We train BERT on the default hyper-parameters using the open-source Hugging Face implementation [42].
3. The \( m \) parameter is set via observing the F1-score curve on the validation set.
5. https://www.unige.ch/cisa
where each sentence is annotated with a single category of basic emotions (i.e., joy, anger, sadness, fear and disgust) and two additional categories (i.e., shame and guilt). The corpus is acquired from questioners based on descriptions of people’s experiences with different cultural backgrounds.

**TEC** stands for “Twitter Emotion Corpus” [5]. The TEC corpus consists of 21,48 tweets self-labelled by the users of such tweets via “hashtags” (e.g., #joy, #glad, #sad, and #anger, among others). The objective was to determine whether or not this method can be used as a surrogate for gathering emotion data automatically.

In this paper, we focus on Ekman’s [48] 6 basic emotions {anger, disgust, fear, joy, sadness, and surprise} because two of the data sets (i.e., IEST and TEC) we used are annotated with those 6 emotions. Table 3 provides a summary of each data set, including the domain (i.e. the source from which the data set is collected), the size, the number of words and the average length of sentences/tweets for each data set.

To pre-processing the data, we utilise the “ekphrasis” tool [19] designed for the specific characteristics of Twitter, e.g., misspellings and abbreviations since two of the data sets we used are collected from Twitter. The tool offers different functionalities, such as tokenisation, normalisation, spelling correction, and segmentation. For all three data sets, we used the tool to tokenize the text, convert words to lower case, and normalise user mentions, URLs and repeated-characters.

**TABLE 3: Statistics of datasets.** Avg.length refers to the average length of sentences/tweets. The number of words as well as the average length of tweets/sentences are counted after applying the “ekphrasis” tool.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>IEST</th>
<th>ISEAR</th>
<th>TEC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain</td>
<td>Tweets</td>
<td>Events</td>
<td>Tweets</td>
</tr>
<tr>
<td># Sentences</td>
<td>30k</td>
<td>7k</td>
<td>21k</td>
</tr>
<tr>
<td># Words</td>
<td>25k</td>
<td>8k</td>
<td>22k</td>
</tr>
<tr>
<td>Avg.length</td>
<td>23.47</td>
<td>25.5</td>
<td>17.56</td>
</tr>
</tbody>
</table>

## 5 Evaluation

### 5.1 Intra- and inter-class evaluation

We evaluate the ability of our method to distinguish between intra- and inter-class variations with respect to each emotion. Since there is no existing metric for evaluating the impact of intra- and inter-class variations on each emotion class, we choose the confusion matrix. The confusion matrix provides a summary of the model performance per class, where correct predictions are represented in the diagonal, while incorrect predictions are shown outside the diagonal. For example, if a row represents joy, we then obtain the values of “joy-to-joy” (i.e., correctly labelled examples), “joy-to-anger” (i.e., mislabelled examples), “joy-to-disgust” (i.e., mislabelled examples), etc. We use the value of correctly labelled examples to represent the intra-class performance, while utilising the values of incorrectly labelled examples to represent the inter-class performance. The inter-class values are then summed up, similar to how they are computed in equation (5). We can then quantify the impact of intra- and inter-class results with respect to each emotion class.

Table 4 presents the results of intra- and inter-class performance per emotion class on all three data sets. We compare the performance of TCL and VTCL, because both are optimised for the objective of minimising the intra-class distance (i.e., within samples sharing the same emotion) and maximising the inter-class distance (i.e., between samples sharing dissimilar emotions). As Table 4 demonstrates, compared to TCL, our VTCL method achieves high values for intra-class distance and low values for inter-class distance among almost all emotion classes apart from the disgust class in the TEC dataset. We attribute this to the small number of tweets as shown in Table 3, which are roughly 761 tweets for both the training and test sets. We observe that some emotions are easier to distinguish than others. For example, the “joy, fear and sadness” emotions achieved higher marginal differences between the intra- and inter-class than anger and disgust. This finding is consistent with the studies of Mohammad and Bravo-Marquez [27] and Agrawal et al. [32], both of which report the same issue with negative emotions of “anger and disgust”, as they are easily confused with each other.

