TERMS: textual emotion recognition in multidimensional space

Yusra Ghafoor[1] · Shi Jinping[2] · Fernando H. Calderon[1] · Yen-Hao Huang[2] · Kuan-Ta Chen[3] · Yi-Shin Chen[4]

Abstract

Microblogs generate a vast amount of data in which users express their emotions regarding almost all aspects of everyday life. Capturing affective content from these context-dependent and subjective texts is a challenging task. We propose an intelligent probabilistic model for textual emotion recognition in multidimensional space (TERMS) that captures the subjective emotional boundaries and contextual information embedded in a text for robust emotion recognition. It is implausible with discrete label assignment;therefore, the model employs a soft assignment by mapping varying emotional perceptions in a multidimensional space and generates them as distributions via the Gaussian mixture model (GMM). To strengthen emotion distributions, TERMS integrates a probabilistic emotion classifier that captures the contextual and linguistic information from texts. The integration of these aspects, the context-aware emotion classifier and the learned GMM parameters provide a complete coverage for accurate emotion recognition. The large-scale experimentation shows that compared to baseline and state-of-the-art models, TERMS achieved better performance in terms of distinguishability, prediction, and classification performance. In addition, TERMS provide insights on emotion classes, the annotation patterns, and the models application in different scenarios.

1 Introduction

With the emergence of social media, a vast amount of big heterogeneous data is generated on various platforms, where users express their opinions regarding almost all aspects of everyday life. An essential piece of information that could be extracted from this user-generated data is the emotional content, which provides very expressive aspects of human lives [1]. In big heterogeneous data (texts, images, videos, and audio), the text is one of the most abundant and effective mediums for understanding emotions. It is succinct in the expressing of opinions; for example, microblogs contain a high density of relevant, sentiment-bearing terms that are readily accessible [2]. In microblogs, such as those found on Facebook and Twitter, emotions are expressed via short and direct text messages containing individual opinions that make them particularly valuable

sources of information for effective emotion recognition [3]. Mining emotions from these large volumes of textual opinions in microblogs can provide expressive information for understanding collective human behavior that can be extremely valuable in many domains, such as product review analysis [4], marketing campaigns [5], political stance detection [6, 7], healthcare [8, 9], stock market analysis [10] etc. Therefore, intelligent textual emotion recognition systems applicable to microblogs are highly desirable.

A great deal of research has been conducted on emotion recognition and the classification of microblog texts, which can be broadly categorized into two computational directions: deterministic and dimensional models [11]. Deterministic models use a discrete and finite set of emotion labels that most fit a given text, based on the strength of the predicted emotion [12, 13]. Such discrete emotion labels are generally taken from pioneering models, such as those of Ekman [14] and Plutchik [15] that specify the primary emotions.

However, the deterministic approach associates each text with a discrete label without the attribute of personalization; in other words, it does not capture personal differences

✉ Yi-Shin Chen
yishin@gmail.com

Extended author information available on the last page of the article.

as the definition of emotions can differ for each individual based on their background and culture. Dimensional models, on the other hand, are flexible in personalizing emotions in terms of valence, arousal, and other dimensions. The dimensional models project each emotion as coordinates in a space of continuous dimensions of valence and arousal as numerical values. Valence (x-axis) represents the pleasantness of a stimulus, and arousal (y-axis) shows the intensity of an emotion provoked by a stimulus [16, 17]. Any affective experience can be expressed as a combination of these two independent dimensions, which is then interpreted as representing a particular emotion [18]. This method enables a personalized and quantified analysis as the emotions are projected in a multidimensional numeric space, which is effective and useful for analyzing the fuzzy boundaries for different emotion features.

The texts on microblogging sites are usually written in a casual style in which the short length and inconsistent language make it difficult to completely recognize and predict the affective information [3]. We anticipated that dealing with informal and ambiguous texts would be crucial in designing a model for accurately identifying emotions in microblogs. Designing such a model, however, is fairly challenging because of the following reasons [19]. First, user-generated text in a microblog may contain linguistic variations and contextual information. For instance, in the text "*Thanxxx mom for cooking the same meal every day,*" the word "thanxxx" is a linguistic variation of thanks, which requires an understanding of the semantic similarity between the two terms. In addition, the term "thanks" is usually associated with joy or a positive sentiment, but in this instance, it refers to an annoyance. Therefore, to accurately classify the user's intended emotions, it is crucial to consider *contextual information*. Second, user-generated text in a microblog can be highly opinionated and subjective in nature, where users may perceive different emotions from the same text [20, 21]. For example, the text "*the virus is spreading*" can communicate emotional states of both fear and sadness, which is partially dependent on the reader's state of mind. Therefore, capturing the *varying emotional perceptions* and fuzzy emotional boundaries is essential for personalized and complete coverage of possible emotional content embedded in a text.

To address this problem, we propose a probabilistic model for textual emotion recognition in multidimensional space **(TERMS)**, which takes the contextual information and subjective nature of the microblog text into account for emotion recognition. The contextual information requires additional details from a text to interpret the given information such as the topic, structure, patterns and sentiment orientation. In view of this, TERMS introduces a probabilistic context-aware emotion classifier that takes syntactic structure and semantic meaning of a text into account to expose the relevant contextual and linguistic information. The syntactic structures automatically captures the pattern of the text via a graph-based algorithm and further enriches them with embeddings to gather semantic content. Second, to cater to the subjectivity of emotions, it is known that emotional perceptions are inherently subjective and cannot be covered by a single point or discrete emotion label. Therefore, we consider varying perceptions and generate them as distributions. A distribution is the exhibition of multiple perspectives and better reflects the nuances of emotion content embedded in a text. TERMS maps the multiple emotional perspectives of every single text as distributions (numerical values) into a multidimensional space, which better personalizes the emotion variations. TERMS models the subjective emotion content of the text as a probabilistic emotion distribution through a Gaussian mixture model (GMM) and learns its parameters for a soft assignment. To effectively recognize emotions, TERMS integrates the proposed context-aware emotion classifier and the GMM modeled probability emotion distribution to describe the emotions thorough low-level textual feature space and high-level emotion space, respectively. Moreover, due to its probabilistic and generative nature, TERMS is conveniently scalable, and assigns soft labels in a multidimensional space.

To our knowledge, a model of this kind that caters to subjectivity by parameterizing emotion distributions in an emotional space has been only applied to music excerpts [22, 23] and speech [24]. Modeling texts has been challenging due to their single modal nature that does not provide added information of tone, expression and prosody to understand the full emotional content as compared to the rich representation of music and speech. The challenges are further escalated owing to microblogs' self-focused topics, short and informal writing format. For microblog texts, the TERMS integrated approach is a novel attempt to model varying perspectives as distributions in emotion space. We cater to these challenges through a context-aware classifier and personalized emotion distributions in TERMS. The main contributions of the article are summarized as follows:

- We propose TERMS, a probabilistic model for textual emotion recognition in multidimensional space, which takes the contextual information and subjective nature of a microblog's text into account.
- We propose the soft modeling of the affective content in a multidimensional space by parameterizing the emotion distribution through a GMM, which provides insight into dealing with subjectivity and indistinct emotional boundaries.
- TERMS integrated approach enhances emotion recognition by estimating emotional weightage combined with multiple emotional perceptions for each text, thus

taking complete advantage of both the models, deterministic and dimensional.

– We annotated our collected data by different annotators in order to conduct large-scale simulations to evaluate the performance of TERMS. Our simulation results show that compared to baseline and state-of-the-art models, TERMS achieves better distinguishability, prediction, and classification performance.

The rest of the article is organized as follows. Section 2 summarizes the related work. Section 3 presents the overview, preliminaries, and the proposed probabilistic TERMS model. Section 4 describes the evaluation, comparative models, performance metrics, setup, and the overall results. Section 5 discusses the predicted results and the impact of annotators' number on model's prediction performance and Section 6 concludes the paper.

2 Related work

Affective computing is an established research field that is burgeoning due to its relevance in many application domains desiring the feature of emotion recognition from different forms of user-generated data such as texts, music, speech, and images [11]. Two of the driving interrelated factors in this flourishing field are social networks and microblogs. Microblogs provide an effective platform for emotion recognition as they provide a wide variety of self-focused topics published in real time [2]. The texts are explicit and succinct with relatively clear projections of users' emotions. The focused nature and higher density of affective terms make these platforms highly useful for emotion recognition as compared to topic-based platforms, such as product and movie reviews [3]. The work on microblog text emotion recognition can be broadly divided into two categories, deterministic and dimensional models. We provide a comprehensive survey on these two categories in this section.

2.1 Deterministic models

There is a substantial well-vetted body of research on microblog emotion recognition, which focuses on classifying texts into a set of discrete emotion classes [25, 26]. Deterministic models utilize supervised, unsupervised, or semi-supervised methods by employing statistical models, such as *machine learning* and *deep learning*.

Using machine-learning models, Meo and Sulis [12] considered structural and lexical-based features from text to automatically identify affective content and compared the results with latent factors and traditional classifiers.

Suttles and Ide [27] recognized emotions from texts based on Plutchik's eight emotional classes by applying distant supervision. Perikos and Hatzilygeroudis [28] used an ensemble classifier schema by combining knowledge-based and statistical machine-learning classification methods for the automatic identification of emotions in text. Symeonidis et al. [29] applied soft computing techniques, namely NB, support vector machines (SVM), logistic regression, and convolution neural networks (CNN) for analyzing emotions. Recent significant additions in emotion classification domain are two of the largest and dynamic emotion corpora, GoEmotions and Vent [30, 31]. GoEmotions is the manually annotated dataset for 58k English Reddit comments, labelled for 27 emotion categories by the readers [30]. Likewise, the Vent dataset contains more than 33M comments from the social media sites, tagged with 705 emotions explicitly by the writer [31]. These datasets are widely being used in the recent academic works [32, 33].

Regarding research with deep learning models, the performance of textual emotion recognition tasks is enhanced due to statistically rich and granular framework of deep learning models [34]. Abdul-Mageed and Ungar [13] proposed a model named Emonet to predict emotions into eight emotional classes based on the gated recurrent neural networks algorithm (GRNN). Another renowned emotion prediction model is DeepMoji presented by Felbo et al. [35], which was trained on billions of emoji-labeled tweets for affective modeling and recognition. Rosenthal et al. [36] identified the sentiment of tweets as per the challenge of SemEval-2017 Task 4: Sentiment Analysis in Twitter. The same series provided SemEval-2018 Task 1: Affect in Tweets, a challenge that organized a subtask of multi-label emotion classification in which teams used state-of-the-art methodologies to predict emotions from microblog affective content [37]. Zhang et al. [38] implemented a multi-layer CNN with an attention mechanism that modelled context representations to perform target-dependent sentiment classification. Sadr et al. [39] proposed a multi-view deep network that takes into account intermediate features extracted from convolutional and recursive neural networks to enhance classification performance. The deep-learning models are effective; however, they require complex computations and extensive training data for better performance, while our proposed model is relatively simple and performs well even on limited data.

