This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3053275, IEEE Transactions on Affective Computing

A Survey of Textual Emotion Recognition and Its Challenges

Jiawen Deng, Fuji Ren

Abstract—Textual language is the most natural carrier of human emotion. In natural language processing, textual emotion recognition (TER) has become an important topic due to its significant academic and commercial potential. With the advanced development of deep learning technologies, TER has attracted growing attention and has significantly been promoted in recent years. This paper provides a systematic survey of latest TER advances, focusing on approaches using deep neural networks. According to how deep learning works at each stage, TER approaches are reviewed on word embedding, architecture, and training levels, respectively. We discussed the remaining challenges and opportunities from four aspects: the shortage of large-scale and high-quality dataset, fuzzy emotional boundaries, incomplete extractable emotional information in texts, and TER in dialogue. This paper creates a systematic and in-depth overview of deep TER technologies. It provides the necessary knowledge and new insights for relevant researchers to understand better the research state, remaining challenges, and future directions in this field.

Index Terms—Textual emotion recognition, Deep learning, Emotional resources, Challenges, Review

1 Introduction

Emotional interaction is a common psychological phenomenon in human's daily life. Accurate emotion recognition is the premise of effective human communication, interaction, and decision making. With the development of big data and artificial intelligence, emotion recognition has become a typical research project in both academia and industry.

As the most basic and direct carrier, textual data generated in emotional communication is often utilized to infer the emotional state of human beings [1], [2], [3]. Moreover, in recent decades, with the rapid development of social media and smart terminals, online social networks have become an unprecedented global phenomenon, such as Facebook, Twitter, Weibo, and Line [4], [5]. People are accustomed to communicating with each other and expressing their thoughts on these platforms. As a result, a large amount of public and personal data is generated. These resources contain rich emotional information and provide a data foundation for emotion-related research [6]. Besides, people pay more attention to understanding themselves better and finding a more effective way to learn, work, and live. This demand has inspired researchers to explore the field of mental health and emotional management [7], [8]. These facts have accelerated the need for fine-grained emotion recognition.

Textual emotion recognition (TER) devote to automatically identifying emotion states in textual expressions, such as Happy, Sad, and Angry. TER is a more in-depth analysis than sentiment analysis[9], [10], and has gained considerable interest from the research community. For humans, emotions can be recognized quickly based on their subjective feelings. Simultaneously, for automatic TER systems, calculation methodologies need to be continuously developed and optimized to achieve more accurate emotion prediction.

In the field of Natural Language Processing (NLP), the majority of reviews about TER are either conducted at coarse-grained sentiment level [11], [12], [13] or focus on the traditional machine learning approaches [14]. Such as the previous work in [15], although they survey the state-of-the-art approaches of TER, deep learning-based TER (deep TER) approaches are only briefly mentioned without an in-depth and comprehensive overview. In recent years, deep learning technologies have overcome the dependence on manual feature extraction and achieved satisfactory performance. Nevertheless, regarding the recent reviews on deep emotion recognition techniques, most studies focus on audio modality [16], visual modality [17], and multimodal [18], [19]. In most cases, they mentioned textual emotion roughly as part of multi-modality, such as the reviews in [20], [21]. Very recently, the works in [22] reviewed the status of emotion recognition in conversation. They introduced the influence of conversation-specific factors during emotion recognition, such as contextual cues and speaker-specific information. However, they are confined to conversation scenarios. They do not outline the advanced TER technology from the perspective of NLP, and the research state of sentence-level TER is not involved as well. These research contents are indispensable and will be reviewed in this paper.

As mentioned above, the existing reviews about deep TER technology only give brief introductions or limited to specific scenarios. In comparison to these reviews, this paper provides a thorough survey of deep TER technology for affective understanding, especially the emerging NLP technology. We aim to create a systematic and in-depth overview for newcomers by discussing the research

- *Jiawen Deng is with the Institute of Technology and Science, Tokushima University, Tokushima, Japan. E-mail: c501847002@tokushima-u.ac.jp.*
- *Fuji Ren is with the Institute of Technology and Science, Tokushima University, Tokushima, Japan. E-mail: ren@is.tokushima-u.ac.jp.*

xxxx-xxxx/0x/$xx.00 © 200x IEEE Published by the IEEE Computer Society

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Robert Gordon University. Downloaded on May 28,2021 at 02:10:33 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3053275, IEEE Transactions on Affective Computing

state, remaining challenges, and future directions of deep TER.

To systematically review the current status, challenges, and future directions of deep TER approaches, we conducted 1) a comprehensive literature search and 2) a further targeted literature selection. It should be noted that our literature search is conducted on the database of Web of Science and Google Scholar. In the process of targeted literature selection for this review, we paid particular attention to those papers published in some top conferences in the field of Natural Language Processing (NLP) and affective computing, including ACL, EMNLP, NAACL, EACL, and CONLING.

1) Comprehensive literature search. During the first round of comprehensive search, we mainly used the keywords: ‘textual emotion’, ‘emotion recognition’, and ‘deep learning’ to search the literature related to deep learning-based TER approaches between 2015 and 2020. To avoid omissions, we also searched some related terms of ‘affective computing’, ‘emotion classification’, ‘emotion detection’, ‘emotion representation’, and ‘emotion embedding’. To reduce overlap with existing reviews, we exclude most literature about non-deep learning and other studies that do not address TER problems.

2) Targeted literature selection. Based on the searched literature, we can systematically understand the research status and development trend of the deep TER technology. Combining with our previous work related to emotion recognition, we think that we can review TER from three aspects and discuss existing challenges from four perspectives, as shown in Figure 1. Therefore, based on the results of the literature search, we conducted the second round of targeted literature selection to include or exclude papers for this review more accurately. We first collected some keywords from various sub-topics of deep TER technology based on the branches shown in Figure 1. This way, we obtained the keywords set: {Language model, Prior knowledge, Emotion knowledge, Transfer learning, Multi-task learning, Joint learning, Cross-language, Emotion annotation, Data imbalance, Low-resource, Multi-label, Label correlation, Emotion intensity, Multi-modal, Dialogue, Conversation, Context Modeling, Dynamic Emotion, Party interaction}. Based on these keywords, we selected targeted literature by quality assessment, selection, and classification. Besides, we also browsed the reference lists of the selected papers, which also yielded some additional literature.

In this paper, we review the research status of deep TER technologies, and our contributions can be summarized as follows:

1) Providing a systematic review of current status, open issues, and research directions of deep TER technologies.

2) We outline the most relevant emotional resources by quantifying the usage in searched literature, including publicly available emotional lexicon and corpora.

3) We review deep TER technologies by focusing on three keys: at the word embedding level, how to obtain word embedding with syntactic contextual information and latent emotional information; at the architecture level, how to design a capable deep learning architecture and incorporate prior knowledge; and at the training level, how to train a TER model based on limited emotional resources efficiently.

4) We conduct further investigations and discussion for the existing challenges and potential opportunities involved in TER from four aspects: The shortage of large-scale and high-quality data; Fuzzy emotional boundaries; Incomplete emotional information in textual expression; and TER in dialogue.

The rest of this paper is organized as follows: Section 2 introduces the background of TER technologies. Section 3 reviews the existing TER approaches based on deep learning technology. Existing challenges are discussed in section 4. Finally, Section 5 concludes this paper.