In contrast to our VTCL method, TCL fails to properly distinguish the difference between intra- and inter-class variations for some emotions in the three data sets. This confirms our observations introduced in section 3.2 that TCL’s selection of the nearest negative centre is problematic for TER, in which it is often important to use multiple centres as negative centres. Nonetheless, VTCL proved effective in increasing the variance between negative emotions, which are often positively correlated with each other, demonstrating the benefits of taking all negative emotions into account instead of only the nearest negative centre as is the case in TCL.

### 5.2 Results

Table 5 presents the performance of VTCL on each data set, in terms of precision, recall and F1-score, and compares it to previously reported state-of-the-art approaches for TER, contextualised embedding and strong loss functions. The results reported in Table 5 are an average of stratified 10-fold cross-validation. In the sections below, we briefly describe the methods that we have compared, including methods that learn a joint loss function to improve the results of emotion classification and those that only use CEL.

#### 5.2.1 Relevant Work

Klinger et al. [29] used a Maximum Entropy classifier (MaxEnt) with a bag of words features for detecting emotion expressions in text. This model exhibits the lowest performance among all compared approaches, as it was trained only on simple features. Islam et al. [34] built a multi-channel-CNN (MCC), which attempts to learn embeddings for each
We also compared and applied our method to BERT for two reasons: i) it can serve as a strong baseline and ii) it can demonstrate the usefulness of VTCL when tested on a different network. To create a sentence representation, we stack a softmax activation layer over the hidden state corresponding to [CLS] in BERT and only consider the “bert-base-uncased” model. As shown in Table 5, BERT trained with the other variants of loss functions apart from VTCL achieves higher results than previously reported approaches to TER on all three data sets. Nevertheless, “CNN (ours)” outperforms the results of BERT trained with the other loss functions apart from VTCL on both IEST and TEC data sets. Although BERT obtains competitive results to “CNN (ours)”, its trained parameters are much larger than those of CNN trained jointly with VTCL.

The fact that BERT scores are higher on the ISEAR data set than “CNN (ours)” maybe because this data set is quite similar to its pre-training corpus. To investigate this, we measured the degree of common word coverage between the “bert-base-uncased” vocabulary and the training set of each data set. We found that the percentage of shared words between the “bert-base-uncased” vocabulary and the training set of ISEAR is 74%, while it is less than 50% for the other two data sets. This confirms our above observation that BERT is pre-trained on a corpus more similar to the ISEAR data set than the IEST and TEC data sets.

We considered further a label semantic (LS) approach which adopted BERT as its encoder and aimed at learning emotion classification (i.e., CEL) and correlation (i.e., Corr) via a joint loss function. Although the LS model used a joint loss function as well as learning the input representation to BERT, it achieved lower performance than our method. It is also worth mentioning that this model takes longer to train than our method because it casts the task as a multiple choice answering task.

5.2.3 Our Method (CEL + VTCL)

Table 5 demonstrates that “CNN (ours)” outperforms all compared models on the IEST and TEC data sets, and all models apart from MTL and BERT when applied to the ISEAR data set. However, when BERT is trained jointly with VTCL, it achieves the highest results across the three data sets. A further observation is that CNN/BERT trained jointly with VTCL.

Finally, we compare strong variants of loss functions aimed at learning intra- and inter-class variations, i.e., including CEL + CL [51] and CEL + TCL [38]. Based on experimental results, we observe that including intra- and inter-class information improves the model performance; CL achieves higher results than TCL on almost all metrics and data sets, proving our earlier hypothesis in section 5.2 that determining the nearest negative centre is indeed not possible for TER. The same patterns are also observed in BERT experiments, which are discussed below.

5.2.2 Contextualised Embeddings

We also compared and applied our method to BERT for two reasons: i) it can serve as a strong baseline and ii) it can demonstrate the usefulness of VTCL when tested on a different network. To create a sentence representation, we stack a softmax activation layer over the hidden state corresponding to [CLS] in BERT and only consider the “bert-base-uncased” model. As shown in Table 5, BERT trained with the other variants of loss functions apart from VTCL achieves higher results than previously reported approaches to TER on all three data sets. Nevertheless, “CNN (ours)” outperforms the results of BERT trained with the other loss functions apart from VTCL on both IEST and TEC data sets. Although BERT obtains competitive results to “CNN (ours)”, its trained parameters are much larger than those of CNN trained jointly with VTCL.