2.2 Dimensional models

Another significant way to represent affective states is *dimensional models*, which provide a continuous fine-grained alternative for conducting affective text analysis [11]. These models contribute in understanding the

conveyance of emotions through language and how the emotional dimensions influence people's behaviour [40]. Russell [41] proposed a dimensional representation model of affect, named the circumplex model, that distinguishes three components: valence, arousal, and dominance (VAD). Studies have shown the modeling of affective states on a valence and arousal map by adopting varying machine-learning approaches [16, 42] and lexicon-based methods [43, 44]. Hasan et al. [45] present a model for real-time emotion tracking by employing [46] and developing an EmotexStream framework. Preotiuc-Pietro et al. [18] predicted valence and arousal on Facebook posts by performing linear regression and released an expert annotated dataset. Mohammad and Bravo-Marquez [47] provided the first emotion intensity dataset (EmoInt) using a best-worst scaling technique. Buechel and Hahn [17, 48] published a benchmark dataset called Emobank (10548 sentences) in which each sentence was manually annotated on the VAD dimensions. Recent studies proposed frameworks that learn from Emobank, the categorical emotion annotations corpus to predict continuous VAD scores [49, 50]. Cheng et al. proposed a Bi-directional Long Short-Term Memory (BiLSTM) model that identifies and forecasts the sentiment information in terms of VA-values and integrated it into a deep learning model to optimise Government social management [51]. Another recent experimental work aimed at testing the role of five emotions (valence, arousal, dominance, approach-avoidant, and uncertainty) on the intervention effect of the Learning Mindset study [52]. The SemEval-2018 Task 1: Affect in Tweets challenge asked for the prediction of intensities (arousal) and valence from a stream of texts in terms of regression and ordinal classification [37, 53]. The winning team [54] proposed a unified architecture for both subtasks by using an ensemble of multiple prediction models and heterogeneous feature extraction methods. Dimensional models provide useful measures of emotions; however, they are unable to capture varying perceptions of emotions, which are subjective and might differ regarding the affective content of the same text.

To address the subjective nature of emotion perceptions, the extension of VA-based models were proposed where the representation of emotions was transformed to probability distributions from points on VA-emotional space [55]. In view of this, recent studies used *Gaussian* parameter-based approaches to estimate emotion distributions on the VA-space that take into account covariance information along with the mean [22, 56]. This approach estimates emotion distribution as a Gaussian with integrated methods. Zhao et al. [57] presented a work that predicts an image's continuous probability distribution by using a GMM in a VA-space. Another work by Sun et al. [58] aimed to unify discrete and dimensional emotion models by introducing a typical fuzzy emotion subspace for affective video

content analysis. Gaussian distributions in dimensional models have also been widely applied in music-listening behavior analysis [22, 23, 59]. In the work conducted by Wang et al. [22], an acoustic GMM was employed to classify music with the utilization of valence and arousal, which increased the accuracy of acoustic classification. Applications of such an approach have also been widely adopted for speech emotion recognition [24, 60, 61]. However, to our knowledge, the Gaussian parameter-based approach has not been applied to microblog texts, which motivated us to personalize this approach for textual emotion recognition. The transformation in mediums has been challenging due to the single modal nature of texts that contain little information to apprehend underlying emotions and intensities relative to speech and music, which are enriched with emotional cues such as tone, expression, accent, prosody etc. The single-mode of information can impact the classification task and annotations. We address this issue by proposing a context-aware emotion classifier with a GMM in the VA-space, which captures the nuances of embedded emotions and varying perceptions in a text.

3 The probabilistic TERMS model

For the purposes of this discussion, the text in a microblog refers to a single statement posted by a locutor. A locutor in this article refers to the person who is writing a text. The text is an expression that reflects the emotionl state of the locutor. The text can be a thought, mood, or an opinion of a locutor based on his or her prevailing emotional state. The emotions relevant for this study are the emotions felt by the locutor that were embedded in the writing of the posted text. The proposed model aims to recognize these embedded emotions from the texts.

The texts posted on microblogs are enriched with emotions, which are seemingly succinct and straightforward. It can be assumed or misunderstood that these explicit texts can be conveniently assigned eight emotion classes defined by Plutchik [15]. The emotional classes are anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. However, a given tweet contains complex granular details and is embedded with (i) contextual information and (ii) multiple perspectives; thus, it is not easy to classify a microblog's text with a straightforward emotion allocation approach.

This study proposes solution to these problems starting with preliminaries in Section 3.1. TERMS is designed to address these problem through three major modules. The first module textual emotion classification (EmoClass) solves the first issue with the help of a context-aware classifier that estimates the emotion probabilities for each class based on syntactic templates and word embeddings (elaborated in Section 3.2). To handle the second issue,

TERMS proposes emotion GMM (EmoGMM) that maps the multiple perspectives of each tweet into VA-space via a GMM and learns its parameters (detailed in Section 3.3) and lastly, the third prediction module jointly exploits the emotion probability and the learned parameters of the GMM to predict the emotion distribution of the text. This is clearly explained in Section 3.4.

3.1 Preliminaries

Before introducing the details of our approach, we highlight a few notable concepts that are useful in understanding it. For clarity, the notations used are explained in Table 1.

We denote the microblog texts as $X = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$, where $x^{(i)} \in \mathbb{R}^M$ represents an M-dimensional feature vector of a tweet i. Let z be the associated discrete emotion or affective state, where $z \in \{1, 2, 3\ldots, K\}$. Consider that (v, a) represents a pair of valence and arousal values (or simply, *VA-value*), where v denotes a valence value and a represents an arousal value. Since each individual tweet is rated by different annotators for diverse perspectives (as explained in Section 4.1); we thus denote the valence and arousal ratings (or simply, VA-ratings) by $y = (v, a)$. y is the position of the text $x^{(i)}$ on the multidimensional VA-space.

Let $g(\theta)$ be an arbitrary probability density function (PDF), parameterized by θ. If the valence and arousal ratings $y = (v, a)$ obey the PDF $g(\theta)$, then an *emotion distribution* is defined as follows:

$$y \sim g(\theta). \tag{1}$$

Since the emotions defined in VA-space are described by a distribution, (1) can be expressed via a mixture model as

follows:

$$p(y) = \sum_{k=1}^{K} \pi_k g(\theta_k), \tag{2}$$

which illustrates that the emotion distribution of a text is a linear combination of K emotion probabilities, where $g(\theta_k)$ is the k-th emotion distribution, termed as the k-th component of the mixture. π_k is called the mixing coefficient, representing the emotion probabilities of the k-th component.

To combine these distributions, we employ the widely used GMM that combines the K Gaussian distributions, also referred to as mixtures of Gaussians. On this account, $g(\theta_k)$ is specified as a bivariate Gaussian distribution as it maps emotions into a two-dimensional VA-space. The reasons for employing the GMM are as follows: (i) the GMM is able to approximate almost any continuous PDF to arbitrary accuracy by using a sufficient number of Gaussians (K) and by adjusting their parameters (θ_k) as well as the coefficients (π_k) [62]; and (ii) the continuous text ratings are well modeled by the GMM, considering they follow a bivariate Gaussian distribution. To verify (ii), we tested if the VA-ratings of each text from different annotators were similar to the bivariate Gaussian distribution. The Mardia multivariate normality test [63] with a significance level of 0.05 was performed on our data to determine the adequacy of the GMM for modeling the emotion distributions. The results achieved were 100%, asserting that all the texts were similar to the bivariate Gaussian distribution, thus making the GMM an obvious and favorable choice.

Technically, TERMS follows a graphical approach with the form $X \rightarrow z \rightarrow y$. $X \rightarrow z$ is carried out via textual emotion recognition through our proposed classifier that outputs the posterior probability of texts into selected affective classes z (as detailed in the next section). $z \rightarrow y$ is the emotion GMM modeling on a VA-space. It maps the associated emotion classes z into VA-space by parameterizing emotion distributions (as described in Section 3.3). The process flow of the TERMS probabilistic model is demonstrated in Fig. 1.

3.2 Textual emotion classification (EmoClass)

To classify the microblogs texts into emotions, $X \rightarrow z$, we propose a classifier named the *Emotion Classifier (EC)*. that outputs the posterior probability of the texts into selected affective classes z, as in (3). Adding posterior probabilities to emotion distributions would enrich the distributions with linguistic and contextual information.

$$p(z = k | x^{(i)}) \sim EC. \tag{3}$$

We refer to the probabilities accumulated from the texts by the emotion classifier as the *emotion probability*. This

Table 1 Notations' Table

| Notation	Description
X	Feature vector of all texts
$x^{(i)}$	Feature vector of text i
$g(\theta_k)$	Emotion distribution for emotion k
π_k	Mixing coefficient for emotion k
μ_k	Mean of k-th emotion distribution
Σ_k	Covariance matrix of k-th emotion distribution
$\mathcal{N}$	Gaussian distribution
Y	Labelled valence and arousal dataset
$y_j^{(i)}$	The i-th text rated by j-th annotator
N_{Ai}	Number of annotators for text i
$\mathcal{L}$	Labelled dataset of texts and VA-ratings
μ^{pre}, Σ^{pre}	Predicted mean and covariance for a text
K	Number of emotional classes
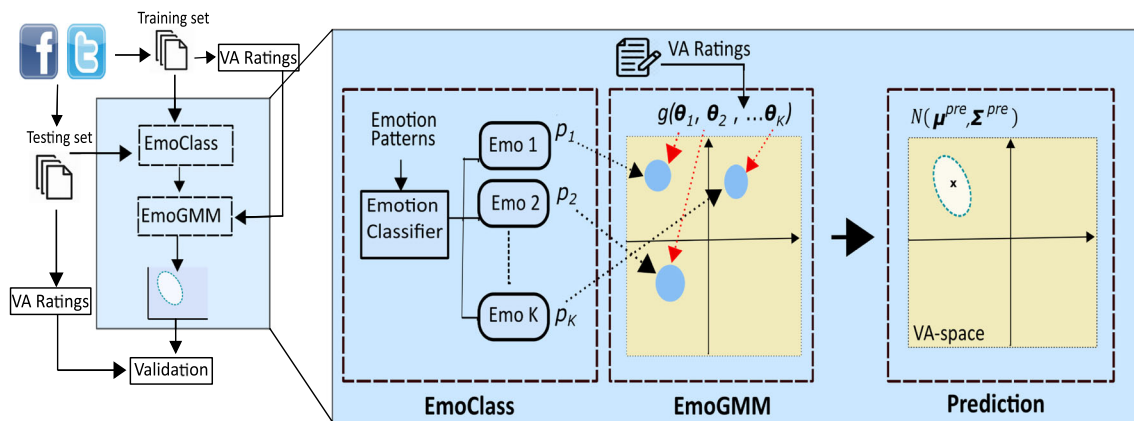
Fig. 1 An illustration of the TERMS probabilistic process. EmoClass is a textual emotion classification module that outputs emotion probabilities for each text into specified affective classes. EmoGMM is an emotion GMM modeling that takes in the probabilities and combines them with VA-ratings to parameterize emotion distributions in a VA-space. The prediction module employs a single affective Gaussian on weighted GMMs to predict an emotion distribution for each unseen text

part of the TERMS model is referred to as textual emotion classification, or *EmoClass*. In the following, we explain the proposed emotion classifier.