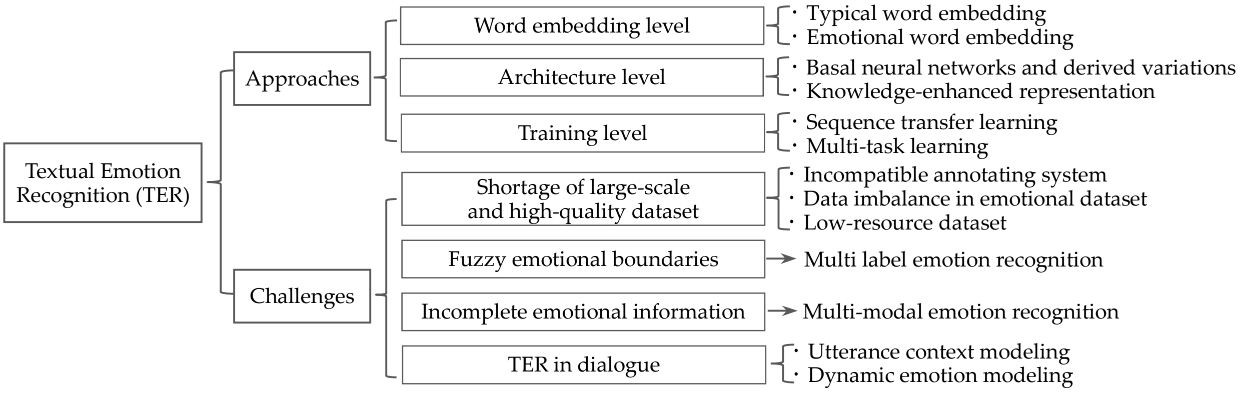
Textual Emotion Recognition (TER)
Approaches
Word embedding level
· Typical word embedding
· Emotional word embedding
Architecture level
· Basal neural networks and derived variations
· Knowledge-enhanced representation
Training level
· Sequence transfer learning
· Multi-task learning
Challenges
Shortage of large-scale and high-quality dataset
· Incompatible annotating system
· Data imbalance in emotional dataset
· Low-resource dataset
Fuzzy emotional boundaries
Multi label emotion recognition
Incomplete emotional information
Multi-modal emotion recognition
TER in dialogue
· Utterance context modeling
· Dynamic emotion modeling

Fig.1 Framework of textual emotion recognition reviewed in this paper

2 Overview

2.1 Textual Emotion Recognition

As an essential element in human nature, emotion analysis has raised increasing attention over the past two decades [23]. The terms ’emotion recognition’ and ‘sentiment analysis’ are often used interchangeably [9] while there are apparent differences between these two concepts [10]. Sentiment analysis mainly measures subjective attitudes

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Robert Gordon University. Downloaded on May 28,2021 at 02:10:33 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3053275, IEEE Transactions on Affective Computing

from the perspective of sentiment polarity. Emotion recognition involves identifying more detailed emotional states. It refers to a wide range of mental states, such as happiness, anger, and fear. Textual emotion recognition (TER) aims to classify a textual expression into one or several emotion categories depending on the underlying emotion theories.

The nature of emotion indicates the potential applications of TER in a variety of fields. In emotional management, people can monitor their work status and mental state by tracking the emotional rhythm in different periods [7], [8]. In marketing communications, analyzing consumer preferences helps to improve business strategies [24]. In Social networks, emotion recognition in a sinister tone helps detect potential criminals or terrorists [4], [5]. By analyzing the data released in social networks, real-time emotion monitoring can be realized, contributing to effective suicide prevention [25]. Emotion detection during a crisis or disaster helps understand peoples' feelings towards a particular situation, contributing to crisis management and critical decision-making [26]. During elections, public emotions can be tracked and predicted based on their speeches and comments online [27]. In human-computer interaction systems, such as dialog systems, and companion robots, the text is a commonly used modality. TER is the most convenient way to understand the emotional dialogue [28], making human-computer interaction more accurate and intelligent [29]. The resulting market demand has extensively promoted TER in Artificial Intelligence (AI) and NLP.

2.2 Emotional Resources

The deep TER system is data-driven and relies on a large amount of data. Standard, free, and generalized databases are the guarantee of model performance. This section discusses the publicly available databases that contain emotion knowledge and are commonly accepted in TER tasks. Existing emotional databases are mainly annotated based on the discrete or dimensional emotion model. Therefore, we introduce emotional resources with categorical annotation and dimensional annotation, respectively. Table 1 and Table 2 provide an overview of some publicly available emotional lexicons and corpora.

2.2.1 Datasets with Categorical Annotation

Discrete emotion models have been commonly accepted in emotion recognition because of their simplicity and intuitiveness. Emotions are classified into several discrete categories. Thus, the emotion recognition task can be interpreted as a classification task that assigns one or several emotion labels to a piece of expression. Until now, there is no consensus on the definition of basic emotion categories. The commonly accepted universal emotions are Paul Ekman's six basic emotions: Anger, Disgust, Fear, Happiness, Sadness, and Surprise [30]. Afterward, considering the influence of emotion correlation, Plutchik proposed a wheel of emotions [31] to describe the relations of basic emotions. The wheel organized eight basic emotions into four bipolar axes: Joy-Sadness, Fear-Anger, Trust-Disgust, and Surprise-Anticipation. He claims other complex emotions can be viewed as combinations of the relevant primary ones with different emotional intensity. This emotion wheel enables researchers to conduct emotion recognition more comprehensively.

Most prominent and available public corpora are annotated based on Ekman and Plutchik's basic emotions or extensions. SemEval-2007 Affective Text Task corpus[1] contains 1250 news headlines annotated with Ekman's basic emotions. These data are collected from news websites and newspapers [32]. ISEAR[2] contains 7600 self-reported experiments of emotion-provoking text, which are generated by their reactions to seven primary emotions. NLPCC-2018 database[3] contains 7928 code-switching texts with five emotion labels, and each text contains more than one language (Chinese and English) [33]. It was the benchmarking data for NLPCC Shared Task of Emotion Detection. Alm's fairy tale dataset consists of 1580 sentences collected from 185 children fairy tales. Their annotation is an extension of Ekman's basic emotions, which further differentiates the polarity of 'Surprise' and supplements the 'Neutral' label [34].

Besides the above-mentioned sentence-level corpus, there are many dialogue corpora categorically annotated. EmotionContext[4] is a collection of three-turns dialogue, and each utterance is annotated with emotion labels [1]. The training data set contain 15k records for emotion classes and 15k records of 'Others'. It was the benchmarking data for SemEval-2019 task 3. DailyDialog[5] contains 13, 118 multi-turn dialogues and 897 scenes about daily life [2]. This dataset is manually annotated with both communication intention and emotion labels. EmoryNLP[6] collects multiparty dialogues [3] from TV show transcripts: 'Friends'. This dataset comprises 97 episodes, 897 scenes, and 12, 606 utterances. EmotionLines [35] contains a total of 29,245 utterances from 2,000 dialogues. They are collected from Friends TV scripts and private Facebook messenger dialogues. The emotions of all utterances are labeled based on the textual content. Multimodal EmotionLines Dataset (MELD)[7] is an expansion of EmotionLines [36]. It contains the same dialogue instances available in EmotionLines, but it also encompasses audio and visual modality along with text.

As the basic emotional knowledge, emotion lexicons are mostly annotated categorically. WordNet-Affect [37] and SentiSense Affective Lexicon [38] are concept-based affective lexicons, providing affective concepts correlated with affective words. NRC Emotion Lexicon [39] (also called EmoLex) is a word-emotion association lexicon available in 40 languages. It is annotated with eight basic emotions and two sentiments. LIWC2015 [40] (Linguistic Enquiry and Word Count, LIWC) is hierarchically annotated with both sentiments and emotion labels. It is provided in English and has been translated into several oth-

[1] Access: *http://web.eecs.umich.edu/~mihalcea/affectivetext/#datasets*
[2] Access: *https://www.unige.ch/cisa/research/materials-and-online-research/research-material/*
[3] Access: *http://tcci.ccf.org.cn/conference/2018/taskdata.php*
[4] Access: *https://www.humanizing-ai.com/emocontext.html*
[5] Access: *http://yanran.li/dailydialog*
[6] Access: *http://doraemon.iis.sinica.edu.tw/emotionlines/index.html*
[7] Access: *https://affective-meld.github.io/*

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Robert Gordon University. Downloaded on May 28,2021 at 02:10:33 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3053275, IEEE Transactions on Affective Computing

