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

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 datasets 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 1, 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.

• *Hassan Alhuzali and Sophia Ananiadou are with NaCTeM, University of Manchester, Manchester M13 9PL, U.K., and also with the Department of Computer Science, The University of Manchester, Manchester M13 9PL, U.K. E-mail: hassan.alhuzali@postgrad.manchester.ac.uk, sophia.ananiadou@manchester.ac.uk.*

Manuscript received 27 Jan. 2021; revised 26 July 2021; accepted 6 Aug. 2021. Date of publication 13 Aug. 2021; date of current version 31 May 2023.
This work was supported by the Alan Turing Institute and in part by Grant MR/R022461/1. A Youth Culturally adapted Manual Assisted Psychological therapy (Y-CMAP) for adolescent Pakistani pati ents with a recent history of self-harm. (Corresponding author: Hassan Alhuzali.)
Recommended for acceptance by C. Clavel.
Digital Object Identifier no. 10.1109/TAFFC.2021.3104720

TABLE 1
Example Tweets From IEST Dataset [3]

#	Sentence	GT
S1	I love you so much and i am [trigger_word] because you do not know that i exist.	sadness
S2	I get so [trigger_word] when parents smoke right next to their little kids.	disgust

GT represents the ground-truth labels.

The paper is organised as follows: Section 2 reviews 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 [7], [8], [10], [29], [30]. For example, Wang *et al.* [8] applied two machine learning algorithms to a large Twitter dataset 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-Mageed 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. First, 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. Second, 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

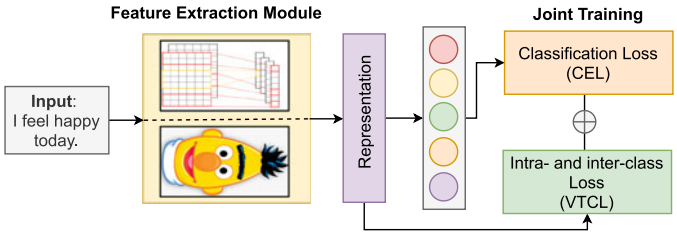

Fig. 1. Illustration of our method. Given the input, we use the feature extraction module to learn the input representation via BERT/CNN and then feed it into our method, which includes the joint supervision of CEL and VTCL, Equation (4).

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:

$$\mathcal{L}_{\text{VTCL}} = \max(\text{intra} + m - \text{inter}, 0), \tag{1}$$

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.

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

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

where B is the training batch size, C corresponds to the number of emotion classes, $f_i \in \mathbb{R}^d$ is the ith 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:

$$\mathcal{L}_{\text{JOINT}} = \mathcal{L}_{\text{CEL}} + \lambda\mathcal{L}_{\text{VTCL}}, \tag{4}$$

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_{CEL} and L_{VTCL}.

$$\mathcal{L}_{\text{CEL}} = -\sum_{i=1}^{B}\sum_{j=1}^{C}\mathbb{1}\{y_i = j\}\log\frac{e^{\mathbf{a}_j^{(i)}}}{\sum_{j=1}^{C}e^{\mathbf{a}_j^{(i)}}}, \tag{5}$$

TABLE 2
Hyper-Parameters

Parameters	CNN	BERT
Window sizes	$\{3, 4, 5\}$	-
Feature maps	100	-
Feature dimension	300	768
Batch size	64	32
Dropout	0.5	0.1
Learning rate	1e-3	2e-5
Margin (m)	$2 \times \lvert NC\rvert$	
Optimiser	Adam	
Early stop patience	10	

$|\text{NC}|$: *denotes the number of negative centres.*

where the indicator function $\mathbb{1}\{condition\} = 1$ 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 TER: the first network is based on a CNN architecture proposed by [41] for text classification, while the other network is based on BERT[1] [43]. Fig. 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 *Word2-Vec* [44] embedding with a size of 300 dimensions,[2] 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.[3] 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.

4.2 Datasets and Task Settings

We evaluated our method on three widely used single-label datasets 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 dataset separately.