Emotion Classifier: To estimate the emotion probabilities $p(z = k|\boldsymbol{x}^{(i)})$, we generalize an emotion classifier from our previous works, Saravia et al. [64, 65]. We employ this classifier as it provides an in-depth contextual information through syntactic templates. For a given text, the classifier assigns probabilities to each associated emotion class z, according to affinity based on the context-aware emotion pattern extracted from the text. Specifically, it is a graph-based algorithm, which constructs syntactic templates from the corpus to extract context-aware emotion patterns. We refer to these features as context-aware as they take syntactic structures and semantic meaning of a text in account to construct pattern-based emotion features. The syntactic structures offered by a graph construction is useful to automatically expose the relevant linguistic information (i.e., contextual and latent information) from a large-scale emotion corpus, whereas to capture and preserve the semantic relationships between patterns, we implement word embeddings on the extracted patterns. This is followed by emotion probability computation, where each pattern is assigned a weight. The weight identifies the relevance of a pattern to an emotion category. In the context of emotion classification, patterns and their weights play the role of features.

The graph-based emotion feature extraction algorithm is summarized in the following steps:

a) **Graph construction.** Given an emotion corpus, we construct a graph $G(V; A)$, where vertices V are a set of nodes that represent the tokens extracted from the corpus, and edges, denoted as A, represent the relationship of words extracted using a window approach [65]. This will help to retain the syntactic structure of the data. For an arc $a_i \in A$, its normalized weight can be computed as:

$$w(a_i) = \frac{freq(a_i)}{max_{j \in A} freq(a_j)}, \tag{4}$$

where $freq(a_i)$ is the frequency of arc a_i.

Token categorization. To extract the emotion patterns, we divide the syntactic structures into two families of words, connector words (cw) and subject words (sw). This provides the foundation for extracting context-aware emotion patterns as the structures are the sequences of these words. The sw correspond to the words that are high on subjective content, while cw reflect the most frequent words in a text that have high connectivity to influential nodes. To find the cw, we use eigenvector centrality, and to estimate sw, we compute the clustering coefficient elaborated in [65].

Pattern extraction. The syntactic templates constructed based on the cw and sw are applied to the dataset, resulting in the patterns. The subject words in the extracted patterns are replaced with an asterisk ("*"), a proxy to cater to linguistic nuances and unknown words that are not present in the training corpus. Furthermore, it enhances the applicability of the model to other domains as well.

b) **Enriched patterns.** The extracted patterns are enriched with word embeddings to make them pertinent for emotion classification and to capture the perspectives and semantic relationships between patterns. We employ agglomerative clustering to link the patterns

to relevant clusters based on the sw component. The details of this procedure can be found in [65].

To this end, the resulting enriched patterns contain both the semantic information provided by the word embeddings and the contextual information gained through the graph components, hence providing *context-aware emotion patterns*.

c) **Emotion probability.** The enriched emotion patterns are then weighted with respect to each emotion category. It exhibits how relevant a pattern is to the respective emotion category. This outputs the score of each emotion for a given text. We refer to score as the *emotion probability*. It is computed as follows:

$$p(z = k|\boldsymbol{x}^{(i)}) \leftarrow \frac{\exp(-ts_k)}{\sum_{k=1}^{K} \exp(-ts_k)}, \tag{5}$$

where s_k is the score of emotion k computed with a customized version of term frequency-inverse document frequency (tf-idf) proposed in [65], and K is the number of emotions. t is an adjusting coefficient that scales the scores, $0 < t \leq 1$.

3.3 Emotion GMM (EmoGMM)

The subjectivity in emotion perceptions is inherent and can be summarised as emotion distributions. The emotion distribution in the VA-space is described as a bivariate Gaussian distribution with $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$ as its parameters associated with emotion k as

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{6}$$

Since the distribution of $\boldsymbol{y}$ given an emotion class $z = k$ is Gaussian, by following [22] for the rest of analysis, we have

$$p(\boldsymbol{y}|z = k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{7}$$

where the parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are associated with the k-th emotion class as well. This transformation of $z \rightarrow y$ in the VA-space is a second module in TERMS, referred to as *EmoGMM*. It maps the associated emotion classes z into VA-space by parameterizing the emotion distributions.

The probability density function for $\boldsymbol{y}$ is then given by the following:

$$p(\boldsymbol{y}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{8}$$

where π_k is a mixing coefficient, which we reparameterize as

$$\pi_k = p(z = k|\boldsymbol{x}^{(i)}). \tag{9}$$

π_k is set as the computed emotion probability (5) from EmoClass. It is used as the weighted mixing coefficient for modeling EmoGMM. We interpret it as the probability of emotion k for a given text.

For any given text, the emotion distribution is denoted as $p(\boldsymbol{y}|\boldsymbol{x}^{(i)})$. An emotion distribution would be a weighted combination of $\{\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^{K}$ that uses $p(z = k|\boldsymbol{x}^{(i)})$ as the weights. Accordingly, by combining (5), (8), and (9), the emotion distribution of $\boldsymbol{y}$ given text $\boldsymbol{x}^{(i)}$ is

$$p(\boldsymbol{y}|\boldsymbol{x}^{(i)}) = \sum_{k=1}^{K} \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) p(z = k|\boldsymbol{x}^{(i)}), \tag{10}$$

where $\{p(z = k|\boldsymbol{x}^{(i)})\}_{k=1}^{K}$ is the weight of the k-th emotion for a given text $\boldsymbol{x}^{(i)}$, stating the emotion probabilities computed via the proposed emotion classifier. The computed $z = k$ connects the EmoClass to an emotional space by parameterizing the emotion probabilities with a GMM. The process of training a GMM with emotion probabilities as input is referred to as EmoGMM (see Fig. 1). This learning process requires annotated VA-ratings of texts for the GMM estimation, where each text is labeled by multiple annotators. With those VA-ratings and emotion probabilities, $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$ can be estimated by the expectation maximization (EM) algorithm [66]. The EM algorithm has been widely adopted to parameterize emotion distributions for music and speech, but rarely employed to map emotional perceptions in VA-space for a text.

The EM algorithm aims to solve the latent parameter estimation problem in a numerical way. It first computes possible values for the parameters to be estimated by taking expectations on all the known variables, which is called the E-step, and secondly, the M-step maximizes the log-likelihood function with the possible values computed in the E-step. Thus, a clear form of the likelihood function is provided for applying the EM algorithm.

We denote $\boldsymbol{y}_j^{(i)}$ as the i-th text rated by the j-th annotator. $\boldsymbol{Y}^{(i)} = \{\boldsymbol{y}_1^{(i)}, ..., \boldsymbol{y}_{N_{Ai}}^{(i)}\}$ is the set of VA-values rated by the annotators, in which N_{Ai} is the number of annotators for text i. Such VA-values are provided by the annotators for all N texts. Let $\mathcal{L} = \{\boldsymbol{x}^{(i)}, \boldsymbol{Y}^{(i)}\}_{i=1}^{N}$ denote the entire annotated dataset.

We first derive the general form of the posterior probability of $z = k$ given $\boldsymbol{y}$, denoted as follows:

$$\begin{aligned} p(z = k|\boldsymbol{y}) &= \frac{p(z=k)p(\boldsymbol{y}|z=k)}{p(\boldsymbol{y},z=k)}, \\ &= \frac{p(z=k)p(\boldsymbol{y}|z=k)}{\sum_{i=1}^{K} p(z=i)p(\boldsymbol{y}|z=i)}, \\ &= \frac{\pi_k \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{i=1}^{K} \pi_i \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}. \end{aligned} \tag{11}$$

In the E-step, according to (11), we compute the posterior probability given $\boldsymbol{y}_j^{(i)}$, as follows:

$$p(z = k|\boldsymbol{y}_j^{(i)}) = \frac{p(z = k|\boldsymbol{x}^{(i)})\mathcal{N}(\boldsymbol{y}_j^{(i)}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} p(z = k|\boldsymbol{x}^{(i)})\mathcal{N}(\boldsymbol{y}_j^{(i)}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \quad (12)$$

In the M-step, the updating forms for the mean vector and covariance matrix are as follows:

$$\boldsymbol{\mu}_k^{new} \leftarrow \frac{\sum_{i,j} p(z = k|\boldsymbol{y}_j^{(i)})\boldsymbol{y}_j^{(i)}}{\sum_{i,j} p(z = k|\boldsymbol{y}_j^{(i)})}, \quad (13)$$

$$\boldsymbol{\Sigma}_k^{new} \leftarrow \frac{\sum_{i,j} p(z = k|\boldsymbol{y}_j^{(i)})(\boldsymbol{y}_j^{(i)} - \boldsymbol{\mu}_k^{new})(\boldsymbol{y}_j^{(i)} - \boldsymbol{\mu}_k^{new})^T}{\sum_{i,j} p(z = k|\boldsymbol{y}_j^{(i)})}. \quad (14)$$

Thus, (12), (13), and (14) are the iteration forms for estimating $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$. We use $\boldsymbol{\mu}_k^{new}$ and $\boldsymbol{\Sigma}_k^{new}$ to compute the log-likelihood function to check if it converges. The general form of the log-likelihood function is given by

$$\begin{aligned} \ell &= \log \prod_{i=1}^{N} \prod_{j=1}^{N_{Ai}} p(\boldsymbol{y}_j^{(i)}|\boldsymbol{x}^{(i)}), \\ &= \sum_{i,j} \log \sum_k \mathcal{N}(\boldsymbol{y}_j^{(i)}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) p(z = k|\boldsymbol{x}^{(i)}). \end{aligned} \quad (15)$$

The pseudocode of the EM algorithm for estimating the EmoGMM parameters by the VA-ratings is shown in Algorithm 1. The algorithm takes the emotion probability $\{p(z|\boldsymbol{x}^{(i)})\}_{i=1}^{N}$ from EmoClass and $\{\boldsymbol{\mu}_k^0, \boldsymbol{\Sigma}_k^0\}_{k=1}^{K}$ as the inputs along with the number of iterations and stopping criteria and outputs the mean and covariance $\{\boldsymbol{\mu}_k^{new}, \boldsymbol{\Sigma}_k^{new}\}_{k=1}^{K}$ parameters of each emotion distribution. We initialize the log-likelihood function l_0 and iterative parameter r in line 1. The learning loop computes the EM algorithm by estimating the posterior probabilities using (12) and updating the mean vector and covariance with (13) and (14) in lines 2–6. Line 7 halts the loop as per the stopping criteria, while line 8 shows the assignment of the computed mean and covariance to the output parameters, which are utilized to map the emotion distributions in VA-space. We implement Algorithm 1 in its standard complexity of $\mathcal{O}(NK)$, where N is the number of tweets and K is the number of emotion classes while $K << N$.

Algorithm 1 EM for the GMM fitting.

INPUT: Prior $\{p(z|\boldsymbol{x}^{(i)})\}_{i=1}^{N}$; initial model parameters $\{\boldsymbol{\mu}_k^0 = \boldsymbol{\mu}_{\mathcal{L}}, \boldsymbol{\Sigma}_k^0 = \boldsymbol{\Sigma}_{\mathcal{L}}\}_{k=1}^{K}$; maximal number of iterations R; threshold of stopping ratio ϵ;

OUTPUT: Model parameters: $\{\boldsymbol{\mu}_k^{new}, \boldsymbol{\Sigma}_k^{new}\}_{k=1}^{K}$

1: Initialize $r \leftarrow 0$ and ℓ_0 using (15)
2: **repeat**
3: Compute the posterior probability using (12) with $\{\boldsymbol{\mu}_k^r, \boldsymbol{\Sigma}_k^r\}_{k=1}^{K}$;
4: $r \leftarrow r + 1$;
5: Update $\{\boldsymbol{\mu}_k^r, \boldsymbol{\Sigma}_k^r\}_{k=1}^{K}$ using (13) and (14);
6: Compute ℓ_r using (15);
7: **until** $r = R$ or $(\ell_r - \ell_{r-1})/|\ell_{r-1}| < \epsilon$;
8: Let $\boldsymbol{\mu}_k^{new} \leftarrow \boldsymbol{\mu}_k^r$ and $\boldsymbol{\Sigma}_k^{new} \leftarrow \boldsymbol{\Sigma}_k^r$;

3.4 TERMS prediction

To demonstrate emotion distribution for each text on VA-space, this module provides statistical estimations. It represents the outcome of the model for the unseen texts by summarizing the weighted GMMs for each text as well as serves as an evaluation of the performance of the emotion distribution on the unseen texts, as shown in the rightmost part of Fig. 1.