TABLE 1
PUBLICLY AVAILABLE DATABASES FOR TEXTUAL EMOTION RECOGNITION

Emotional corpora	Ekman's 6 basic emotions						Other Emotions	Number	Language
	A	D	F	J	Sad	Sur			
SemEval-2007	√	√	√	√	√	√	--	1250	English
ISEAR	√	√	√	√	√	--	Shame, Guilt	7600	English
NLPCC-2018	√	--	√	--	√	√	Happiness	7928	Chinese English
EmoInt Dataset	√	--	√	√	√	--	Sentiment with intensity	7100	English
Ren-CECps	√	--	--	√	--	√	Anxiety, Hate, Love, Sorrow, Expect, Sentiment with intensity	1487 blogs	Chinese
Alm's fairy tale	√	√	√	--	√	√	Happiness, Neural	1580	English
EEC	√	√	√	√	√	√	Love, Optimism, Pessimism, Trust, Anticipation, Sentiment with intensity	8640	English, Arabic, Spanish
EmotionContext	√	--	--	--	√	--	Happiness, Others	30k records	English
DailyDialog	√	√	√	√	√	√	Neural	13k dialogues	English
MELD	√	√	√	√	√	√	Neural, Positive, Negative,	13k utterances	English
EmoryNLP	--	--	--	√	√	--	Neutral, Mad, Scared, Powerful, Peaceful	12.6k utterances	English
EmotionLines	√	√	√	--	√	√	Neural, Happiness	2.9k utterances	English
SEMAINE Database	--	--	--	--	--	--	Dimensional	240 dialogues	English
IEMOCAP	√	--	√		√	√	Dimensional, Happiness, Excitement, Frustration, Other and Neutral	5.5 K utterances	English
EMOBANK	√	√	√	√	√	√	Dimensional	10k	English

Ekman's 6 basic emotions: A-Anger, D-Disgust, F-Fear, J-Joy, Sad-Sadness, Sur-Surprise

er languages, including Arabic, Korean, Turkish, and Chinese.

Some more detailed corpora are annotated with both categorical emotional labels and corresponding emotion intensity. Equity Evaluation Corpus (EEC)[8] [41] was the benchmarking data for SemEval-2018 Task 1: Affect in Tweets. 8,640 English sentences are carefully chosen in this dataset to tease out biases towards races and genders. Both 11 emotion or sentiment and their intensity are provided. EmoInt Dataset[9] (also called Tweets-2016) annotates 7,100 English tweets with four emotions: joy, sadness, fear, and anger [42]. The emotion intensity is assigned between -1 and 1, indicating the sentiment with -1 being negative and +1 being positive. It was the benchmarking data for WASSA-2017 Shared Task on Emotion Intensity. Ren-CECps [43], [44] contains 1487 blogs with 34, 719 sentences in Chinese. It is hierarchically annotated with eight basic emotions and corresponding intensity in the document, paragraph, and sentence level.

Some word-level emotion lexicons are providing both emotion and intensity as well. NRC Emotion Intensity Lexicon [45] provides real-valued intensity scores from 0 (lowest) to 1(highest) of four basic emotions (anger, fear, sadness, joy). DepecheMood++ [46] is built in a completely automated and domain-agnostic fashion. It is annotated with eight emotions and contains three types of word representations: 188k tokens, 186k lemmas, and 285k lemma#PoS in English. It is also available in Italian.

2.2.2 Datasets with Dimensional Annotation

The dimensional emotion model measures emotion states with numerical dimensions, and each emotion state is represented as a multi-dimensional vector. In each dimension, the value is continuously changed to distinguish the nuances of emotion, and the extremes of two directions mean two polarities. PAD model is a commonly used dimensional model [47], representing emotions with three dimensions: pleasure, arousal, and dominance [48]. Russell's circumplex model [49] thinks that the two dimensions of VA (Valence and Arousal) could represent the most different emotions.

Compared with discrete emotional corpora, it is more challenging to construct a high-quality dimensional emotional corpus, and the available public corpora are relatively limited. SEMAINE Database[10] contains approximately 240 character conversations [50]. They are the interaction records between a human user and a human operator who pretends to be an artificially intelligent agent with a prototypic emotional character. This study is still ongoing, and currently, approximately 80 conversations have been fully dimensional annotated.

There are some dimensional annotated emotion lexicons. ANEW [51] (Affective Norm for English Words) and NRC_VAD Lexicon [52] are annotated with valence, arousal, and dominance scores. They contain near 2k words and 20k English words, respectively. SenticNet-6

[8] Access: *https://competitions.codalab.org/competitions/17751#learn_the_details-datasets*
[9] Access: *http://saifmohammad.com/WebPages/EmoInt2017.html*

[10] Access: *https://ibug.doc.ic.ac.uk/research/semaine-database/*

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Robert Gordon University. Downloaded on May 28,2021 at 02:10:33 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3053275, IEEE Transactions on Affective Computing

[53] is an affective commonsense knowledge graph consisting of 200k commonsense concepts. It is automatically annotated based on the Hourglass of Emotions [54] and is available in 40 different languages.

TABLE 2
PUBLICLY AVAILABLE EMOTIONAL LEXICON

Emotional lexicons	Scale	Annotation
WordNet-Affect [11]	2874 synsets, 4787 words	Hierarchical affective labels
SentiSense Affective Lexicon [12]	5,496 words, 2,190 synset	14 emotions
NRC Emotion Lexicon [13]	14k words, 25k senses	8 emotions
LIWC2015 [14]	6400 words, word stems, and emoticons	Hierarchically annotated
NRC Emotion Intensity Lexicon [15]	6k entries	4 emotions
DepecheMood++ [16]	188k tokens, 186k lemmas, 285k lemma#PoS	8 emotions
SenticNet [17]	200k concepts	Dimensional
NRC_VAD Lexicon[18]	20k words	Dimensional
ANEW [19]	2k words	Dimensional

2.2.3 Categorical v.s. Dimensional Emotion Model

Some works annotate emotional datasets with both categorical emotion labels and dimensional values to explore the associations between two annotation strategies. The Interactive Emotional Dyadic Motion Capture (IEMOCAP) Database[20] is an acted, multimodal, and multispeaker database [55]. It contains approximately 12 hours of audiovisual data, including video, speech, motion capture of face, and text transcriptions. IEMOCAP is manually annotated with dimensional and categorical labels. EMOBANK corpus[21] contains 10,548 texts and is manually annotated according to the psychological VAD scheme from both writer and reader [56]. Because a subset of EMOBANK has been previously annotated with Ekman's 6 Basic Emotions, it becomes possible to do the mapping between two annotation strategies.

The psychological emotion model is the basis of a high-quality emotional corpus. Although many emotion models have been proposed, there is no uniform standard of emotion annotation. In existing research, the discrete emotion model is commonly accepted due to its intuitive and straightforward advantages. However, human's natural emotion is very complicated. The generation, development, and disappearance of emotions is a continuous process, and it is challenging to be described by a discrete emotion model [57]. In recent years, the dimensional emotion model has received more and more attention with its stronger representation ability. Especially in dynamic scenarios, such as dialogue or human-computer interaction, dimensional emotion models are more conducive to tracking the dynamic emotion states than discrete models.

3 DEEP LEARNING APPROACHES FOR TER

This section reviews some deep learning-based technologies commonly utilized in TER. We first introduce word embedding technology, which maps the input sequence to a continuous vector and generates the underlying input representation. Then we introduce some neural network architectures and knowledge enhanced representation for emotion recognition. They focus on creating and training an effective network to learn multi-layered emotional features automatically. We also reviewed some transfer learning techniques that can alleviate the problems caused by scarce training data.

3.1 Word Embedding

Word embedding is a technique based on distributional semantic modeling and aimed to learn latent, low-dimensional representations from the language structure. Pre-trained word embedding alleviates the problems of sparse features and high-dimensional representation in traditional bag-of-words models. Some well-established embedding models are widely used and have been proven to boost the performance of NLP tasks.

According to the different emphasis of encoding information, word embedding can be divided into two kinds: typical word embedding and emotional word embedding. The former focuses on learning continuous word embedding by modeling the general semantic and contextual information, while the latter focuses on encoding emotional information into word embedding.