IEST. [4] 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

1. We train BERT on the default hyper-parameters using the open-source Hugging Face implementation [42].
2. https://code.google.com/archive/p/word2vec/.
3. The m parameter is set via observing the F1-score curve on the validation set.
4. http://implicitemotions.wassa2018.com/data/

TABLE 3
Statistics of Datasets

Dataset	IEST	ISEAR	TEC
Domain	Tweets	Events	Tweets
# Sentences	30k	7k	21k
# Words	25k	8k	22k
Avg.length	23.47	25.5	17.56
# Sentences/Tweets per class			
Anger	5,000	1,096	1,555
Disgust	5,000	1,096	761
Fear	5,000	1,095	2,816
Joy	5,000	1,094	8,240
Sadness	5,000	3,285	3,830
Surprise	5,000	—	3,849

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.

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. [5] 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, 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 questionnaires based on descriptions of people' experiences with different cultural backgrounds.

TEC. [6] stands for "Twitter Emotion Corpus" [6]. The TEC corpus consists of 21,048 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 datasets (i.e., IEST and TEC) we used are annotated with those 6 emotions. Table 3 provides a summary of each dataset, including the domain (i.e. the source from which the dataset is collected), the size, the number of words and the average length of sentences/tweets for each dataset.

To pre-processing the data, we utilise the "ekphrasis[7]" tool [49] designed for the specific characteristics of Twitter, e.g., misspellings and abbreviations since two of the datasets we used are collected from Twitter. The tool offers different functionalities, such as tokenisation, normalisation, spelling correction, and segmentation. For all three datasets, we used the tool to tokenise the text, convert words to lower case, and normalise user mentions, URLs and repeated-characters.

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 (3). 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 datasets. 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 datasets. 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 dataset, in terms of precision, recall and F1-score, and compares it to previously reported state-of-the-art approaches to 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. https://www.unige.ch/cisa
6. http://saifmohammad.com/WebDocs/Jan9-2012-tweets-clean.txt.zip
7. https://github.com/cbaziotis/ekphrasis

TABLE 4
The Results (%) of Intra- and Inter-Class Values on Three Datasets (i.e., IEST, ISEAR, and TEC)

Dataset	Loss	Mode/Label	anger	disgust	fear	joy	sadness	surprise
IEST	TCL	intra (↑)	49.80	49.40	57.20	64.20	58.40	61.00
		inter (↓)	50.20	50.60	42.80	35.80	41.60	39.00
		Δ (↑)	-0.40	-1.20	14.40	28.40	16.80	**22.00**
	VTCL	intra (↑)	54.20	53.00	62.60	65.60	61.80	56.60
		inter (↓)	45.80	47.00	37.40	34.40	38.20	43.40
		Δ (↑)	**8.40**	**6.00**	**25.20**	**31.20**	**23.60**	13.20
ISEAR	TCL	intra (↑)	46.36	40.91	48.18	59.09	77.03	—
		inter (↓)	53.64	59.09	51.82	40.91	24.22	—
		Δ (↑)	-7.27	-18.18	-3.64	18.18	**52.81**	—
	VTCL	intra (↑)	51.82	52.73	60.91	62.73	66.26	—
		inter (↓)	48.18	47.27	39.09	37.27	33.74	—
		Δ (↑)	**3.64**	**5.45**	**21.82**	**25.45**	32.52	—
TEC	TCL	Intra (↑)	44.32	41.67	47.87	55.83	57.70	55.54
		inter (↓)	55.68	58.33	52.13	44.17	42.30	44.46
		Δ (↑)	-11.36	-16.66	-4.26	11.65	15.40	11.08
	VTCL	intra (↑)	54.19	47.56	64.77	65.90	61.88	56.88
		inter (↓)	45.81	52.44	35.23	34.10	38.12	43.12
		Δ (↑)	**8.39**	-4.88	**29.54**	**31.80**	**23.76**	**13.77**

Note that ↑ after the mode indicates the larger the better, while ↓ after the mode indicates the smaller the better. Δ represents the difference between intra- and inter-class scores. The best scores are highlighted in bold. It should be mentioned that ISEAR is not annotated with the "surprise" emotion.