Consider $p(z = k|\boldsymbol{x}_{unseen})$ as the unseen text emotion probability that is calculated as shown in Section 3.2, and $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the estimation of the GMM model derived in Section 3.3; thus, the weighted GMM for unseen text is represented as follows:

$$p(\boldsymbol{y}|\boldsymbol{x}_{unseen}) = \sum_{k=1}^{K} p(z = k|\boldsymbol{x}_{unseen})\mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \quad (16)$$

To summarize the weighted GMM for the unseen text, $p(\boldsymbol{y}|\boldsymbol{x}_{unseen})$, we estimate a single affective Gaussian represented as $N(\boldsymbol{\mu}^{pre}, \boldsymbol{\Sigma}^{pre})$ and thus approximated as follows:

$$\boldsymbol{\mu}^{pre} = \sum_{k=1}^{K} p(z = k|\boldsymbol{x}_{unseen})\boldsymbol{\mu}_k, \quad (17)$$

$$\boldsymbol{\Sigma}^{pre} = \sum_{k=1}^{K} p(z = k|\boldsymbol{x}_{unseen})(\boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k^{*T}\boldsymbol{\mu}_k^{*}), \quad (18)$$

where $\boldsymbol{\mu}_k^{*} = \boldsymbol{\mu}_k - \boldsymbol{\mu}^{pre}$.

The above computations indicate the position and shape of an unseen text in the VA-space. An affective Gaussian on the weighted GMM estimates a single mean and a covariance, thus providing a single distribution as the prediction outcome. This makes the evaluation between the predicted emotion distribution and the ground truth easier to estimate and comprehend in VA-space.

4 Performance evaluation

In this section, we report on the performance evaluation of TERMS that was conducted with large-scale simulations.

4.1 Data collection

For the experimental analysis, we collected data from Twitter, where texts have rich affective content. To collect relevant data, we retrieved sentiment-related hashtags placed at the end of the text, which conveyed the emotion in the text is felt by the locutor as stated in [13]. Based on this method, after some refinement, we gathered 4000 texts from Twitter with labels that were the same as the eight emotions in the wheel of emotion model presented by Plutchik [15]. The eight emotion candidates were *anger, anticipation, disgust, fear, joy, sadness, surprise*, and *trust*. The number of affective text selections was designed to maintain balance among all the classes of sentiments. The statistics of the emotion distributions are shown in Table 2.

Each of the selected texts was rated with VA-values by five different annotators who passed a sample qualification test on Amazon Mechanical Turk (AMT), which is considered a reliable service to obtain high-quality data inexpensively and rapidly. The ratings by five different annotators for each text makes the collection of 20000 rating for the given 4000 texts. We adopted [67] to design an affective slider (AS) in the form of two slider bars to rate valence and arousal independently. The ranges of the valence and arousal were set as $v \in [1, 9]$ and $a \in [1, 9]$. The rating interface is shown in Fig. 2.

4.2 Comparative models

For comparative evaluations, we tested TERMS with baseline models as well as state-of-the-art models. We implemented baseline models that are known to perform well in classification tasks and had been extensively used for emotion recognition. The baseline classifiers used for the comparative analysis are elaborated below.

Baseline Classifiers For baseline models, we implemented four prevalent supervised models to compute emotion probabilities and parameterize distributions. The classifiers employed are multinomial naïve Bayes (NB) [1], support vector machine (SVM) [16], gradient boosting (GBM) [68], and convolution neural network (CNN) [65]. All these approaches directly output the probability of each emotion category for a given text; thus, their outputs were directly used as emotion probabilities.

State-of-the-art models We also compared our model with four benchmark studies. The first was the DeepMoji

emotion prediction model [35]. The second was that of the winning team of the emotion classification subtask in SemEval-2018 Task 1: Affect in Tweets challenge [37, 69]. The third study is a semi-supervised approach for valence and arousal prediction based on variational autoencoder model [70] and the fourth is a context-aware model for emotion classification and sentiment score prediction [71].

DeepMoji It is an established model and has been used as a foundation in many recent studies. It has been trained on billions of tweets and uses the GRNN algorithm for emotion prediction. We used the model[1] available on the GitHub platform and finetuned it with our dataset.

NTUA-SLP_NBOW and NTUA-SLP_LSTM The second comparative study is related to the SemEval-2018 Task-1 challenge, which proposed five subtasks related to intensity (arousal) and valence detection and multi-label emotion classification. The first four subtasks required the identification of arousal and valence scores in tweets in terms of regression values (Subtasks 1 and 3) and ordinal classification (Subtasks 2 and 4), and the fifth subtask was emotion classification, the assignment of multiple labels to the tweets based on the best fit. We compared our TERMS model with the results of the fifth subtask and arousal and valence regression subtasks (Subtasks 1 and 3). The winning team for the fifth subtask was NTUA-SLP [69], which also took second and fourth place in Subtasks 1 and 4, respectively. We obtained the team's pretrained model[2] and implemented it on our data. The team had implemented two approaches: NTUA-SLP_NBOW and NTUA-SLP_LSTM. NTUA-SLP_NBOW used neural bag-of-words model (NBOW) with word2vec and affective word embeddings fed into an SVM classifier. NTUA-SLP_LSTM employed a transfer learning model, which consisted of a two-layer bidirectional long short-term memory (LSTM) with a deep self-attention mechanism. We evaluated the NTUA-SLP model for both the implemented approaches, NTUA-SLP_NBOW and NTUA-SLP_LSTM for comparative evaluation.

SRV-SLSTM It is a semi-supervised regression variational autoencoder (SRV) that identifies VAD scores. The model architecture consist of three modules, encoder, sentiment prediction and decoder. Encoder uses LSTM to encode text into hidden vectors, a sentiment prediction module scores text via a 2-layer stacked Bi-LSTM and decoder reconstructs the original text. We use SRV-SLSTM model

---

[1] https://github.com/bfelbo/deepmoji

[2] https://github.com/cbaziotis/ntua-slp-semeval2018

Table 2 Emotion distribution statistics

Emotions	Anger	Anti.	Disgust	Fear	Joy	Sad.	Surprise	Trust	Total
No. of texts	535	482	481	539	495	511	470	487	4000

publicly available at GitHub platform[3] and employed it on our dataset.

Context-LSTM-CNN (C-LSTM-CNN) The model combines the strength of LSTM and CNN with the lightweight context encoding algorithm Fixed Size Ordinally Forgetting (FOFE) for emotion classification and sentiment score prediction based on contexts and long-range dependencies. The model used for comparative evaluation is available at GitHub platform[4].

4.3 Evaluation measurements

We used the following performance metrics to evaluate the proposed TERMS and comparatives models.

Distinguishability: This shows the average distance among the K emotions: the greater the average distance, the higher the distinguishability of emotions. We denote the average distance between the emotion distributions on VA-space by AEmoD, which is computed as follows:

$$\text{AEmoD} = \frac{1}{N_{pair}} \sum_{i \neq j}^{K} ||\boldsymbol{\mu}_i - \boldsymbol{\mu}_j||, \tag{19}$$

where $N_{pair} = \frac{K(K-1)}{2}$, and $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$ are the means of emotion i and j, respectively.

Prediction Correctness: This shows the correctness of the predicted emotions with respect to the direct observations, which were provided by the annotators. The ratings obtained from the annotators were averaged for each text and used as the ground truth for the comparative evaluation. To quantify the prediction correctness, we used the average Kullback-Leibler (AKL) divergence, average Euclidean distance (AED), and Pearson correlation coefficient (PCC). The AKL divergence [72] measures the distance and similarity between two distributions expressed as an average difference. A smaller AKL indicates the two distributions are similar, hence implying the predicted emotion distribution is close to the ground truth. AKL is a notable measure for evaluation as it takes both the mean and covariance of distributions into account for the correctness test. In addition to AKL divergence, we also calculated the AED, which shows the mean square difference between

the two emotion distributions. A smaller value of AED indicates higher prediction correctness. PCC, denoted as r, was utilized to measure the correlation between the predicted emotion and direct observations. It was used with valence and arousal independently. Differing from the AKL, the PCC is only concerned with the position of emotion distributions on VA-space, by measuring how close the predictions are to the direct observations.

Classification Performance To evaluate the performance of the classifiers employed for soft emotion classification, we use standard evaluation metrics, such as precision, recall, and F1-score computed with macro-averaging. The reason to use macro-averaging for these metrics is the balanced structure of emotion classes in the dataset. Precision (P_e) denotes the fraction of true positives predicted in the processed data, whereas recall (R_e) measures the fraction of true positives predicted from all the positives in the ground truth data [61]. The F1-score is the harmonic mean of the precision and recall. These performance metrics are estimated as follows adapted from [37]:

$$P_e = \frac{\text{No. of texts correctly assigned to emotion class } e}{\text{No. of texts assigned to emotion class } e} \tag{20}$$

$$R_e = \frac{\text{No. of texts correctly assigned to emotion class } e}{\text{No. of texts in emotion class } e} \tag{21}$$

$$F_e = \frac{2 \times P_e \times R_e}{P_e + R_e}, \tag{22}$$

$$F1 - Score = \frac{1}{|E|} \sum_{e \in E} F_e \tag{23}$$

To further validate the classification performance, the Jaccard index is computed as in [37]. The Jaccard index computes the accuracy of the models by dividing the intersection size of the predicted and ground truth labels with the size of their union as shown in (24), where t refers to a text, G_t is the set of ground truths, and P_t is the set of predicted labels.

$$Jaccard = \frac{1}{|T|} \sum_{t \in T} \frac{G_t \cap P_t}{G_t \cup P_t} \tag{24}$$

The described evaluation metrics are considered effective in assessing the efficiency of classifiers and have been

---

[3] https://github.com/wuch15/SRV-DSA

[4] https://github.com/deansong/contextLSTMCNN

Fig. 2 Valence and arousal rating interface. Top: arousal. Bottom: valence

TextID: 1082

I get to spend tomorrow with some of my favorite people

Move the slider to rate Arousal after reading the text above
Arousal represents the intensity of the affective content in a text.

Move the slider to rate Valence after reading the text above
Valence represents the sentiment polarity of the affective content in a text.

used in many pioneering studies [53, 38]. We selected these evaluation metrics as higher scores in all of them represented higher classification performance.