The fact that BERT scores are higher on the ISEAR data set than “CNN (ours)” maybe because this data set is quite similar to its pre-training corpus. To investigate this, we measured the degree of common word coverage between the “bert-base-uncased” vocabulary and the training set of each data set. We found that the percentage of shared words between the “bert-base-uncased” vocabulary and the training set of ISEAR is 74%, while it is less than 50% for the other two data sets. This confirms our above observation that BERT is pre-trained on a corpus more similar to the ISEAR data set than the IEST and TEC data sets.

We considered further a label semantic (LS) approach which adopted BERT as its encoder and aimed at learning emotion classification (i.e., CEL) and correlation (i.e., Corr) via a joint loss function. Although the LS model used a joint loss function as well as learning the input representation from BERT, it achieved lower performance than our method. It is also worth mentioning that this model takes longer to train than our method because it casts the task as a multiple choice answering task.
to CNN/BERT trained on the other loss functions (i.e., CEL, CEL+CL and CEL+TCL). This proves the strength of VTCL against these loss functions as well as against both the MTL and LS approaches. In addition, VTCL does not rely on any external resources unlike MTL, which relies on emotion lexicons to generate label distribution. Moreover, VTCL only requires a small number of parameters to be trained, equivalent to the number of emotion classes multiplied by the size of the feature dimension. Even though VTCL is tested on the simple CNN network architecture, it shows strong performance because, unlike other approaches, it benefits from taking into account intra- and inter-class variation, whose impact on model performance is assessed in the next section via an ablation study.

5.3 Ablation Study

We undertake an ablation study of the results using two settings: firstly, the model is trained without inter-class and subsequently, it is trained without intra-class. Training the model without these two types of information is equivalent to training it only with cross-entropy loss.

TABLE 6: Ablation experiment results. The proportions in parentheses indicate the relative change with respect to ours.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>IEST</th>
<th>ISEAR</th>
<th>TEC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>F1 (%)</td>
<td>F1 (%)</td>
<td>F1 (%)</td>
</tr>
<tr>
<td>CNN (ours)</td>
<td>57.71</td>
<td>65.79</td>
<td>58.19</td>
</tr>
<tr>
<td>-inter</td>
<td>56.63 (1.2%)</td>
<td>64.27 (1.2%)</td>
<td>56.96 (1.2%)</td>
</tr>
<tr>
<td>-intra</td>
<td>55.22 (4.4%)</td>
<td>63.66 (4.3%)</td>
<td>53.55 (4.8%)</td>
</tr>
<tr>
<td>BERT (ours)</td>
<td>59.38</td>
<td>68.89</td>
<td>59.47</td>
</tr>
<tr>
<td>-inter</td>
<td>56.98 (4.4%)</td>
<td>67.42 (4.2%)</td>
<td>57.93 (4.3%)</td>
</tr>
<tr>
<td>-intra</td>
<td>56.27 (5.0%)</td>
<td>66.92 (5.3%)</td>
<td>57.67 (5.3%)</td>
</tr>
</tbody>
</table>

As Table 6 shows, the results of CNN and BERT drop by 2-4% f1-score when the inter-class is removed. When the intra-class is additionally removed, the performance drop increases to 3-8% in f1-score. These results demonstrate the benefits of incorporating intra- and inter-class variations into TER, supporting our hypothesis that taking into account both types of information can improve the model performance substantially.

6 ANALYSIS

6.1 Model Predictions

We analysed the model predictions on two different objectives: firstly, the model is trained only with the cross-entropy loss, and subsequently, it is jointly trained with VTCL. Our research hypothesis is that including VTCL in the emotion classification loss (i.e., cross-entropy) can generate more discriminative features and thus increase the model prediction scores. For this analysis, we use the CNN network architecture and hyper-parameters discussed in section 4.

For each data set, we randomly selected one example per emotion class whose scores are correctly predicted by the two objectives mentioned above and extracted their prediction scores with respect to each emotion class.