TABLE 5
Comparison of Our Method to Previous Approaches as Well as Contextualised Embedding Applied to IEST, ISEAR, and TEC Datasets

Dataset	IEST			ISEAR			TEC		
Model/Metric	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
Relevant Work									
MaxEnt	49.41	48.90	48.85	61.20	63.60	62.20	49.83	48.50	49.00
CNN (CEL)	55.74	55.20	55.22	64.21	63.97	63.66	57.79	51.43	53.55
MCC (CEL)†	56.20	56.95	56.09	—	—	—	55.90	**56.50**	55.60
MTL (CEL+KL)‡	57.17	56.93	56.95	67.11	**66.91**	**66.80**	62.10	52.57	56.94
CNN (CEL+CL)	57.20	56.62	56.63	65.80	64.12	64.27	64.47	53.96	56.96
CNN (CEL+TCL)	57.27	56.56	56.60	65.30	64.03	63.59	63.27	54.18	56.90
CNN (ours)	**58.08**	**57.77**	**57.71***	**70.35**	64.14	65.79*	**65.85**	54.95	**58.19***
Contextualised Embeddings									
BERT (CEL)	56.84	56.23	56.27	68.60	67.50	66.92	58.71	58.72	57.67
BERT (CEL+CL)	59.09	56.77	56.98	68.52	67.84	67.42	60.12	57.87	57.93
BERT (CEL+TCL)	57.88	56.70	56.85	68.03	67.01	66.83	59.07	56.46	57.31
LS (CEL+Corr)	57.74	57.00	57.06	67.32	67.10	67.08	59.31	56.51	57.08
BERT (ours)	**60.20**	**59.34**	**59.38***	**70.73**	**69.44**	**68.89***	**60.18**	**60.24**	**59.47***

* *We follow this work [52] to compute the significance tests.*
(P%), (R%) and (F1%): refers to precision, recall and f1-score. Note that † indicates that the model uses hand-crafted features (e.g., tweet-specific, affect and sentiment features), while ‡ indicates that the model utilises emotion lexicons to generate label distribution. * *denotes significance at* $p < 0.05$ *compared to "CEL +CL". The best result in each part is marked in bold.*

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.* [36] built a multi-channel-CNN (MCC), which attempts to learn embeddings for each sample as well as additional features that occur in the same sample (e.g., emojis, emoticons and hashtags). This model achieves better results than the MaxEnt classifier, and it also obtains the highest recall apart from the results of BERT over all models on the TEC dataset. The fact that MCC used additional hand-crafted features (e.g., tweet-specific, affect and sentiment features) may explain its high recall on this dataset. We only report on the results of MCC on the IEST and TEC datasets because it was specifically developed for Twitter. Zhang *et al.* [50] proposes a multi-task-loss approach (MTL) involving learning of both emotion distribution and classification (i.e., CEL plus Kullback-Leibler (KL) loss). The MTL model achieves higher recall and f1 scores than "CNN (ours)" on the ISEAR dataset. However, it should be noted that, in contrast to our

approach, it relies on emotion lexicons to generate label distributions.

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 datasets, proving our earlier hypothesis in Section 3.2 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 datasets. Nevertheless, "CNN (ours)" outperforms the results of BERT trained with the other loss functions apart from VTCL on both IEST and TEC datasets. 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 dataset than "CNN (ours)" may be because this dataset 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 dataset. We found that the percentage of shared words between the "bert-base-uncased" vocabulary and the training set of ISEAR is 74 percent, while it is less than 50 percent for the other two datasets. This confirms our above observation that BERT is pre-trained on a corpus more similar to the ISEAR dataset than the IEST and TEC datasets.

We considered further a label semantic (LS) approach [53] 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.

5.2.3 Our Method (CEL+VTCL)

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

TABLE 6
Ablation Experiment Results

Dataset	IEST	ISEAR	TEC
Model	F1 (%)	F1 (%)	F1 (%)
CNN (ours)	**57.71**	**65.79**	**58.19**
-inter	56.63 (↓ 2%)	64.27 (↓ 2%)	56.96 (↓ 2%)
-intra	55.22 (↓ 4%)	63.66 (↓ 3%)	53.55 (↓ 8%)
BERT (ours)	**59.38**	**68.89**	**59.47**
-inter	56.98 (↓ 4%)	67.42 (↓ 2%)	57.93 (↓ 3%)
-intra	56.27 (↓ 5%)	66.92 (↓ 3%)	57.67 (↓ 3%)

The proportions in parentheses indicate the relative change with respect to ours.