Another evaluation metric that is essential to signify the better classification performance of TERMS model is Bayesian analysis [73]. In Bayesian analysis, the experiment is summarised by the posterior distribution. The posterior describes the distribution of the mean difference of accuracies between the two classifiers. Formally, the interval $[-0.01, 0.01]$ defines a region of practical equivalence (rope) for classifiers [73, 74]. By querying the posterior distribution, we infer the probability that TERMS is better than other comparative models, if the posterior probability of the mean difference are positive, namely the integral of the posterior on the interval $[0.01, \infty]$. Alternatively, if the mean difference is negative (interval $[-\infty, -0.01]$), it states the proposed model is not better, and lastly, if over the rope interval ($[-0.01, 0.01]$) means the posterior probability of the two classifiers are equivalent [73].

4.4 Setup

Since none of the models use a GMM to map the (elliptical) emotion distributions in the VA-space, we utilized all the described baseline models and DeepMoji to map the emotion distributions in the VA-space as had been done with the TERMS model. The baseline classifiers (NB, GBM, and SVM) use the bag-of-words (BoW) model with term frequency features to train the classifiers. The classifiers employed were MultinomialNB, GradientBoostingClassifier, and SVC(linear) respectively from the Python sklearn toolkit. For the parameter setting of the classification models, we used GridsearchCV that exhaustively evaluates all the parameter combinations and retains the best combination to fit the data. For CNN, the

TextCNN algorithm with Adamax optimization is used with word embeddings (128 dimensions) as features, batch size 100, and layers for kernel sizes 2 to 5 were included.

To train the models for emotion probability estimation, we collected another data set with similar textual content. The data set was gathered from Facebook and Twitter, which, after refining, was reduced to 14350 texts. The texts were labeled with eight emotions (as per the wheel of emotion model) by three psychological experts from the field and were also verified by the authors themselves. This data set was merely used for training models in order to compute emotion probabilities for the primary data set (4000 texts). Once the emotion probabilities were estimated, they were infused into a GMM like the proposed model with the same VA-annotations for comparative evaluation. The state-of-the-art models NTUA-SLP, SRV-SLSTM, and C-LSTM-CNN performed the prediction of valence and arousal in their own setting, therefore, we did not infuse it into our model. NTUA-SLP_LSTM used its multilayered design with three main steps: word-embedding pre-training, transfer learning, and fine-tuning. The first two steps of the model were implemented likewise in [69]. For the transfer learning approach, the biLSTM network with deep self-attention mechanism was pre-trained on the Semeval 2017 Task 4A dataset (SA2017). The pre-trained model was combined with the final layer of the model, which was attributed to the subtasks, such as predicting valence and arousal and multi-label classification. We have fine-tuned the final layer of the model for our dataset with respect to each subtask. The same 4000 rated texts were used to fine-tune valence and arousal prediction subtasks. The experimental settings for SRV-SLSTM and C-LSTM-CNN had been kept same as in the original works as the models seemed to perform best on the specified settings. SRV-SLSTM was trained for various ratios of labeled training data; however, it showed best performance on 40% of

labelled data; therefore, we compared our model to those scores. Each experiment was performed 10 times for SRV-SLSTM and the average results were added in the paper. The approach for C-LSTM-CNN model was modified in a similar way to [75] in order to return the dimensional emotion scores.

In addition, we did not assess these models (NTUA-SLP, SRV-SLSTM, and C-LSTM-CNN) for the metrics of distinguishability and prediction with AKL and AED, as the model's architectures were not designed for mapping emotion distributions in VA-space and had their own function for computing the VA-values. This eliminated the need to test it in our setting and enabled us to evaluate our model in the dynamic environment.

To evaluate the prediction performance of the TERMS and comparative models, five-fold cross-validation was carried out on the 4000 rated texts. The data was split into an 80/20 ratio, where for each fold, 80% was used as the training data, and the remaining 20% was used as the testing data. The validation process was completed five times, with each 20% of the set serving once as the testing data, in order to gather the overall results.

4.5 Results

The main take-away messages and simulation results are provided in this section. We first demonstrate results for distinguishability, followed by prediction correctness, and at the end the classification performance of the TERMS and comparative models are elaborated.

4.5.1 Distinguishability

This part compares the distinguishability achieved by TERMS and all the other models as displayed in Fig. 3. Figure 3a illustrates that all the emotion distributions for the proposed TERMS model are well separated and have a better adjustment (i.e., positive emotions on the right and negative on the left in all four quadrants of the VA-space), thus, exhibiting well-discriminated emotion distributions. The deep learning models such as, CNN (Fig. 3b) and DeepMoji (Fig. 3c) show good distinguishability compared to other baseline models, where all the emotion distributions lie correctly on the valence dimension with better clarity. The DeepMoji model blended fairly well in the TERMS
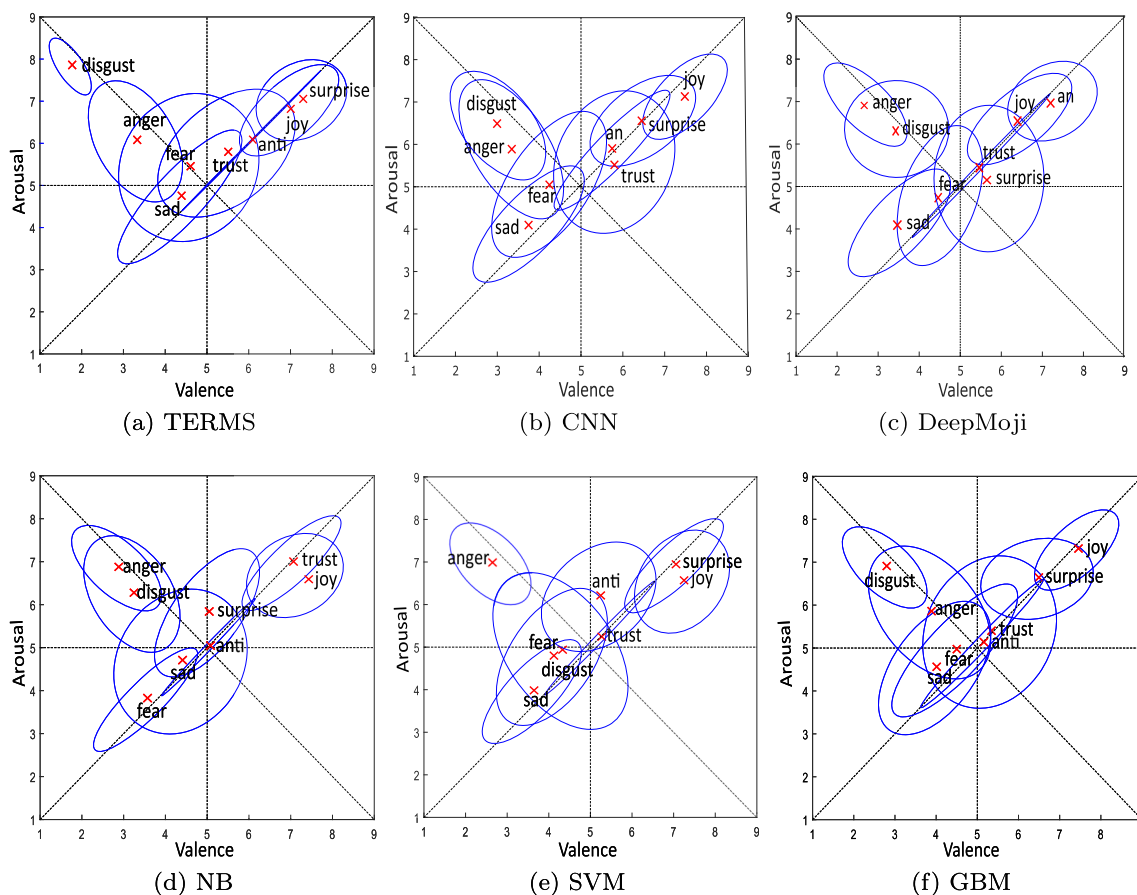

Fig. 3 Distinguishability results

setting with an appropriate allocation of emotion polarities in the VA-space. The baseline models (Fig. 3d–f) also show fair adjustment, however with marginal difference, they fell short of distinct projections of emotion distributions. Upon close inspection, we observe that compared to all the other models, our proposed TERMS model have higher distinguishability.

To quantify distinguishability, we computed the AEmoD for each model via (19). Figure 4 shows the achieved results. A higher value of AEmoD indicates more scatteredness and distinguishable emotion distributions. From Fig. 4, we can see that the deep learning models performed well; however, the TERMS model achieved the highest distinguishability score of 2.642, while the other models scored lower. The graph-based approach of the TERMS emotion classifier provides better coverage by capturing rare words through syntactic relationships and disambiguating emotional meaning using the enriched and refined contextual information of the patterns. The emotion patterns capture fine-grained linguistic affect information, which helps in distinguishing the emotions.

4.5.2 Prediction correctness

We evaluated the prediction performance of TERMS and comparative models by computing the distance between the ground truth and the predicted distributions via AKL and AED. Table 3 lists the AKL, AED, and the correlation coefficient of r for valence and arousal for each model. We found that among all the models, the proposed TERMS model achieved the lowest AKL and AED scores (4.71 and 1.32, respectively) and achieved the highest correlation for valence (0.60) and the third best for arousal (0.30). The results show the predicted distributions for our model
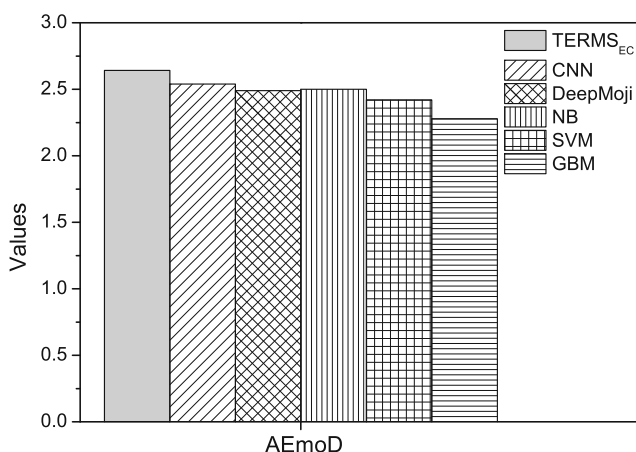
were closest to the actual ratings, thus indicating the better prediction performance of TERMS over the baseline and state-of-the-art models. The integration of a context-aware emotion classifier with the varying emotion perceptions modeled via the GMM distributions provided an edge to TERMS in capturing the nuances of embedded emotions. The architecture of the proposed emotion classifier and the emotion patterns acted as the key components resulting in the higher prediction performance of TERMS, compared to other models. NTUA-SLP_LSTM performed very well with the highest correlation in arousal prediction and the second-best for valence after TERMS. We believe the 2-layer bidirectional LSTM (BiLSTM) with a deep self-attention mechanism captured the salient words in tweets by gathering the information from both directions of text. It provided fair estimation of important words that were highly indicative of certain emotions. NTUA-SLP_NBOW also performed well, which can be attributed to the fact that the pre-trained word2vec embeddings combined with the 10 affective dimensions enabled the model to encode the correlation of each word with different affective dimensions that could result in better intensity performance. SRV-SLSTM and C-LSTM-CNN also showed greater prediction performance compared to the baseline models. The results also indicated that arousal was more challenging to predict compared to valence as the r of arousal was lower than that of valence for all the models.