In Figure 2, the graphs illustrate prediction scores when the model is trained without VTCL (left-hand graphs) and with VTCL (right-hand graphs). The sub-figures from top to bottom correspond to the instances extracted from IEST, ISEAR and TEC data sets, respectively. In the top group of sub-figures (corresponding to examples from the IEST data set), it can be observed that for the model trained without VTCL overlaps with other emotion classes and as a result, the prediction score for the correct emotion class is low and is close to the prediction scores for other emotion classes. However, when the model is trained jointly with VTCL, a much higher prediction score is achieved for the correct emotion, which is well distinguished from all the other emotion classes. Figure 2b shows the scores of the “disgust” class (i.e., without vs with VTCL), demonstrating
This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

Fig. 2: Prediction scores (y-axis) across emotions (x-axis). Each sub-figure shows the scores of the two evaluated objectives, i.e., without VTCL (left) vs with VTCL (right). The corresponding instance to be classified is included at the bottom of each sub-figure. The frames from top-to-bottom represents instances belonging to IEST, ISEAR and TEC data sets, respectively. ‘...’: refers to the removed triggered word from the IEST data set.
TABLE 7: Analysis of the model predictions trained on two settings (i.e., w/o VTCL vs w/ VTCL). The actual label for each example is also included.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Text</th>
<th>Actual</th>
<th>w/o VTCL</th>
<th>w/ VTCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEST</td>
<td>So i was [triggered_word] because i told him to charge it before he left so i am already in a bad mood. I get so [triggered_word] when parents smoke right next to their little kids. I really get [triggered_word] when someone says i will be waiting for that day. to do what exactly? I think i will finally be [triggered_word] when i go to a fête. just need to get rid of this stress. I love you so much and i am [triggered_word] because you do not know that i exist. I was so [triggered_word] when she said, it had no relevancy to my statement.</td>
<td>anger</td>
<td>fear</td>
<td>anger</td>
</tr>
<tr>
<td></td>
<td>disgust</td>
<td>disgust</td>
<td>anger</td>
<td>disgust</td>
</tr>
<tr>
<td></td>
<td>joy</td>
<td>joy</td>
<td>sadness</td>
<td>joy</td>
</tr>
<tr>
<td></td>
<td>joy</td>
<td>joy</td>
<td>sadness</td>
<td>joy</td>
</tr>
<tr>
<td></td>
<td>sadness</td>
<td>sadness</td>
<td>joy</td>
<td>sadness</td>
</tr>
<tr>
<td></td>
<td>surprise</td>
<td>surprise</td>
<td>disgust</td>
<td>surprise</td>
</tr>
<tr>
<td>ISEAR</td>
<td>Someone told me that i was chosen for english lectures because the class leader is going out with me (not true). When i heard that a woman of my community had aborted and got rid of the foetus by throwing it in the drain. When there was a bomb threat in &lt;place&gt; hall. this was the first time that i felt my life could be in danger. Doing unexpectedly well in an examn. Not having good marks like other people for homeworks.</td>
<td>anger</td>
<td>disgust</td>
<td>anger</td>
</tr>
<tr>
<td></td>
<td>disgust</td>
<td>disgust</td>
<td>sadness</td>
<td>disgust</td>
</tr>
<tr>
<td></td>
<td>fear</td>
<td>fear</td>
<td>joy</td>
<td>fear</td>
</tr>
<tr>
<td></td>
<td>joy</td>
<td>joy</td>
<td>sadness</td>
<td>joy</td>
</tr>
<tr>
<td></td>
<td>sadness</td>
<td>sadness</td>
<td>disgust</td>
<td>sadness</td>
</tr>
<tr>
<td></td>
<td>surprise</td>
<td>surprise</td>
<td>anger</td>
<td>surprise</td>
</tr>
<tr>
<td>TEC</td>
<td>The cock who keeps pushing his chair onto my legs needs to stop. I am honestly ashamed to be living in the same state as &lt;place&gt; state. i cannot even imagine being a student there expect respect. No sleep today. cannot even rest when the sun's down. That feeling you get when you open up a bill and there's a credit. no payment required. Ever wish you could go back a few years , and do it all differently. When you think if a year ago someone told you this was going to happen, you would not believe them.</td>
<td>anger</td>
<td>sadness</td>
<td>anger</td>
</tr>
<tr>
<td></td>
<td>disgust</td>
<td>disgust</td>
<td>fear</td>
<td>disgust</td>
</tr>
<tr>
<td></td>
<td>fear</td>
<td>fear</td>
<td>sadness</td>
<td>fear</td>
</tr>
<tr>
<td></td>
<td>joy</td>
<td>joy</td>
<td>anger</td>
<td>joy</td>
</tr>
<tr>
<td></td>
<td>sadness</td>
<td>sadness</td>
<td>joy</td>
<td>sadness</td>
</tr>
<tr>
<td></td>
<td>surprise</td>
<td>surprise</td>
<td>anger</td>
<td>surprise</td>
</tr>
</tbody>
</table>