5.3 Ablation Study

We undertake an ablation study of the results using two settings: first, 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.

As Table 6 shows, the results of CNN and BERT drop by 2-4 percent f1-score when the inter-class is removed. When the intra-class is additionally removed, the performance drop increases to 3-8 percent 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: first, 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 dataset, 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 Fig. 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

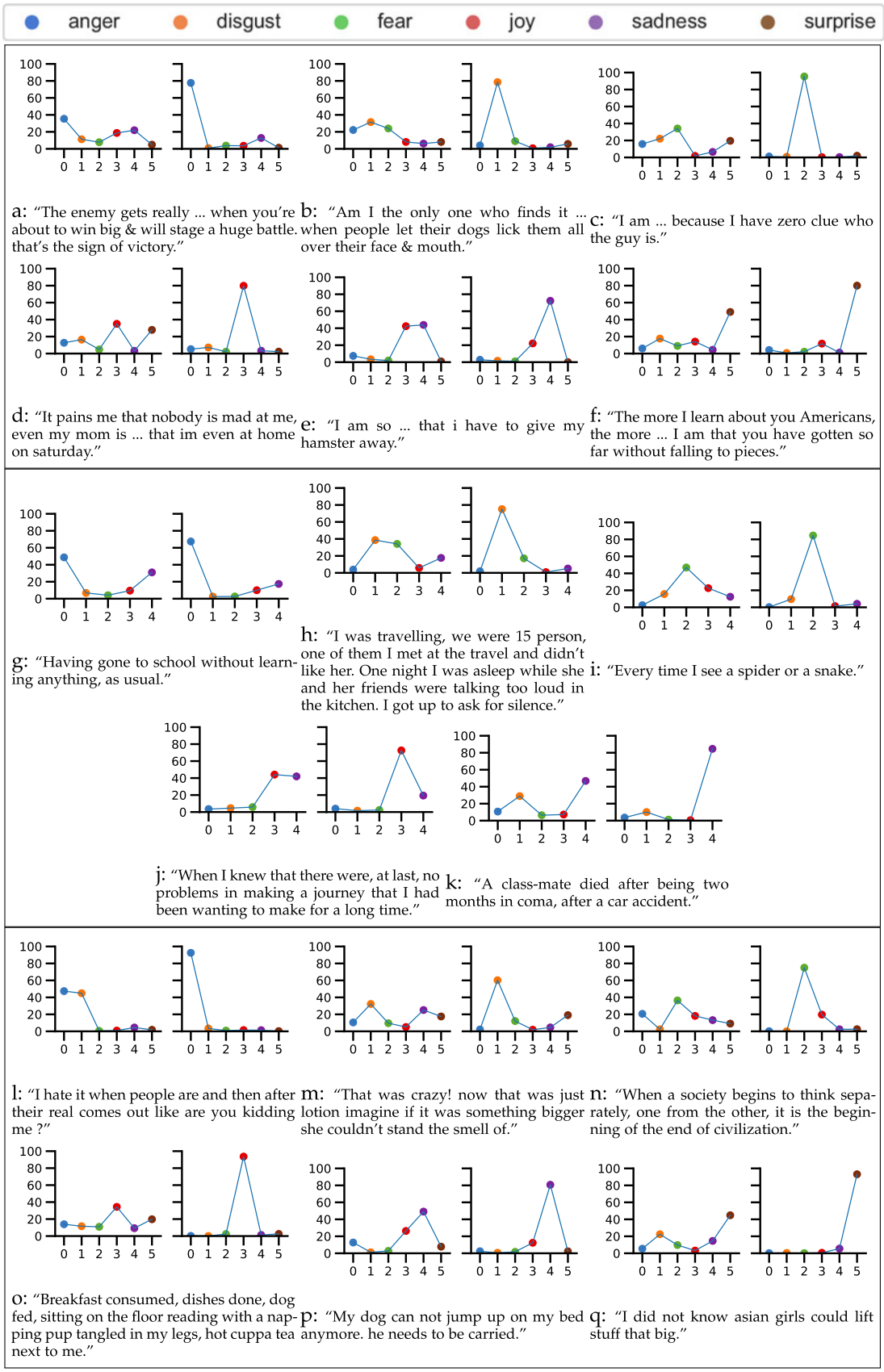

Fig. 2. Prediction scores (y-axis) across emotions (x-axis). Each sub-figure shows the scores of the two evaluated objectives, i.e., without VTCL (left) versus with VTCL (right). The corresponding instance to be classified is included at the bottom of each sub-figure. The frames from top-to-bottom represents instances belonging to IEST, ISEAR and TEC datasets, respectively. "...": refers to the removed triggered word from the IEST dataset.

TABLE 7
Analysis of the Model Predictions Trained on Two Settings (i.e., w/o VTCL Versus w/ VTCL)

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

The actual label for each example is also included.

and TEC datasets, respectively. In the top group of sub-figures (corresponding to examples from the IEST dataset), 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. Fig. 2b shows the scores of the "disgust" class (i.e., without versus