4.5.3 Classification performance

We evaluated the performance of emotion classification for the proposed TERMS model and all the comparative models. Figure 5 presents the calculated results of precision, recall, F1-score, and Jaccard. We found that the TERMS emotion classifier achieved higher values for precision (0.66), recall (0.65), F1-score (0.64), and Jaccard (0.49). In contrast, the comparative models achieved lower scores than the TERMS model. Thus, TERMS outperformed all the comparative models in classification. This is due to the context-aware emotion patterns that captured the building blocks in text by creating the syntactic patterns of connector words and subject words with clear distinction. This helped to expose the contextual and latent information, which was followed by the enrichment with word embeddings to provide semantic relationships. The enriched emotion patterns offered to capture the minute details of embedded emotions in a text, such as emotional intensity expressed through repeating characters in words like "looove" or similar emotion-relevant verbs like "desire" and "fancy" that were useful for interpreting context. This attribute of gathering the embedded emotional information enabled the emotion classifier to more effectively recognize the emotions relative to other models.

Fig. 4 AEmoD for each model to determine distinguishability; the larger the value, the better the clarity in the emotion distributions on VA-space

Table 3 Overall performance of prediction

Method	AKL	AED	r valence	r arousal
GBM	5.97	1.51	0.34	0.26
SVM	4.88	1.36	0.53	0.25
NB	5.45	1.42	0.52	0.23
CNN	5.07	1.35	0.58	0.24
DeepMoji	4.81	1.35	0.54	0.23
NTUA-SLP_NBOW	NA	NA	0.56	0.39
NTUA-SLP_LSTM	NA	NA	0.59	**0.40**
SRV-SLSTM	NA	NA	0.53	0.26
C-LSTM-CNN	NA	NA	0.56	0.28
TERMS	**4.54**	**1.30**	**0.60**	0.30

The model that performed the closest to TERMS in classification performance was C-LSTM-CNN. C-LSTM-CNN model's architecture combined with FOFE algorithm effectively captured the large context of the focus sentence that helped in better identification of emotions. NTUA-SLP_NBOW, NTUA-SLP_LSTM and CNN also showed satisfactory classification performance. NTUA-SLP_LSTM performed better on its own dataset for all the subtasks provided by SemEval-2018 Task 1. However, in our setting, in contrast, NTUA-SLP_NBOW performed better in terms of classification performance. The deep learning models, CNN and DeepMoji's classification performance was substantially better than the conventional baseline models, which showed a severe setback in performance for this task. Altogether, we observed that TERMS scored higher in classification evaluations followed by the state-of-the-art and deep learning models, and with a large margin to baseline models.

In addition to macro-averaging classification metrics for precision, recall, and F1-score, we evaluated the classification performance with micro-averaging metrics as well. The results are displayed in Fig. 6, which shows that the difference between the macro and micro-averaging scores is trivial, ascertaining the minor impact of averaging methods on balanced structure of emotion classes in the dataset.

To end, we evaluated TERMS model with other comparative models for Bayesian analysis and the results are elaborated in Table 4. The Table shows that the TERMS performs better than the other models as the posterior probability of the mean difference of accuracies are all positive and above 0. All the posteriors are towards the right of the rope i.e. on the interval of $[0.01, \infty]$ shown in last two columns of the Table 4. The test results estimated further strengthened the better performance of the proposed model relative to comparative models.
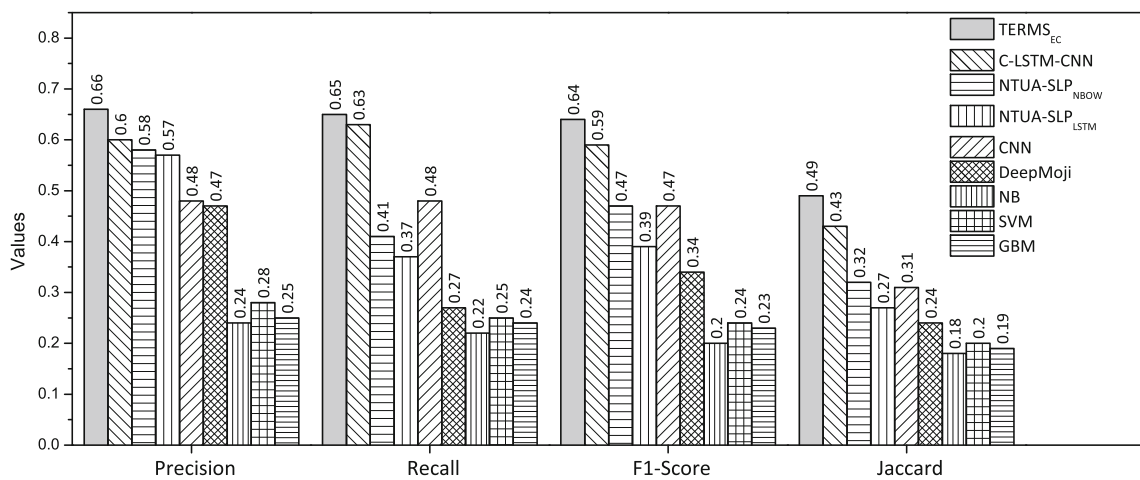

Fig. 5 Classification evaluation metrics for TERMS and all the comparative models. TERMS performs better by demonstrating higher precision, recall, F1-score, and Jaccard
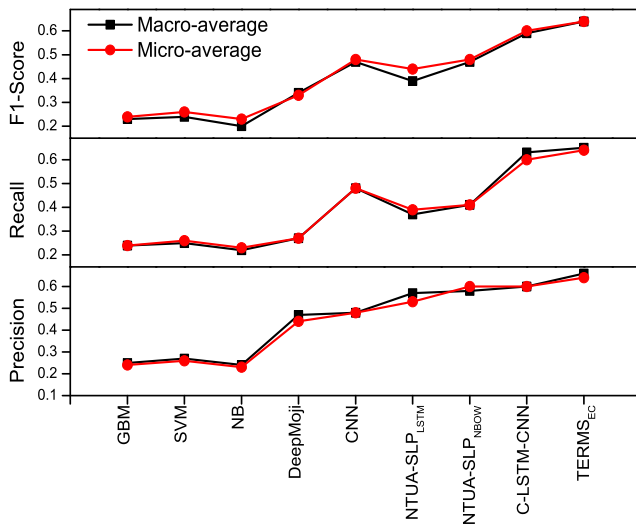
Fig. 6 Classification evaluation metrics with macro and micro-averaging scores

5 Discussion

This section discusses TERMS with different aspects to provide insights on emotional classes, VA-annotations and TERMS in various emotion prediction problems.

Recall by emotion class. The statistical performance of the TERMS emotion model is shown but it would be interesting yet essential to discuss which emotion classes from the dataset were mainly misclassified by the model. In order to do so, we estimated recall by class for the proposed model to determine which emotion class has higher count of false negatives i.e. the class with lowest recall. The results for recall by class are shown in Table 5. From the table, the emotion classes that shows lowest recalls are sadness (0.46) and anger (0.53), which specifies the misclassifications were made in respective classes. The error analysis is provided further to identify the underlying causes of misclassification in these classes.

Both classes with the lowest recall belong to the negative polarity. We believe the texts related to negative emotions

have a high element of sarcasm, satire and irony in them that makes these emotion classes difficult to comprehend. Sarcasm or sardonic statements depend on the prosodic information or non-verbal aspects of communication such as tone, pitch, volume, timbre, facial expressions etc. Lack of these paralinguistic dimensions for anger and sadness can complicate the identification of such emotions from the texts. We provide the sarcastic misclassified texts for recall from our dataset to support our reasoning in Table 6.

Apart from humour and sarcastic comments, the openness of natural language invites ambiguity and misunderstanding. Lack of explicitness in a statement can make it difficult for the emotion detection model to interpret the fuzzy margin between nature of emotions. Table 7 shows the examples from our dataset that were miss calculated due to the lack of explicitness. Lastly, we believe the minor reason that led to low recall was the word sense disambiguation. The miscassified texts based on word ambiguity are stated in Table 8.

VA-annotations. Another aspect of this study that needs an argumentative analysis is valence and arousal annotations. Figure 7 shows TERMS predicted valence and arousal values relative to the ground truth ratings gathered from AMT. From the figure, we can observe that in general the predictions follow the curve of the ratings in the ground truth. For valence, the overall difference between the predicted values and ground truth is smaller as compared to arousal. Predictions for arousal seem more conservative and restricted to individual differences. The arousal dimension normally is widely subjective and shows subtle variations among individuals, which makes this parameter challenging to comprehend.

Furthermore, this study analyzed the impact of a number of annotators on the VA-rating prediction. The model was trained for the reduced number of annotators i.e. 3 and 2 to analyze the influence of annotators' number on prediction performance. Figure 8 shows the PCC curves of valence and arousal for a varied number of annotators. It is explicable that the model performed better for the

Table 4 Bayesian analysis comparative results

Proposed	Others	t-value	p-value	Mean diff.	Lower	Upper
TERMS	C-LSTM-CNN	3.12	0.00	0.32	0.18	0.53
	NTUA-SLP_NBOW	57.10	0.00	2.46	2.41	2.64
	NTUA-SLP_LSTM	30.46	8E-18	0.23	0.21	0.23
	CNN	5.79	8E-09	0.29	0.18	0.38
	DeepMoji	15.87	7E-56	0.78	0.70	0.88
	GBM	7.64	2E-14	0.37	0.28	0.46
	SVM	42.92	0.00	1.88	1.81	1.97
	NB	40.19	0.00	1.76	1.99	2.16

Table 5 Recall by emotion class

| Emotions	Anger	Anti.	Disgust	Fear	Joy	Sad.	Surprise	Trust
Recall	**0.53**	0.64	0.70	0.61	0.69	**0.46**	0.73	0.82

Table 6 Misclassified texts (Sarcasm & satire)

Texts	Actual	Predicted
I love when i can't sleep.	anger	trust
seriously?! we had to turn around because my mom forgot the chicken in the freezer.	anger	sadness
ummmm grow up? please. thank you!	anger	joy
sorry sweetheart you downgraded	anger	joy
lol oh really? is that what its all about?!! hahahaha	anger	sadness

Table 7 Misclassified texts (Lack of explicitness)

Texts	Actual	Predicted
royal mail... why you loose my parcel?	anger	trust
do some girls really think its attractive to look like prostitutes on a daily basis...	anger	disgust
that's fucked up..	anger	joy
i don't even know you anymore.	sadness	trust
guess i'm not good enough for you...	sadness	trust

Table 8 Misclassified texts (Word sense disambiguation)

Texts	Actual	Predicted
this **walking dead** is very disappointing	sadness	fear
there are so many disrespectful and **disgusting** men in this world.	anger	disgust
had to say it....because this generation is going straight down the **drain.**	sadness	disgust
i **hate** it when my comforter **smells** like someone that i miss.	sadness	disgust
i was in such a **good mood**...that's gone out the window!	anger	joy