3.1.1 Typical Word Embedding

Early word embedding models are usually trained based on the syntactic context. They believe that frequently cooccurred words are often similar in some semantic criteria. Such as Word2vec [58], [59] and GloVe [60], they are trained on a large scale of unlabeled data and aimed to capture fine-grained syntactic and semantic regularities. The pre-trained word embedding model performs better than randomly initialized word vectors and has achieved great success in NLP tasks. However, early word embedding models assume that 'A word is represented by a unique vector' and ignore the influence of different contextual information. They embed each word into a unique vector, whether monosemous or polysemy. This way, antonyms words with the same language structure often have similar vectors [61]. This limitation of meaning conflation hampered The effectiveness of the early word embedding model.

Inspired by the successful transfer learning of CNNs from the image field to other computer vision fields, the emergence of pre-trained language model opened the pre-training era in the NLP field. They generate contextualized word embedding with general knowledge that can be easily transferred to almost all downstream tasks. The works in [62] learn contextualized word vectors (CoVe)

[11] Access: *http://wndomains.fbk.eu/wnaffect.html*
[12] Access: *http://nlp.uned.es/~jcalbornoz/SentiSense.html*
[13] Access: *http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm*
[14] Access: *http://liwc.wpengine.com/compare-versions/*
[15] Access: *http://saifmohammad.com/WebPages/AffectIntensity.htm*
[16] Access: *https://github.com/marcoguerini/DepecheMood/releases*
[17] Access: *https://sentic.net/downloads/*
[18] Access: *http://saifmohammad.com/WebPages/nrc-vad.html*
[19] Access: *https://csea.phhp.ufl.edu/media/anewmessage.html*
[20] Access: *https://sail.usc.edu/iemocap/index.html*
[21] Access: *https://github.com/JULIELab/EmoBank*

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Robert Gordon University. Downloaded on May 28,2021 at 02:10:33 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3053275, IEEE Transactions on Affective Computing

by training an encoder in the supervised machine translation system. Their CoVe achieved satisfactory results when transferred to other NLP tasks. Subsequent works devoted to training language models directly on large amounts of unlabeled data. ELMo (Embeddings from Language Models) [63] is a deep bidirectional language model. ELMo dynamically generates word embedding by capturing the variation of word meaning depending on its context. ELMo can be easily transferred to existing models and significantly improved the state-of-the-art across six challenging NLP problems. Soon afterward, transformer-based language model GPT [64] is proposed by OpenAI. GPT's intelligent generation performance demonstrated the powerful characterization capabilities of the transformer. BERT [65], the Bidirectional Encoder Representations for Transformers, achieved state-of-the-art performance on several NLP tasks when published, demonstrating its effectiveness in encoding contextual knowledge. Through a deep network architecture, BERT tries to predict unseen words in the context by unsupervised learning. A large amount of unlabeled data is utilized during the training, which helps the model learn useful linguistic knowledge. The emergence and success of GPT and BERT caused an upsurge in NLP and inspired more and more pre-trained models, such as GPT-2 [66], GPT-3 [67], Transformer- XL [68], XLNet [69], MASS [70], and UNILM [71]. Pre-trained language model has been shown to work well in practice on multiple tasks and has become a new paradigm for natural language processing.

3.1.2 Emotional Word Embedding

Emotional knowledge can be represented in different ways. Motivated by typical word embedding, emotional word embedding has shown outstanding contributions in different emotion-related tasks, such as emotion classification and emotion intensity prediction.

To learn generalized emotion representation, Emo2Vec is proposed in [72] to encode emotional semantics into vectors. This model is pre-trained in six emotion-related tasks by multi-task learning. Emo2Vec is often utilized by combing with other typical word embeddings, contributing to more competitive performances. Emojis have been widely used in electronic messages to express emotions and have been a large part of popular culture. To directly map emojis to continuous representations, Emoji2vec [73] is released for all Unicode emoji representations. DeepMoji [74] generates representations with rich emotional information. It is pre-trained on 1.2 billion tweets by utilizing a two-layer BiLSTM network along with the attention mechanism. In [75], a sentiment-specific word embedding (SSWE) model is proposed. This model integrates sentiment information into loss function, distinguishing words with similar syntactic context but opposite sentiment polarity. A domain-sensitive and sentiment-aware embedding (DSE) model is proposed in [76], which jointly models the sentiment semantics and domain specificity of words. These works incorporate emotional information into word embedding, leading to satisfying performance of emotional recognition.

Through transfer learning, pre-trained emotional word embedding contributes to various target tasks related to emotion understanding. The works in [77] compared the performance of several well-known pre-trained embedding models, including DeepMoji, ELMo, GLoVe, Emo2vec, BERT, and Emoji2Vec. Their experimental results demonstrate that DeepMoji features outperform others by a large margin. According to their discussion, the reason may be that DeepMoji is trained on a large emotion corpus, and their training data is similar to the target task. Therefore, ensuring the relevance of the source task and the target task can further contribute to feature representation. It is essential to select an appropriate pre-trained language model to generate word embedding.

3.2 Basal Neural Networks and Derived Variations

Deep neural networks have become a prevalent method for automatic feature representation. In general, neural networks consists of successive building blocks following by non-linear activation functions. Due to the multilayer nature, each successive layer is hypothesized to represent the data more abstractly [19]. The outputs of final or penultimate neural layers are often extracted as textual feature representation. Then, they are fed into classifiers for final prediction.

Pooling operation on the embedding layer is the simplest way for input sequence encoding [78]. Simple pooling works well for long documents (hundreds of words). However, its performance on TER is often unsatisfactory. The pooling operation is too simplified to ignore the word order information, while emotion-related tasks are more sensitive to word-order features. In recent years, various neural network models have been proposed to extract sequence information and useful semantic information, such as Recurrent neural networks (RNNs), Long-short term memory (LSTM), Gated Recurrent Neural Nets (GRNNs), and other variations. These recurrent networks are often severed as the core of typical emotion recognition systems. They have achieved highly competitive performance in sequence modeling and context encoding [79], [80]. As the emotion recognition system proposed in [81], their model mainly consisting of three layers: pre-trained ELMo layer for word embedding, BiLSTM layer for context learning, and dense layer for final prediction. As the works in [82], they develop an emotion prediction model based on GRNN, acquiring a superior accuracy on recognizing 24 fine-grained emotions. Besides recurrent networks, convolutional neural network (CNN) is efficient at text encoding as well. CNN extracts features through deep-layered convolutional kernels and pooling operation [83]. CNN deals with input sentences of different lengths and generalizes a feature map to decipher abstract concepts within a sentence [84].

The attention mechanism is a hot topic in deep learning, contributing to the TER model focusing more on emotion-related words. Transformer [85] is a self-attention mechanism-based architecture with powerful characterization capabilities and improved training speed. It solves the tricky long-term dependency problem in NLP and abandons the dependency of the traditional

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Robert Gordon University. Downloaded on May 28,2021 at 02:10:33 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3053275, IEEE Transactions on Affective Computing

NLP model on CNN and RNN networks. Transformer is the basis of pre-trained language models that have recently received widespread attention and has been a powerful representation learning model in most NLP tasks.

The hierarchical structure is a natural characteristic of textual data: words form phrases, phrases form sentences, and sentences form documents. Modeling this structured knowledge could contribute to better feature representation. Inspired by this fact, some works try to stack deep learning architectures to provide a specific understanding at each level [86]. Hierarchical Attention Network (HAN) [87] mirrors the hierarchical structure of documents. HAN mainly consists of two GRU-based attention layers applied at the word and sentence level, respectively. TreeLSTM [88] is a generalization of LSTMs to tree-structured network topologies. They think the typical LSTM network is a linear chain, while the tree-like syntactic properties (word-phrase-sentence) exist in natural language. Their experiments on sentiment classification demonstrated the effectiveness of their TreeLSTM architecture on semantic representations.

Affective modifiers play an essential role in the task of emotion recognition. The recurrent networks, such as RNN and LSTM, retain the affective semantic features at various time steps. However, they ignored the effect of modifiers, such as negation and intensification, which should receive more attention. In [89], a modified GRNN: a gated recurrent neural network with sentimental relations (GRNN-SR) is proposed to model the context. There are two hidden states in each time step, and the information of sentiment polarity and sentiment modifier context are separately encoded.