with VTCL), demonstrating 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 dataset (i.e., IEST) and to the ISEAR and TEC datasets, 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 dataset 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 dataset 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 datasets. 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

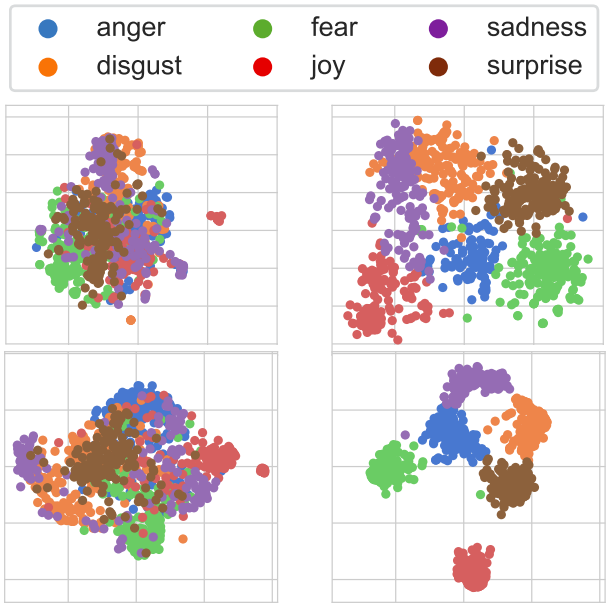

Fig. 3. t-SNE feature visualisation of CNN (top graph) and BERT (bottom graph). The left- and right-hand graphs illustrate features of the model trained without VTCL and with VTCL, respectively. All four plots share the same colour scheme, as defined at the top of the figure.

layer of BERT and CNN, and then used t-SNE[8] [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 versus with VTCL).

Fig. 3 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 it achieves the highest results when trained jointly with our method. Overall, the visualisations serve to reinforce the benefits of our method in terms of decreasing intra-class variance between examples sharing the same emotion as well as increasing their inter-class variances with other emotions.

6.4 Selecting the Number of Negative Centres

Fig. 4 presents the results of selecting varying numbers of negative centres for each dataset. It should be noted that ISEAR contains a maximum of four negative centres, because it only consists of five classes of Ekman's [48] basic emotions (i.e., anger, disgust, fear, joy and sadness).