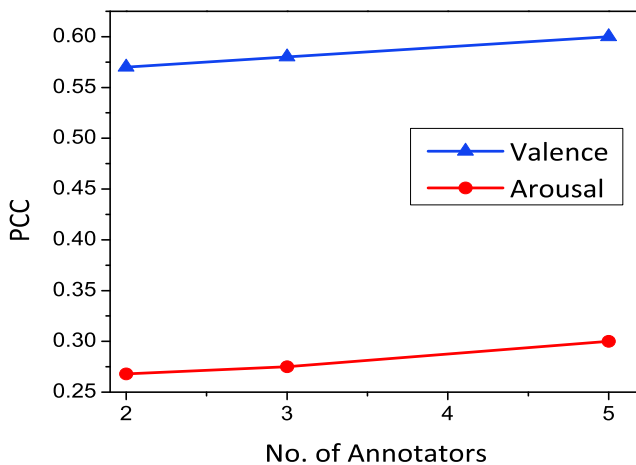**Fig. 7** Predicted values of valence and arousal by TERMS
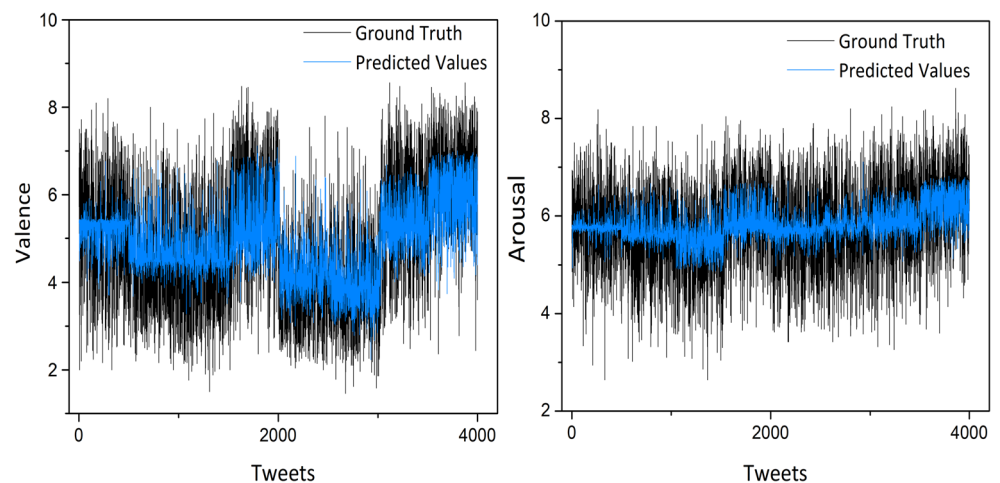
Fig. 8 PCC curves of valence and arousal at varied number of annotators.

highest number of annotators. The difference in prediction performance between the number of annotators is evident that ascertains the increased number of annotators would enhance the quality of the model's performance. However, it is notable that the minor variation in annotators' number has resulted in a significant improvement in models prediction performance. We anticipate that model's prediction quality and capacity to capture individual differences would enhance substantially with a slight increase in the number of annotators for future implementations.

Personalizing VA-annotations For this study, the VA-ratings used for experiments were annotated through AMT. The quality checks for the VA-ratings were maintained during the annotation process; however, it is inferred that the annotation could be influenced by annotators' personality traits or culture differences. To test this inference, a small experiment was conducted with 200 texts from Twitter, annotated again for VA-ratings on AMT; however, before annotation the personality test was conducted by Big Five Inventory (BFI) to get the scores of the personality in each dimension of the Big Five (extraversion, agreeableness, conscientiousness, neuroticism and openness to experience) [76]. The results show the personalities do influence the va-ratings but in different manner. It shows that people who score high in the neuroticism dimension would signify the negative emotions, which can lead to lower VA-ratings for positive emotion. In contrast, annotators high in agreeableness tend to be more exciting and pleasant for positive emotions (higher arousal and valence) and calmer for negative emotions. Extraversion and openness to experience have minimal impact on the VA-ratings for negative emotions and the last personality trait exhibits the annotators high in conscientiousness have higher VA-ratings for positive emotions; however,

for negative emotions, they tend to signify unpleasantness (lower valence). The findings conclude that the personality traits have a moderate influence on VA-rating behaviour. The model limits in covering the personality difference in VA-annotations and the impact it can have on emotion distributions. In future, we would like to integrate an aspect of personality variation and its influence in recognizing emotions on dimensional VA-space.

TERMS in emotion prediction problem TERMS have an absolute significance in the prevailing global crisis, the Coronavirus pandemic (COVID-19). The model is well equipped to identify the nuances of emotions and is applicable for any emotion prediction problem as severe as the COVID-19 crisis [77]. The trauma of COVID-19 has spread uncertainty and extreme emotional distress among people. The variation and uncertainty in emotional states would be essential to identify and understand the emotional needs of people in the crisis. The TERMS context-aware emotion classifier can be effectively used to capture the emotions from the microblog text before COVID-19 and during the pandemic to analyse prevailing emotion dynamics. The resulting emotional classes can be scaled with any variables that are significant to the pandemic (such as population, density, migration, etc.) for any city or location through linear regression to study its impact on emotional states during and before the pandemic. This will provide an overview of the emotional standing or the cognitive narrative of the respective cities, which is essential in this global crisis to provide reassurance and designate the contingency plan as per the emotional needs of city dwellers. Moreover, the model can significantly be employed for any political scenario. The microblog texts related to political debates are high on emotion and sentimental content. The texts contain varying perspectives, fuzzy opinions, and linguistic variations that would require a probabilistic context-aware model and dimensional mapping integrated in a model to capture the nuance, depth and dimensions of emotions embedded in a text.

6 Conclusion

Microblog texts are explicit, relevant, and rich in emotional content; however, their aberrant and informal language makes emotion recognition a challenging task to be employed in real-world systems. It is essential to understand the contextualized information and linguistic variation with a complete coverage of varying emotional perceptions towards the same text in order to recognize emotions from texts accurately. In this article, we propose a probabilistic emotion recognition model TERMS that addresses the

above challenges. In particular, the TERMS model captures the rare and refined contextual emotional information through the proposed emotion classifier. To capture and learn from varying perceptions, TERMS utilizes a GMM to derive the emotion distribution in a VA-space. The emotional information in the probabilistic form is merged with learned GMM parameters from the VA-ratings to generate emotion distributions in VA-space to cover the varying emotional perceptions. We validate the significance of emotion distributions through a detailed comparative analysis with baseline and state-of-the-art models. The results show that TERMS achieved the best performance relative to other models based on the performance metrics of distinguishability, prediction, and classification performance. Furthermore, the proposed model is scalable and adaptable since different classifiers can be implemented to compute emotional probabilities as well as due to the transparent learning process of the GMM. TERMS paves the way for the affective modeling of texts by parameterizing emotion distributions with applications to behavior analysis, forecasting, healthcare, and affective human-computer interaction.

References

1. Perikos I, Hatzilygeroudis I (2018) A framework for analyzing big social data and modelling emotions in social media. In: IEEE Proceedings of BigDataService, pp 80–84
2. Basile P, Basile V, Nissim M, Novielli N, Patti V et al (2018) Sentiment analysis of microblogging data
3. Bermingham A, Smeaton A (2010) Classifying sentiment in microblogs: Is brevity an advantage? In: ACM Proceedings of CIKM, pp 1833–1836
4. Rintyarna BS, Sarno R, Fatichah C (2020) Enhancing the performance of sentiment analysis task on product reviews by handling both local and global context. Int J Inform Decis Sci 12(1):75–101
5. Dini L, Bittar A, Robin C, Segond F, Montaner M (2017) Soma: The smart social customer relationship management tool: Handling semantic variability of emotion analysis with hybrid technologies. In: Sentiment Analysis in Social Networks, pp 197–209
6. Ghanem B, Buscaldi D, Rosso P (2019) Textrolls: Identifying russian trolls on twitter from a textual perspective. arXiv:1910.01340
7. Abdullah M, Hadzikadic M (2017) Sentiment analysis of twitter data: Emotions revealed regarding Donald Trump during the 2015-16 primary debates. In: IEEE Proceedings of ICTAI, pp 760–764
8. Calvo RA, Milne DN, Hussain MS, Christensen H (2017) Natural language processing in mental health applications using non-clinical texts. Nat Lang Eng 23(5):649–685
9. Carrillo-de Albornoz J, Rodríguez Vidal J, Plaza L (2018) Feature engineering for sentiment analysis in e-health forums. PLoS One 13(11):e0207996
10. Torres EP, Torres EA, Hernández-Álvarez M, Yoo SG (2020) Emotion recognition related to stock trading using machine learning algorithms with feature selection. IEEE Access 8:199719–199732
11. Calvo RA, Mac Kim S (2013) Emotions in text: dimensional and categorical models. Comput Intell 29(3):527–543
12. Meo R, Sulis E (2017) Processing affect in social media: A comparison of methods to distinguish emotions in tweets. ACM T Internet Techn 17(1):1–25
13. Abdul-Mageed M, Ungar L (2017) Emonet: Fine-grained emotion detection with gated recurrent neural networks. In: Proceedings of ACL, pp 718–728
14. Ekman P, Sorenson ER, Friesen WV (1969) Pan-cultural elements in facial displays of emotion. Science 164(3875):86–88
15. Plutchik R (2001) The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. AmSci 89(4):344–350
16. Paltoglou G, Thelwall M (2013) Seeing stars of valence and arousal in blog posts. IEEE Trans Affect Comput 4(1):116–123
17. Buechel S, Hahn U (2017) Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In: Proceedings of EACL (Short Papers), pp 578–585
18. Preotiuc-Pietro D, Schwartz HA, Park G, Eichstaedt JC, Kern M, Ungar L, Shulman EP (2016) Modelling valence and arousal in facebook posts. In: ACL Proceedings of WASSA, pp 9–15
19. Mohammad SM (2017) Challenges in sentiment analysis. In: A practical guide to sentiment analysis. Springer, pp 61–83
20. Mulcrone K (2012) Detecting emotion in text. UMM CSci Senior Seminar
21. Liu B (2010) Sentiment analysis and subjectivity. Handb Nat Lang Process 2(2010):627–666
22. Wang J-C, Yang Y-H, Wang H-M, Jeng S-K (2015) Modeling the affective content of music with a gaussian mixture model. IEEE Trans Affect Comput 6(1):56–68
23. Vinayagasundaram B, Mallik R, Aravind M, Aarthi RJ, Senthilrhaj S (2016) Building a generative model for affective content of music. In: IEEE Proceedings of ICRTIT, pp 1–6
24. Pribil J, Pribilova A, Matousek J (2019) Artefact determination by GMM-based continuous detection of emotional changes in synthetic speech. In: IEEE Proceedings of TSP, pp 45–48
25. Giachanou A, Crestani F (2016) Like it or not: A survey of twitter sentiment analysis methods. ACM Comput Surv 49(2):1–41
26. Seyeditabari A, Tabari N, Zadrozny W (2018) Emotion detection in text: A review. arXiv:1806.00674v1
27. Suttles J, Ide N (2013) Distant supervision for emotion classification with discrete binary values. In: Proceedings of CICLing, pp 121–136
28. Perikos I, Hatzilygeroudis I (2016) Recognizing emotions in text using ensemble of classifiers. Eng Appl Artif Intel 51:191–201
29. Symeonidis S, Effrosynidis D, Arampatzis A (2018) A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. Expert Syst Appl 110:298–310
30. Demszky D, Movshovitz-Attias D, Ko J, Cowen A, Nemade G, Ravi S (2020) Goemotions: A dataset of fine-grained emotions. arXiv:2005.00547
31. Lykousas N, Patsakis C, Kaltenbrunner A, Gómez V (2019) Sharing emotions at scale: The vent dataset. In: Proceedings of the International AAAI Conference on Web and Social Media, vol 13, pp 611–619
32. Alvarez-Gonzalez N, Kaltenbrunner A, Gómez V (2021) Uncovering the limits of text-based emotion detection. arXiv:2109.01900
33. Malko A, Paris C, Duenser A, Kangas M, Mollá D, Sparks R, Wan S (2021) Demonstrating the reliability of self-annotated emotion data. In: Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access, pp 45–54