Multi-feature fusion is commonly utilized to enhance feature representation. The works in [90] conduct feature fusion by integrating general knowledge provided by fine-tuned BERT and domain knowledge extracted by a convolutional network. Their experimental results highlight the importance of domain knowledge in domain-specific applications. In [91], an architecture with two parallel attention-LSTM towers is proposed to focus its encoding on emotion-specific words. Two kinds of word embeddings are fed into parallel towers separately. Then, feature fusion and max-pooling are operated on the outputs of parallel towers to extract the most prominent features.

The ensemble model combines multiple individual models to generate a more robust and comprehensive model. Some works realize ensemble by varying training data with the same classifier. They divide the training data into multiple subsets, and each subset is utilized for training an individual model. Other works prefer to build an ensemble model by varying individual classifiers with the same training data. There are often differences between these individual classifiers, such as the dimension of the hidden layer, the overall structure, and even the learning rate during training. In [92], an ensemble model is proposed for intensity prediction on four emotions: anger, fear, joy, and sadness. Their model consists of three individual models, including feed-forward neural networks, multi-task learning networks, and CNNs-LSTMs based networks. Their final prediction is a weighted average among individual models. This work achieved the best performance in the WASSA 2017 shared task on emotion intensity. In [93], the ensemble model is built based on three classifiers, including a knowledge-based classifier and two statistical classifiers (Naïve Bayes and Maximum Entropy learner). Their final prediction is obtained by the majority voting approach. In [94], five independent models are trained on image and textual features, including Naive Bayes, Bayesian Network, decision tree, KNN, and SVM, which are then ensembled by the Bayesian model averaging method [95]. In an ensemble model, each classifier can complement each other, contributing to a more heterogeneous set of mapping functions.

3.3 Knowledge Enhanced Representation

With the development of deep learning networks, prior knowledge bases are often incorporated as auxiliary information and have been proven to be essential for TER tasks. The prior knowledge includes emotional lexicon resources, commonsense, linguistic patterns, affective semantic rules, or any other emotion-related knowledge. Incorporating prior knowledge can enhance emotional feature representation of textual data for deeper understanding [96].

Deep learning-based features are often combined with lexicon-based features. It is the most direct way to introduce emotional resources into deep learning networks and realize emotion enhancement [91]. The combined features are fed into a deeper network to generate more high-level and abstract features or fed into classifiers directly for final prediction. In [97], to detect emotion states from health-related posts, they combined lexicon-based features and the output of CNN network, which are then fed into LSTM network for further learning. In [98], they extract feature representation from an intermediate layer of their pre-trained network and then concatenate the extracted features with some hand-crafted feature vectors, such as TF-IDF weighted word vector and lexicon-based features. In [99], to alleviate the problem caused by misspelling and out-of-vocabulary words, they combined lexical features with neural features to boost performance. Incorporating lexicon-based emotional knowledge helps capture the hidden semantics and provides a more insightful understanding of emotional texts.

Some works concentrate on enhancing word-level representation with external knowledge. It is easier to infer the implicit emotion in textual expression through the rich information of relevant common sense knowledge. In [100], they try to explore implicit emotion by utilizing the external commonsense knowledge from ConceptNet and emotion intensity information from the NRC_VAD lexicon. With a context-aware affective graph attention mechanism, they dynamically retrieve the context-aware concepts and obtain concept-enriched word representation. In [101], a knowledge-enriched two-layer attention network is proposed. Their primary word-level attention is applied to input word and related terms obtained by searching WordNet and Distributed Thesaurus, generating word embedding enhanced by the knowledge graph.

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Robert Gordon University. Downloaded on May 28,2021 at 02:10:33 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3053275, IEEE Transactions on Affective Computing

The secondary attention mechanism works on sentence-level for further context learning. Their proposed system performed remarkably well on the benchmark datasets of SemEval 2017 Task 5. In [92], their emotional features of tweets are generated based on seven lexicons with the filter in the AffectiveTweets: TweetToLexiconFeatureVector [42]. In [102], NTUA-SLP embedding is proposed to incorporate emotional seed words into word embedding. Start from a set of seed words with affective ratings ([-1,1]), the embeddings for new words are estimated by considering both semantic similarity and ten affect-related features, including valence, dominance, arousal, pleasantness, anger, sadness, fear, disgust, concreteness, and familiarity.

It is practical and convenient to generate emotional word embedding from a pre-trained word embedding model. In [74], pre-trained DeepMoji is utilized and contributes to learning richer emotional context representation at the sentence level. In [103], Sentiment and Semantic-Based Emotion Detector (SS-BED) is proposed to learn sentiment and semantic features. It contains two LSTM layers that take two sequences of word embedding as input separately. One is sentiment representation obtained by SSWE (Sentiment Specific Word Embedding), and another is semantic word representation obtained by pre-trained Word2Vec, Glove, and FastText [104]. This approach is evaluated on real-world dialogue datasets and significantly outperforms traditional machine learning baselines.

Rule-based representation is another commonly accepted form of knowledge representation. In [105], rule-embedded neural networks (ReNN) are proposed to encode domain knowledge and commonsense information. The rule-based knowledge reduces computing complexity and helps to train a better model with a smaller scale of the dataset. To perform knowledge base completion, ITransF is proposed in [106], which discovers hidden concepts of relations and transfers statistical strength by sharing concepts.

Linguistic patterns play an important role in emotion recognition because of their ability to shift the emotional tendency of the entire sentence [107], [108]. The works in [109] propose a linguistically regularized LSTMs, including sentiment, negation, and intensity regularizers. Their model addresses the sentient shifting effect of linguistic roles and enhances the sentence-level emotion representation.

3.4 Transfer Learning for Emotion Recognition

Emotional resources are the guarantee of satisfying performance in the TER system. Collecting and annotating large amounts of data is time-consuming and expensive. The shortage of high-quality emotional datasets is always the most urgent problem. How to train an effective TER model with smaller datasets has attracted increasing attention.

By transfer learning, the knowledge learned from the source domain is transferred to the target domain, realizing performance improvement. In this paper, the target task refers to TER, and the source task can be any other related NLP tasks, including sentiment analysis and machine translation. The valid information learned from related tasks can be transferred into the target TER model. Transfer learning can alleviate the problems caused by scarce training data and speed up training, which improved the performance of deep learning models. Gupta [110] investigates the application of semi-supervised and transfer learning methods in low-resource sentiment classification tasks and demonstrates that transfer learning could significantly improve the performance compared with the simple supervised method.

There are various kinds of architectures proposed based on transfer learning. They can be categorized into inductive transfer learning, transductive transfer learning, and unsupervised transfer learning [111]. Among them, training data of the target domain is labeled in inductive transfer learning while unlabeled in another two. Recent works about TER with transfer learning are mainly based on inductive transfer learning. Based on some different situations, related works can be summarized into two sub-cases. The first is sequential transfer learning, by which source and target tasks are learned successively. Another is multi-task learning, by which source and target tasks are learned simultaneously.

3.4.1 Sequential Transfer Learning

Sequential transfer learning is arguably the most frequently used transfer learning scenario in the NLP field. The process generally consists of two stages. The first is pre-training on the source tasks, such as emotion-related tasks or other natural language understanding related tasks. The second is the transfer phrase, in which the knowledge learned in the source domain is transferred to the target TER task.

During the pre-training phase, many efforts are devoted to train language models with universal knowledge of the natural language. As the aforementioned pre-trained language models (detailed in Section 3.1), they can be transferred to almost all NLP tasks and have advanced multiple state-of-the-art performances. Some works are devoted to pre-training their model on emotion-related source tasks with sufficient training data, such as sentiment classification and emotion intensity regression.

In the transfer phase, there are mainly two ways to realize the transformation. One is taking the pre-trained model as a feature extractor, and all layers are frozen. The generated feature representations are fed into an upper-level TER model for further learning and final prediction. Another is fine-tuning the pre-trained network on the target task. This operation is often accompanied by minor modifications of the network architecture, such as replacing the prediction layer and fine-tune the network. Layer-wise fine-tuning can adjust the individual patterns across the network with a reduced risk of overfitting. Some commonly used fine-tuning strategies are compared in the task of cross-lingual emotion classification [26]. Their experimental results show that the performance of gradual unfreezing [112] and single top-down unfreeze [74] are slightly better than others on the fine-tuning phase.