Fig. 4 shows that the greater the number of negative centres that are combined together in the computation of inter-class distance, the better the performance; the same

8. We use the scikit-learn library [54] to generate the t-sne visualisation and follow the default setting.

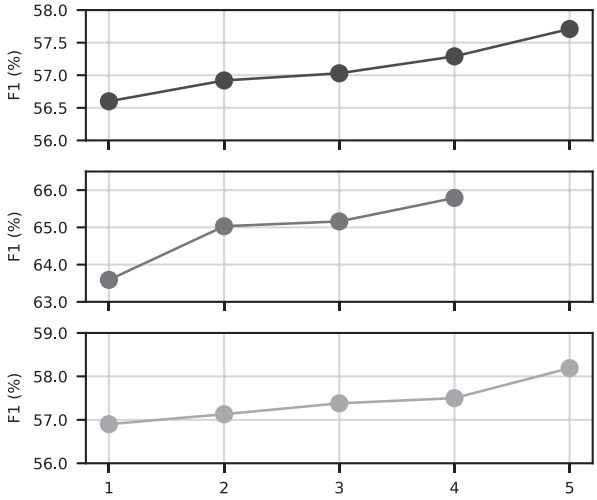

Fig. 4. Our method with a range of C negative centres (x-axis). For the computation of inter-class, the numbers from 1-4 represents the top-k negative centres, while the last one combines all negative centres via summation (i.e. VTCL). The sub-figures from top to bottom represent IEST, ISEAR and TEC datasets, respectively.

trend is observed across all three datasets. These findings also confirm our hypothesis that combining all negative centres (i.e., VTCL) via summation helps to simplify our method as well as to ensure that the intra-class distance is minimised within the same emotion class, while the inter-class distance is maximised between different emotion classes. In other words, our method optimises the inter-class distance to be larger than the intra-class distance plus the margin. In short, our method addresses the problem of selecting the nearest negative centre, which is the case in TCL, by combining all negative centres. This is especially beneficial for TER, where multiple centres can be potentially used as negative ones.

7 Conclusion

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 performance as well as to introduce discriminative features. Our experiments also demonstrated the advantages and utility of VTCL as an auxiliary loss for emotion classification.

Emotion recognition as a single-label classification problem often focuses on maximising the probability of the correct emotion label, but overlooks the important role of intra- and inter-class variations within and between emotions. However, in this work, we have shown the effectiveness of incorporating both intra- and inter-class information into TER, demonstrating the ability of this information not only 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.

ACKNOWLEDGMENTS

The authors would like to thank Paul Thompson and Fenia Christopoulou for their valuable comments and suggestions. The authors would also like to thank the anonymous reviewers for their careful reading and insightful feedback. The work of Hassan Alhuzali was supported by a doctoral fellowship from Umm Al-Qura University.

REFERENCES

[1] H. Alhuzali and S. Ananiadou, "SpanEmo: Casting multi-label emotion classification as span-prediction," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2021, pp. 1573–1584.

[2] S. Akhtar, D. Ghosal, A. Ekbal, P. Bhattacharyya, and S. Kurohashi, "All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2019.2926724.

[3] R. Klinger, O. de Clercq, S. Mohammad, and A. Balahur, "IEST: WASSA-2018 implicit emotions shared task," in *Proc. 9th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, 2018, pp. 31–42.

[4] H. Alhuzali, M. Abdul-Mageed , and L. Ungar, "Enabling deep learning of emotion with first-person seed expressions," in *Proc. 2nd Workshop Comput. Model. People's Opin., Pers., Emot. Social Media*, Jun. 2018, pp. 25–35. [Online]. Available: https://www.aclweb.org/anthology/W18–1104

[5] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word–emotion association lexicon," *Comput. Intell.*, vol. 29, no. 3, pp. 436–465, 2013.

[6] S. M. Mohammad, "#Emotional tweets," in *Proc. 1st Joint Conf. Lexical Comput. Semantics–Volume 1: Proc. Main Conf. Shared Task, Volume 2: Proc. 6th Int. Workshop Semantic Eval.*, 2012, pp. 246–255.

[7] D. Tang, B. Qin, T. Liu, and Z. Li, "Learning sentence representation for emotion classification on microblogs," in *Proc. Int. Conf. Natural Lang. Process. Chin. Comput.*, 2013, pp. 212–223.

[8] W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth, "Harnessing twitter "big data" for automatic emotion identification," in *Proc. Int. Conf. Privacy, Secur., Risk Trust, Int. Conf. Social Comput.*, 2012, pp. 587–592.

[9] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," in *Proc. ACM Symp. Appl. Comput.*, 2008, pp. 1556–1560.