34. Peng S, Cao L, Zhou Y, Ouyang Z, Yang A, Li X, Jia W, Yu S (2021) A survey on deep learning for textual emotion analysis in social networks. Digital Communications and Networks

35. Felbo B, Mislove A, Søgaard A, Rahwan I, Lehmann S (2017) Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. arXiv:1708.00524

36. Rosenthal S, Farra N, Nakov P (2017) Semeval-2017 Task 4: Sentiment analysis in twitter. In: ACL Proceedings of SemEval-2017, pp 502–518

37. Mohammad S, Bravo-Marquez F, Salameh M, Kiritchenko S (2018) Semeval-2018 Task 1: Affect in tweets. In: ACL Proceedings SemEval, pp 1–17

38. Zhang S, Xu X, Pang Y, Han J (2020) Multi-layer attention based cnn for target-dependent sentiment classification. Neural Process Lett 51(3):2089–2103

39. Sadr H, Pedram MM, Teshnehlab M (2020) Multi-view deep network: A deep model based on learning features from heterogeneous neural networks for sentiment analysis. IEEE Access 8:86984–86997

40. Mohammad SM (2021) Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text. In: Emotion Measurement. Elsevier, pp 323–379

41. Russell JA (1980) A circumplex model of affect. J Pers Soc Psychol 39(6):1161–1178

42. Mohammad SM (2016) Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In: Emotion measurement. Elsevier, pp 201–237

43. Warriner AB, Kuperman V, Brysbaert M (2013) Norms of valence, arousal, and dominance for 13 915 English lemmas. Behav Res Methods 45(4):1191–1207

44. Mohammad SM (2018) Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In: Proceedings of ACL, pp 174–184

45. Hasan M, Rundensteiner E, Agu E (2018) Automatic emotion detection in text streams by analyzing twitter data. Int J Data Sci Anal 7(1):35–51

46. Hasan M, Rundensteiner E, Agu E (2014) Emotex: Detecting emotions in twitter messages. In: Proceedings of ASE, pp 1–10

47. Mohammad SM, Bravo-Marquez F (2017) WASSA-2017 shared task on emotion intensity. arXiv:1708.03700

48. Buechel S, Hahn U (2016) Emotion analysis as a regression problem-dimensional models and their implications on emotion representation and metrical evaluation. In: ACM Proceedings of ECAI, pp 1114–1122

49. Park S, Kim J, Ye S, Jeon J, Park HY, Oh A (2021) Dimensional emotion detection from categorical emotion. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp 4367–4380

50. Rawat T, Jain S (2021) A dimensional representation of depressive text. In: Data Analytics and Management. Springer, pp 175–187

51. Cheng Y-Y, Chen Y-M, Yeh W-C, Chang Y-C (2021) Valence and arousal-infused bi-directional lstm for sentiment analysis of government social media management. Appl Sci 11(2):880

52. Li M (2022) Application of sentence-level text analysis: The role of emotion in an experimental learning intervention. J Exp Soc Psychol 99:104278

53. Mohammad SM, Bravo-Marquez F (2017) Emotion intensities in tweets. arXiv:1708.03696

54. Duppada V, Jain R, Hiray S (2018) Seernet at semeval-2018 Task 1: Domain adaptation for affect in tweets. arXiv:1804.06137

55. Zhao S, Jia G, Yang J, Ding G, Keutzer K (2021) Emotion recognition from multiple modalities: Fundamentals and methodologies. IEEE Signal Proc Mag 38(6):59–73

56. Yang Y-H, Chen HH (2011) Prediction of the distribution of perceived music emotions using discrete samples. IEEE T Audio Spe 19(7):2184–2196

57. Zhao S, Yao H, Jiang X (2015) Predicting continuous probability distribution of image emotions in valence-arousal space. In: ACM Proceedings of MM, pp 879–882

58. Sun K, Yu J, Huang Y, Hu X (2009) An improved valence-arousal emotion space for video affective content representation and recognition. In: IEEE Proceedings of ICME, pp 566–569

59. Yang Y-H, Liu J-Y (2013) Quantitative study of music listening behavior in a social and affective context. IEEE T Multimed 15(6):1304–1315

60. Huang Z, Epps J (2016) Detecting the instant of emotion change from speech using a martingale framework. In: IEEE Proceedings of ICASSP, pp 5195–5199

61. Trabelsi I, Ayed DB, Ellouze N (2018) Evaluation of influence of arousal-valence primitives on speech emotion recognition. Int Arab J Inf Technol 15(4):756–762

62. Bishop CM (2006) Pattern recognition and machine learning. Springer, New York

63. Mardia KV (1970) Measures of multivariate skewness and kurtosis with applications. Biometrika 57(3):519–530

64. Saravia E, Argueta C, Chen Y-S (2016) Unsupervised graph-based pattern extraction for multilingual emotion classification. Soc Netw Anal Min 6(1):1–21

65. Saravia E, Liu H-CT, Huang Y-H, Wu J, Chen Y-S (2018) Carer: Contextualized affect representations for emotion recognition. In: ACL Proceedings of EMNLP, pp 3687–3697

66. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. J R Stat Soc: Ser B (Methodol) 39(1):1–22

67. Betella A, Verschure PFMJ (2016) The affective slider: A digital self-assessment scale for the measurement of human emotions. PLoS One 11(2):e0148037

68. Tavares G, Mastelini S et al (2017) User classification on online social networks by post frequency. In: Anais Principais do XIII Simpósio Brasileiro de Sistemas de Informação, pp 464–471

69. Baziotis C, Athanasiou N, Chronopoulou A, Kolovou A, Paraskevopoulos G, Ellinas N, Narayanan S, Potamianos A (2018) Ntua-slp at semeval-2018 Task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning. arXiv:1804.06658

70. Wu C, Wu F, Wu S, Yuan Z, Liu J, Huang Y (2019) Semi-supervised dimensional sentiment analysis with variational autoencoder. Knowl-Based Syst 165:30–39

71. Song X, Petrak J, Roberts A (2018) A deep neural network sentence level classification method with context information. arXiv:1809.00934

72. Hershey JR, Olsen PA (2007) Approximating the Kullback Leibler divergence between Gaussian mixture models. In: IEEE Proceedings of ICASSP, pp IV–320

73. Benavoli A, Corani G, Demšar J, Zaffalon M (2017) Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. J Mach Learn Res 18(1):2653–2688

74. Kruschke JK (2015) Tutorial: Bayesian data analysis. In: CogSci

75. Zhu S, Li S, Zhou G (2019) Adversarial attention modeling for multi-dimensional emotion regression. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp 471–480

76. Guan J (2017) Proving personality-related differences in valence and arousal annotations in social media tasks, National Tsing Hua University, Hsinchu City

77. Ghafoor Y, Calderon FH, Chen LS-W, Chen Y-S (2021) Emotion interaction in cities. In: 2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI). IEEE, pp 91–98

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Yusra Ghafoor is a Ph.D. candidate at Institute of Information Science, Academia Sinica and Institute of Information Systems and Applications, National Tsing Hua University (NTHU). She is a Taiwan International Graduate Program (TIGP) student and studies in program Social Networks and Human Centered Computing (SNHCC). She received her M.S. in Electrical Engineering and Computer Science from National Taipei University of Technology (NTUT), Taiwan. She achieved Excellent student award and Best thesis award in her Master studies. Her research interests include data analysis, artificial intelligence, machine learning and natural language processing.

Fernando H. Calderon is a Honduran, born in 1987. He is a PhD student and Teacher Assistant with the Social Networks and Human Centered Computing program at Academia Sinica and the Institute of Information Systems and Applications, National Tsing Hua University since 2015. His current research interests include behavioral and sentiment analysis from social media, emotion recognition and computational mental health.

Yen-Hao Huang was born in Taiwan, in 1993. He is a PhD student and Teacher Assistant with the Institute of Information Systems and Applications, National Tsing Hua University since 2015. His current research interests include sentiment analysis, emotion recognition, computational mental health, natural language processing and deep learning.

Shi Jinping received the B.E degree in Science and Technology of Remote Sensing from China University of Geosciences, Wuhan, China, in 2010, and the M.S. degree from National Tsing Hua University, Hsinchu, Taiwan, in 2017. From 2010 to 2014, he was a Software Engineer/Product Consultant for a GIS company in China. He is currently an Algorithm Engineer in an e-commerce company, Shenzhen, China, focusing on natural language processing and recommendation systems.

Kuan-Ta Chen (a.k.a. Sheng-Wei Chen) was a Research Fellow at the Institute of Information Science and the Research Center for Information Technology Innovation (joint appointment) of Academia Sinica. Dr. Chen received his Ph.D. in Electrical Engineering from National Taiwan University in 2006, and received his B.S. and M.S. in Computer Science from National Tsing-Hua University in 1998 and 2000, respectively. He received the Young Scholar's Creativity Award from Foundation for the Advancement of Outstanding Scholarship in 2013, and IEEE ComSoc MMTC Best Journal Paper Award in 2014. He was an Associate Editor of IEEE Transactions on Multimedia (IEEE TMM) and an Associate Editor of ACM Transactions on Multimedia Computing, Communications, and Applications (ACM TOMM).

Yi-Shin Chen joined National Tsing Hua University (NTHU) at 2004. Dr. Chen received her Ph.D. in Computer Science from University of Southern California in 2002, and received her B.B.A. and M.B.A. in information management from National Central University in 1996 and 1997, respectively. Currently, she is the principal investigator of the Artificial Intelligence Talent Cultivation Project for AI Techniques and Application Courses (funded by Minister of Education of Taiwan), and the standing director of Taiwanese Association for Artificial Intelligence. She is passionate about increasing society's benefits through her research efforts. For avoiding a media monopoly, she focused her research efforts on Web intelligence and integration. Currently, she has applied natural language processing techniques in understanding the characteristics of music therapy, emotion recognitions, and mental illness.

Affiliations

Yusra Ghafoor[1] (ID) · Shi Jinping[2] · Fernando H. Calderon[1] · Yen-Hao Huang[2] · Kuan-Ta Chen[3] · Yi-Shin Chen[4]

Yusra Ghafoor
yusraghafoor15@gmail.com

Shi Jinping
jping.sh@gmail.com

Fernando H. Calderon
fhcalderon87@gmail.com

Yen-Hao Huang
yenhao0218@gmail.com

Kuan-Ta Chen
swc@iis.sinica.edu.tw

[1] Social Networks and Human-Centered Computing, Taiwan International Graduate Program, Institute of Information Science, Academia Sinica. Institute of Information Systems and Applications, National Tsing Hua University, Hsinchu City, Taiwan

[2] Institute of Information Systems and Applications, National Tsing Hua University, Hsinchu City, Taiwan

[3] Institute of Information Science, Academia Sinica, Taipei, Taiwan

[4] Department of Computer Science, National Tsing Hua University, Hsinchu City, Taiwan