Sentiment analysis and intensity regression are often

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Robert Gordon University. Downloaded on May 28,2021 at 02:10:33 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3053275, IEEE Transactions on Affective Computing

severed as source task in transfer learning-based TER. Compared with existing emotional corpus, sentiment related datasets with polarity annotation are abundant and are widely applied in sentiment analysis tasks. Although their content and labeling system are different, both of them contain rich emotional characteristics. Such information can be utilized by knowledge transfer and enhance the target TER model [113]. With transfer learning, the deep attentive RNNs model proposed in [102] ranked 1st in semeval-2018 task 1 'Multi-Label Emotion Classification'. They first obtain affective word embedding based on a small number of emotional seed words and then pretrain their model on the sentiment dataset: Semeval 2017 Task 4A. Their final layer is replaced with a task-specific layer model and then further fine-tuned with two fine-tuning schemes. In [114], an ensemble of transfer learning techniques is proposed to predict the emotions of removed emotion trigger words. They utilize three different pre-trained models to initialize some specific layers of their networks, including a language model, a word embedding model, and a sentiment model. Then, they ensemble and fine-tuning these models in their dataset and have achieved competitive experimental results. In [115], a dual attention-based transfer learning approach is proposed for multi-label emotion classification. They respectively captured typical sentiment features and emotion-specific features with a shared attention layer, which are respectively fed to the task-specific layer by a dual attention mechanism. Experimental results show that their dual attention transfer architecture can bring consistent performance gains compared to several existing transfer learning approaches.

In TER, the existing emotional resources are mainly in English, while there are low resources in most European languages. The quality of the TER model with few resources is often limited. Therefore, to those low-resource languages, it is advantageous to leverage the emotional information from resource-rich languages. The works in [26] try to solve the low-resources problem in Hindi emotion detection and propose a deep transfer learning framework to transfer emotional knowledge from English to Hindi. They train cross-lingual word-embeddings by mapping each monolingual word-embedding into a shared space with the transformation strategy of alignment matrices [116]. Therefore, relevant information can be captured through the shared space. Based on the cross-lingual word embedding, the deep learning model is pre-trained on English emotional datasets, and then fine-tuning is done in Hindi datasets.

Cross-domain transfer learning aimed to transfer knowledge across different domains, utilizing a small amount of labeled data from the target domain and abundant labeled data from a different source domain. The data distribution and labeled emotions from different domains have a significant impact on the performance of transfer learning. In [117], they try to transfer emotional knowledge from the source domain through joint learning with a domain classifier, promoting the performance of emotion classification through the sharing of domain-specific representation. In [118], to make transfer learning methods more robust to unseen data, Adversarial Discriminative Domain Generalization (ADDoG) is proposed to generalize the representation of cross-corpus. ADDoG follows a 'meet in the middle' approach, iteratively moves their dataset representations closer to one another, and improves the cross-dataset generalization. In [119], transfer learning is utilized to address the task of cross-domain and cross-category emotion tagging for comments on online news. They achieve domain adaption through reweighting instances from the source domain by modeling the distribution difference. They also model the relationship between different sets of emotion categories from each domain, enabling project data from one domain into the label space of another domain.

3.4.2 Multi-Task Learning

Multi-Task Learning (MTL, also known as joint learning) is another form of inductive transfer. Various studies have shown that MTL dramatically improves the performance of TER systems than single-task learning. In MTL, target and source tasks are related and trained simultaneously[120]. Through underlying shared representations, sub-tasks promote and supply each other to learn more relevant information. Compared to the single-task framework, multi-task learning on related tasks can significantly reduce the risk of overfitting, contributing to better generalization performance and the improvement on all sub-tasks.

MTL framework targets to enhance the generalization performance by leveraging the inter-relatedness of multiple tasks [121], [122]. Taking emotion-related tasks as auxiliary tasks is ideal for the MTL-based emotion recognition system, which indirectly realized emotional information integration from different resources [123]. Most works are conducted based on the structure consisting of shared networks and some task-specific layers. In the Emo2Vec model proposed in [72], to encode emotional semantics into word-level representations, six different emotion-related tasks are trained simultaneously. Their generalized emotion representation outperforms multiple existing affect-related representations, such as DeepMoji, but with much smaller training data. In [124], a two-stage multi-task learning structure is proposed to complement the feature representation in the dimensional model with the knowledge transferred from the discrete model, thereby establishing a relationship between discrete and dimensional emotion. Rather than parameter sharing, the works in [125] realize label transformation from sentiment label to emotion label by joint learning and have improved the performance of emotion classification.

To address the time-consuming problem in emotion annotation, multi-task active learning for regression (ALR) is proposed in [126]. The most beneficial samples are selected in their model, benefiting the emotion estimation in three dimensions (valence, arousal, and dominance) simultaneously. In [98], a multi-task ensemble learning framework is proposed for several tasks related to category/dimensional emotion, sentiment, and intensity. They firstly pre-trained three individual networks by multi-task learning and obtained three task-aware deep repre-

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Robert Gordon University. Downloaded on May 28,2021 at 02:10:33 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3053275, IEEE Transactions on Affective Computing

sentations. These representations are combined with other hand-crafted features and then fed into a multi-task ensemble model for further learning. This multi-task ensemble framework helps in achieving generalization and contributes to superior results.

The difference of label distributions between the training and test sets is considered in [127]. They find that most of the errors are raised by recognizing the 'Others' category. They think their performance could be better if firstly conduct the binary classification 'Others' versus 'Not-Others'. This problem is considered in [128]. They achieve better performance by utilizing a multi-learning network and can better detect emotions from the 'Others' class.

Sub-tasks in MTL are highly correlated, and label relationships among all tasks can provide useful information as well. To address emotion ambiguity in the textual expression, a multi-task CNN model is proposed in [129], which learns emotion label distribution and emotion classification simultaneously. These two tasks boost each other and thereby generate a robust text representation. In [130], An Adversarial Attention Network (AAN) is proposed to conduct adversarial learning between each pair of emotional dimensions. They conduct multidimensional emotion regression tasks by multi-task learning, and they perform well on the EMOBANK corpus. In [131], a joint label space is induced to enable multi-task learning from both labeled and unlabeled data. They exploit the relationships between different labels from all tasks according to a label transfer network, demonstrating that potential synergies between label spaces can be leveraged for label transformation.

TABLE 3
A SUMMARY OF TER METHODS

Techniques	Details (Strength and weakness)	Examples
Typical Word Embedding	• Learn latent, low-dimensional representations from language structure. • Pre-trained language model provides contextual word embedding and relieves the restriction of meaning conflation deficiency. • Pre-trained on large amounts of datasets containing useful linguistic knowledge. It can be easily transferred to TER tasks.	Word2vec [58], [59], GloVe [60]. CoVe [62], ELMo [63], GPT [64], BERT [65], GPT-2, [66], Transformer- XL [68], XLNet [69], MASS [70] and UNILM [71].
Emotional Word Embedding	• Pre-trained on emotional resources, contains both semantic similarity and prior emotional knowledge • Can be fine-tuned to downstream tasks.	Emo2Vec [72], Emoji2vec [73], DeepMoji [74], SSWE model [75], DSE model [76], NTUA-SLP embedding [102], AffectiveTweets [42].
Basal neural networks and derived variations	• Be able to generalize and capture context over the sentence and hierarchical level. • Attention mechanism makes networks focus more on emotion-related words. • Data-driven and abundant resources are necessary while collecting and annotating large amounts of data are time-consuming and expensive.	RNN (LSTM or GRU) [79], [80], [81], [82], CNN [83], [84], Hierarchical Network [86], [87], [88], Modified network [89], Feature fusion [90], [91], Ensemble network [92], [93], [94], [95].
Knowledge-enhanced text representation	• Prior knowledge bases are utilized as auxiliary information for deeper understanding. • Compensate for the lack of training data to some extent. • Enhance emotional feature representation.	Fusion of emotional features and deep features [91], [97], [98], [99], Utilize emotional word embedding [74], [103], Knowledge-enriched network [92], [100], [101], [102], [105], [106], [107], [108], [109].
Sequential transfer learning	• Transfer emotional information from the emotion-related tasks and general information from other NLP tasks. • Target and source tasks are related and trained successively.	Transfer from pre-trained language models [26], [114], emotion-related source tasks [102], [113], [114], [115], Cross-language [26], Cross-domain [117], [118], [119].
Multi-task learning	• Enhance the generalization performance by leveraging the inter-relatedness of multiple tasks. • Target and source tasks are related and trained simultaneously through underlying shared representations	Joint learning for knowledge transfer [72], [98], [122], [123], [124], [126], [128], Label transfer [129], [130], [131].