[10] S. Aman and S. Szpakowicz, "Identifying expressions of emotion in text," in *Proc. Int. Conf. Text, Speech Dialogue*, 2007, pp. 196–205.

[11] H. Alhuzali and S. Ananiadou, "Improving classification of adverse drug reactions through using sentiment analysis and transfer learning," in *Proc. 18th BioNLP Workshop Shared Task*, Aug. 2019, pp. 339–347. [Online]. Available: https://www.aclweb.org/anthology/W19–5036

[12] M. E. Aragón, A. P. López-Monroy, L. C. González-Gurrola, and M. Montes-y-Gómez, "Detecting depression in social media using fine-grained emotions," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 1481–1486.

[13] X. Chen, M. D. Sykora, T. W. Jackson, and S. Elayan, "What about mood swings: Identifying depression on twitter with temporal measures of emotions," in *Proc. Companion Proc. Web Conf.*, 2018, pp. 1653–1660.

[14] H. Khanpour and C. Caragea, "Fine-grained emotion detection in health-related online posts," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 1160–1166.

[15] S. Volkova and Y. Bachrach, "Inferring perceived demographics from user emotional tone and user-environment emotional contrast," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 1567–1578.

[16] S. M. Mohammad and S. Kiritchenko, "Using nuances of emotion to identify personality," 2013, *arXiv:1309.6352*.

[17] H. Rashkin, A. Bosselut, M. Sap, K. Knight, and Y. Choi, "Modeling naive psychology of characters in simple commonsense stories," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2289–2299.

[18] P. Fung, "Robots with heart," *Sci. Amer.*, vol. 313, no. 5, pp. 60–63, 2015.

[19] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 2000.

[20] C. Voigt, B. Kieslinger, and T. Schäfer, "User experiences around sentiment analyses, facilitating workplace learning," in *Proc. Int. Conf. Social Comput. Social Media*, 2017, pp. 312–324.

[21] C. S. Montero and J. Suhonen, "Emotion analysis meets learning analytics: Online learner profiling beyond numerical data," in *Proc. 14th Koli Calling Int. Conf. Comput. Educ. Res.*, 2014, pp. 165–169.

[22] Q. Li and S. Shah, "Learning stock market sentiment lexicon and sentiment-oriented word vector from stocktwits," in *Proc. 21st Conf. Comput. Natural Lang. Learn.*, 2017, pp. 301–310.

[23] Y. Mansar, L. Gatti, S. Ferradans, M. Guerini, and J. Staiano, "Fortia-FBK at SemEval-2017 task 5: Bullish or bearish? Inferring sentiment towards brands from financial news headlines," in *Proc. 11th Int. Workshop Semantic Eval.*, 2017, pp. 817–822.

[24] B. Liu, R. Govindan, and B. Uzzi, "Do emotions expressed online correlate with actual changes in decision-making?: The case of stock day traders," *PloS One*, vol. 11, no. 1, 2016, Art. no. e0144945.

[25] Y. B. Alaluf and E. Illouz, "Emotions in consumer studies," in *The Oxford Handbook of Consumption*. New York, NY, USA: Oxford Univ. Press, 2019, p. 239.

[26] J. Herzig, G. Feigenblat, M. Shmueli-Scheuer , D. Konopnicki, and A. Rafaeli, "Predicting customer satisfaction in customer support conversations in social media using affective features," in *Proc. Conf. User Model. Adapt. Personalization*, 2016, pp. 115–119.

[27] S. M. Mohammad and F. Bravo-Marquez, "Emotion intensities in tweets," in *Proc. 6th Joint Conf. Lexical Comput. Semantics*, Vancouver, BC, Canada, 2017, pp. 65–77.

[28] S. Poria, A. Gelbukh, E. Cambria, A. Hussain, and G.-B. Huang, "EmoSenticSpace: A novel framework for affective common-sense reasoning," *Knowl.-Based Syst.*, vol. 69, pp. 108–123, 2014.

[29] L.-A.-M. Bostan and R. Klinger, "An analysis of annotated corpora for emotion classification in text," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 2104–2119.