the improvement brought by our approach in increasing the correct prediction score, as well as reducing the overlap with other highly correlated emotions (e.g., anger and fear). A similar pattern can be observed for the other instances belonging both to the same data set (i.e., IEST) and to the ISEAR and TEC data sets, thus supporting our hypothesis that the incorporation of both intra- and inter-class variations into the task of TER increases performance by introducing discriminative features.

6.2 Qualitative analysis

In order to better understand the performance of our method on the same two objectives discussed in section 6.1, we carried out a qualitative analysis of the predictions made by each objective. We observe that in many cases, the second objective (i.e., training the model with VTCL) performs better than the first objective (i.e., training it without VTCL). Table 7 presents the analysis. Since some emotions share similarities in linguistic expressions, the model can easily confuse and mislabel emotions. This problem mainly appears in negative emotions (i.e., anger, fear, disgust and sadness). We also note that the main sources of errors made by the first objective are cases involving strong expressions of one emotion over another, implicit emotions and certain lexical units.

For example, the first, second and final examples of the IEST data set show implicit emotion instances, which led the model to select incorrect predictions. Moreover, the presence of the strong expressions “love you so much” and “get rid of this stress” in the fourth and fifth examples of the IEST data set confuse the model with the first objective, such that it selects incorrect predictions. In contrast, the model trained with the second objective is able to overcome these potential confusions and predict the correct emotion. Similar patterns are also seen in ISEAR and TEC data sets. Overall, introducing discriminative features helps the model overcome the above-discussed problems and predict the correct emotion labels with high probabilities, thus supporting our hypothesis regarding the importance of incorporating intra- and inter-class variations for TER.

6.3 Visualisation of Learned Representations

To provide insights into the ability of our method to introduce discriminative features, we selected the penultimate layer of BERT and CNN, and then used t-SNE [55] to visualise the learned features. For this analysis, we randomly chose 1,000 examples from the test set of IEST data and then trained models by following the same two objectives discussed in section 6.1 (i.e., training the model without VTCL vs with VTCL).

Figure 8 visualises the learned features for each emotion label, from which we observe some positive properties: i) The first objective performs poorly in learning compact and discriminative features, whereas the second one is able to simultaneously create compact and more clearly separated clusters. In other words, our method ensures that the learned embeddings of the same emotion label are as close as possible to each other, but also as distant as possible from other emotions. ii) The deeply learned representations from BERT are more clearly separated and compact than the ones obtained from CNN, which is not surprising, as

8. We use the scikit-learn library [54] to generate the t-sne visualisation and follow the default setting.
We have proposed a novel loss function focused on taking into account intra- and inter-class variations within and between emotions. To achieve this, we introduced variant triplet centre loss (VTCL) as an auxiliary task for emotion classification loss (i.e., cross-entropy loss). We showed the effectiveness of incorporating both intra- and inter-class variations into TER, demonstrating their ability to increase model prediction scores, but also to more clearly distinguish between different emotions, especially those highly associated with each other. It is hoped that the results of our study will stimulate further investigation into the usage of metric learning in TER, as well as other related tasks in NLP. As future work, we will extend our method for application to multi-label emotion classification.

8 ACKNOWLEDGEMENTS

We would like to thank Paul Thompson and Fenia Christopoulou for their valuable comments and suggestions. We also thank the anonymous reviewers for their careful reading and insightful feedback. The first author is supported by a doctoral fellowship from Umm Al-Qura University.

REFERENCES

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3104720, IEEE Transactions on Affective Computing