4 CHALLENGES

TER task brings some open challenges for NLP researchers. Beyond the effectiveness of the TER algorithm itself, the challenges of TER can be attributed to many problems, from the shortage of high-quality data to the complex nature of textual emotion expression. In this section, existing challenges and potential opportunities are discussed from the following four aspects:

1) Shortage of high-quality datasets. High-quality emotional datasets are the guarantee of a high-performance TER system. Although many public emotion-related databases have been proposed, most of them are limited by

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Robert Gordon University. Downloaded on May 28,2021 at 02:10:33 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3053275, IEEE Transactions on Affective Computing

the quantity. There is no uniform emotional annotation scheme, which leads to incompatibility between different corpora. Besides, the imbalanced distribution of data between each emotion category also affects the performance of the TER model.

2) Fuzzy emotional boundaries. The nature of human emotional expression is complex. There is no clear dividing line between emotions, such as love and happiness. Besides, due to people's own experiences and current feelings, the emotions could be different even in the same expression. The multi-label emotion recognition task is derived from this challenge. Thereby the contained emotions can be described in more detail by assigning multiple emotional labels to a textual expression.

3) Incomplete emotional information in textual expression. In actual emotional interactions, human emotional expressions are more introverted, and it is not comprehensive to make emotion prediction based solely on textual information. Therefore, some researchers focus on emotion recognition by using multi-modal information, intending to alleviate the one-sided problem of emotion information in textual modal to a certain extent.

4) Textual emotion recognition in dialogue. TER is an essential part of a successful and intelligent dialogue system. Most TER is conducted based on the assumption that the sentence is separated from the context and expresses static emotion. However, textual emotion is dynamic and highly correlated with the contextual information, making TER task more challenging.

4.1 Shortage of Large-Scale and High-Quality Dataset

4.1.1 Incompatible Annotating System

Discrete and dimensional emotion models are widely used in existing TER tasks. However, there is no uniform annotation standard, which results in incompatibilities between different databases. For example, SemEval-2007 [32] is annotated with Ekman's six basic emotions, EmoInt Dataset [42] is annotated with four emotions, and Ren-CECps [43] is annotated with eight emotions. Due to the different emotion labels utilized, cross-corpus emotion resources are often not compatible. Incompatible annotation can explain why there are still insufficient training data, although many emotional corpora are released.

In future work, building an emotion recognition benchmark, or creating a link between the annotation schemes of different emotion corpora, will contribute to integrating existing corpus resources and creating a more substantial database covering all emotion categories.

4.1.2 Data Imbalance in Emotional Dataset

Many works devote to training a TER model by minimizing misclassification errors. It is noteworthy that some inherent properties of emotional datasets should be investigated to further enhance the generalization performance. In most research, it is assumed that the data distribution of each emotion in the corpus is balanced, which means that the number of texts annotated with each emotion is roughly similar. However, this is often not the case. Most of the existing emotional datasets are manually collected and annotated. Thereby the data distribution of each emotion is always unbalanced. For example, in RenCECps [43], 3% of sentences are annotated with 'Surprise' while 22% of sentences are annotated with 'Love'. While the dataset is imbalanced, maximizing the overall accuracy may not be the best approach. The classifier could tend to predict the emotion labels with more data [132], which is always unavoidable in the TER task.

The most direct way to eliminate data imbalances is to adjust the data distribution by sampling techniques such as under-sampling and over-sampling. The under-sampling technique eliminates the data of majority classes to balance the data distribution in each class. However, in actual emotion recognition tasks, emotional resources with annotation are often limited. To maximize the usage of labeled data, under-sampling is usually not the best option. The over-sampling technique often replicates the data in the minority classes to achieve an overall relative balance distribution. In [133], the authors think that data imbalance has a significant impact on the performance of neural network models and conclude that over-sampling is the most effective way to combat this problem. In [134], a data augmentation method with generative adversarial networks (GAN) is proposed to balance label distribution. During training, they generate the auxiliary data of minority classes with CycleGAN [135], which supplements low-dimensional data manifold and helps find the margins of neighboring classes. In [136], to address data imbalance in NLP tasks, they discuss four data augmentation techniques: synonyms replace, randomly insert, randomly swap, and randomly delete. Their experimental results demonstrate that simple data augmentation operations could boost performance and reduce overfitting when training on smaller datasets. These over-sampling methods can alleviate the problem of data imbalance in TER tasks to a certain extent. However, if the training data is large enough, the model itself could already be able to generalize, and over-sampling may not work. On the other hand, when the training data is small (this is often the case for TER tasks), over-sampling could lead the model to be overfitting and increase the computation cost and training time [137]. How to prevent overfitting and improve the generalization ability of the model has been a critical issue in research.

During training a TER model, it is often assumed that the cost of misclassification of each emotion label is the same. However, in the case of data imbalance, we usually hope the classifier pays more attention to recognizing the minority emotion labels. For example, the recognition accuracy of 'Surprise' is often lower than other emotions because of its rare training data. The minority emotion labels should be given more weight during the training so that the classifier could be more sensitive to their recognition. Some works attempt to modify the classification algorithm to pay more attention to minority classes during training, such as modifying the loss function. In [113], the weighted loss function is proposed for unbalanced data. In their loss function, the weight of each category is inversely proportional to the number of samples under the corresponding category. When the model is optimized

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Robert Gordon University. Downloaded on May 28,2021 at 02:10:33 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3053275, IEEE Transactions on Affective Computing

iteratively, more attention can be paid to identifying the minority emotion categories.

Addressing the challenges brought by data balance is crucial for improving the overall performance of TER model. It contributes to recognizing both common emotions and rare emotions accurately.

4.1.3 Low-Resource Dataset

In the TER model, a large scale of labeled emotion corpus guarantees the successful emotion prediction, especially for the deep TER model, which is data-driven. The annotation of an emotional corpus is a complicated task even for humans. Most of the existing emotional resources are manually labeled, and the large-scale emotional corpus is scarce [34], [138], [139]. Although manual labeling can guarantee accuracy to some degree, the labeling process consumes a lot of time and labor. Moreover, due to the influence of annotators' subjective feelings, the inter-annotator agreement cannot be guaranteed. The low-resource emotional dataset has been a critical challenge. How to alleviate low-resource problems and efficiently create high-quality emotional corpora is the key to improving the overall performance of the TER model.

Any attempt to create a large-scale emotional dataset can provide a bright future for this field. Seeking an efficient method for annotating emotional corpora has attracted researchers' attention, such as automatic or semi-automatic emotion annotation [140]. In [141], Mohammad tries to automatically create an emotion corpus from Twitter posts by exploiting emotion word hashtags. Automatic labeling with the hashtag is a simple and efficient technology in terms of time and cost. However, hashtags can only be utilized in social networks, and whether hashtags are correctly utilized is also a concern. The pre-annotation technique is a semi-automatic methodology, which employs automatic processes to help human annotators in manual tasks [142], [143]. As the works in [144], their EmoLabel provides large-scale emotional corpora to lessen and counteract the challenge of emotion annotation.