[30] J. S. Y. Liew and H. R. Turtle, "Exploring fine-grained emotion detection in tweets," in *Proc. 16th Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2016, pp. 73–80.

[31] R. Xia and Z. Ding, "Emotion-cause pair extraction: A new task to emotion analysis in texts," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1003–1012.

[32] A. Agrawal, A. An, and M. Papagelis, "Learning emotion-enriched word representations," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 950–961.

[33] H. Alhuzali, M. Elaraby, and M. Abdul-Mageed, "UBC-NLP at IEST 2018: Learning implicit emotion with an ensemble of language models," in *Proc. 9th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, Oct. 2018, pp. 342–347. [Online]. Available: https://www.aclweb.org/anthology/W18–6250

[34] M. Abdul-Mageed and L. Ungar, "EmoNet: Fine-grained emotion detection with gated recurrent neural networks," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 718–728.

[35] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann, "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1615–1625.

[36] J. Islam, R. E. Mercer, and L. Xiao, "Multi-channel convolutional neural network for twitter emotion and sentiment recognition," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, , 2019, pp. 1355–1365.

[37] E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen, "CARER: Contextualized affect representations for emotion recognition," in *Proc. Empirical Methods Natural Lang. Process.*, 2018, pp. 3687–3697. [Online]. Available: http://aclweb.org/anthology/D18–1404

[38] X. He, Y. Zhou, Z. Zhou, S. Bai, and X. Bai, "Triplet-center loss for multi-view 3D object retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1945–1954.

[39] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.

[40] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 499–515.

[41] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Oct. 2014, pp. 1746–1751. [Online]. Available: https://www.aclweb.org/anthology/D14–1181
[42] T. Wolf *et al.*, "HuggingFace's transformers: State-of-the-art natural language processing," 2019, *arXiv:1910.03771*.
[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.
[44] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.
[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[46] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances Neural Inf. Process. Syst.*, vol. 32, pp. 8026–8037, 2019.
[47] K. R. Scherer and H. G. Wallbott, "Evidence for universality and cultural variation of differential emotion response patterning," *J. Pers. Social Psychol.*, vol. 66, no. 2, pp. 310–328, 1994.
[48] P. Ekman, "An argument for basic emotions," *Cogn. Emot.*, vol. 6, no. 3–4, pp. 169–200, 1992.
[49] C. Baziotis, N. Pelekis, and C. Doulkeridis, "DataStories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis," in *Proc. 11th Int. Workshop Semantic Eval.*, 2017, pp. 747–754.
[50] Y. Zhang, J. Fu, D. She, Y. Zhang, S. Wang, and J. Yang, "Text emotion distribution learning via multi-task convolutional neural network," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 4595–4601.
[51] S. Tripathi, A. Ramesh, A. Kumar, C. Singh, and P. Yenigalla, "Learning discriminative features using center loss and reconstruction as regularizer for speech emotion recognition," in *Proc. Workshop Artif. Intell. Affect. Comput.*, 2020, pp. 44–53.
[52] R. Dror, G. Baumer, S. Shlomov, and R. Reichart, "The hitchhiker's guide to testing statistical significance in natural language processing," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2018, pp. 1383–1392. [Online]. Available: https://aclanthology.org/P18–1128
[53] R. Gaonkar, H. Kwon, M. Bastan, N. Balasubramanian, and N. Chambers, "Modeling label semantics for predicting emotional reactions," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2020, pp. 4687–4692.
[54] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
[55] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

Hassan Alhuzali received the MS degree in information science from Indiana University-Bloomington, USA, in 2016. He is currently working toward the PhD degree with the Department of Computer Science, The University of Manchester. He was a visiting student with Positive Psychology Centre, UPENN, USA, and UBC, CA. His research interests include natural language processing and text mining, with a focus on affective computing and textual emotion recognition/analysis.

Sophia Ananiadou is currently the director of the UK National Centre for Text Mining and a professor of computer science with The University of Manchester. She is also a turing fellow with Alan Turing Institute and a visiting professor with AIST/AIRC. Her research interests include the areas of biomedical text mining and natural language processing.