Many related works have proved that transfer learning can sufficiently alleviate the limitations brought by resource shortage. As mentioned in Section 3.4, transfer learning can transfer knowledge from other related tasks to target TER tasks to improve generalization performance. For example, it is possible to learn general language knowledge through pre-training models in other NLP tasks, such as reading comprehension and machine translation, which have abundant training resources. Emotional information can be learned in emotion-related tasks, such as sentiment classification. Even in a low-resource language, it is possible to leverage the resources and tools available in other resource-rich languages, such as transfer knowledge from English to Hindi. In future work, the TER task based on transfer learning may be developed around the following steps: 1) Pre-train a general model with sufficient unsupervised data. 2) Fine-tuning the general model by utilizing other related tasks with sufficient labeled data. 3) Further fine-tuning the model on TER task by multi-task or single-task learning.

The scale and quality of emotional resources can directly affect the final performance of the TER system. According to expanding the emotional corpus and borrowing other resources, it is possible to alleviate the limitations caused by the lack of emotional resources.

4.2 Fuzzy Emotional Boundaries and Multi-label Emotion Recognition

In existing emotion categorical models, emotions are divided into several categories, such as Ekman's six basic emotions, which are oversimplified and ignore emotion's diversity. Human emotion is complex in reality. Many emotional categories have a particular connection, and there is no distinct boundary. Thus it is difficult to match an accurate label for emotional expression. Besides, emotions are very subjective feelings. Treating the same textual expression, humans may feel different emotions according to their own experience. Thereby, TER becomes challenging because of fuzzy emotional boundaries and human's subjective feelings. Many researchers trend to describe emotions in more detail by assigning multiple emotion labels simultaneously [145]. Multi-label emotions are the combination of several basic emotions, which can describe complex emotions more comprehensively and alleviate the above limitations with a compromise approach. In recent years, multi-label emotion recognition (ML-EC) has attracted more attention, aiming to identify all possible emotions in textual expression [146].

Plenty of multi-label classification (MLC) techniques has been employed to identify multiple emotions [147]. Through commonly used problem transformation algorithm, MLC task is transformed into other well-established learning scenarios [148], such as multiple binary problems or multi-class problems. In [149], they transform ML-EC task into one binary classification problem by reconstructing the input training data. In their preprocessing, each multi-emotion sample is transformed into multiple xy-pairs with single-emotion. Thus they take the xy-pair as input, and the binary output indicates whether the input emotion label y is included in the input instance x. Problem transformation-based methods are conceptually simple and high efficient. Although they can achieve satisfactory performance, these approaches can be less effective and poor generalization due to their ignorance of label correlations. To capture label correlations, Classifier chains (CC) transforms MLC into a chain of binary classification problems. However, the wrong prediction of the previous classifier in CC can propagate the error to the latter ones along the chain.

With the development of deep learning, multi-class emotion classification algorithms can be easily extended to ML-EC task. The most common and direct way is replacing the prediction layer with ML-EC task-specific layer [102]. For example, modify the number of neurons in the output layer (usually the same as the number of emotional labels), set appropriate nonlinear activation functions (such as tanh, sigmoid, and ReLu), and thresholds. Such a framework focuses more on underlying emotional feature extraction, and the underlying layers are usually similar to multi-class emotion classification [115]. ML-EC for tweets is presented as a sub-task in SemEval-

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Robert Gordon University. Downloaded on May 28,2021 at 02:10:33 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3053275, IEEE Transactions on Affective Computing

2018 Task 1: Affect in Tweets. This task aims to identify whether emotions are included in tweets and then further identify all possible emotion labels. Most of the proposed works are based on the above framework. Competition participants widely use the pre-trained embedding models and existing emotion lexicons to generate robust text representation [91], [139]. The works in [150] consider that humans categorize the sentence to each emotion separately but not all at once. Thereby they realize ML-EC task by imitating how humans comprehend and then classify emotions with multiple independent CNNs.

Plenty of deep TER tasks focus more on generating robust text representation. In most cases, the emotional information of text is often encoded together into a representation vector and then directly fed into classifiers [151], [152], [153]. They treat ML-EC as a general classification task, without considering emotion correlation in both feature extraction and prediction steps. However, there is some relatedness among emotions, and the latent emotion correlation is an important clue. The performance of ML-EC model can be improved by emotion correlation learning. For example, as our common sense, emotions 'Joy' and 'Love' tend to appear simultaneously. It means that if 'Joy' is detected in a text, 'Love' may be contained with a higher probability than 'Hate'. Although the above common knowledge of emotion correlations increased the complexity, they can be utilized to facilitate more in-depth emotion analysis.

Label correlation is always a critical problem in MLC tasks [154]. Some approaches attempt to implicitly estimate label correlations by the modification of loss functions [155], such as the label-correlation sensitive loss function [156] and joint binary cross-entropy (JBCE) loss [153]. Deep canonical correlation analysis (DCCA) contributes to both feature-aware label embedding and label-correlation aware prediction [157], [158]. MLC and label correlations learning can also be realized in a joint learning framework [159], [160].

Some studies try to explore the label correlations by transform MLC tasks into ranking or generation tasks. In [161], they transform ML-EC into a relevant emotion ranking task, aiming to generate a ranked list of relevant emotion labels according to emotional intensity. Moreover, the prior knowledge of emotion co-occurrence is incorporated in their emotion-aware loss function as constraints. Such emotional relationships can provide essential cues for more accurate emotion prediction. In [162], MLC is regarded as a sequence generation task, and Sequence Generation Model (SGM) is proposed to consider label correlations. They introduced global embedding to alleviate the exposure bias caused by mispredictions made in the previous time steps. After that, some attempts based on the seq2seq model are proposed. In [163], CNN is employed as the encoder to capture high-level local sequential semantic information, while the global label correlation information is modeled by RNN network in the decoder. They also stacked an Initialized Fully Connection (IFC) layer in the decoder to incorporate prior knowledge of emotion label dependency information.

Emotional label-importance is an inherent property in ML-EC tasks [164]. There may be a dominating emotion with stronger intensity and others with a relatively weaker intensity in a textual expression with multiple emotions. However, most of the existing works are conducted based on the assumption of equal labeling-importance. Emotion-importance learning can often be regarded as emotion intensity prediction. By distinguishing the different importance of each emotional label in semantics representation, the generalization performance of the ML-EC system can be further improved.

Multi-label emotion describes human emotions more intuitively and in more detail. Effectively emotion correlation learning is the focus of future research, contributing to a more accurate ML-EC system.

4.3 Incomplete Emotional Information and Multimodal Emotion Recognition

Human perception in real life usually takes many forms, such as through vision and audio information. In real-time emotional interaction, it is challenging to achieve accurate emotion recognition only through textual information due to the relatively introverted emotional expression. For example, the emotion expressed in textual data is Joy, but this may also be the forced laughter, and the inner emotion maybe sorrow. In this case, it is difficult to accurately recognize the complex emotion merely through textual information, even for humans. Like the works in the manually labeled emotional corpus SemEval2007 [32], the inter-annotator agreement ranges from 0.36 (Surprise) to 0.68 (Sadness) with Pearson correlation measure, while other emotions are around 0.50. These results suggest that emotional information in textual expression is incomplete, and the TER task is difficult even for humans. An emotional recognition system that relies solely on text modality is hard to meet the requirements of robustness and accuracy, which significantly limits its application.

With the increasing amounts of installed webcams, an increasing amount of affective information is posted on social networks in the forms of audio and audiovisual rather than purely textual basis. To alleviate the one-sided problem of textual emotion information to a certain extent, more and more researchers have begun to use multimodal information for emotion prediction. Textual information is usually combined with other modalities, such as visual [17], audio [16], [165], body posture, and psychological single [19], [166]. The information of multiple modalities can mutually confirm and complement each other, providing more comprehensive information for emotion recognition. When textual emotional resources are limited (such as lack of annotated data or unreliable labels), the knowledge of other resource-rich modalities can be utilized to supplement textual modality according to deep learning technologies. Multi-modal information contributes to the overall performance of emotion prediction [167].

Visual, audio, and text are the most commonly combined modalities in multimodal emotion recognition [168]. Information is usually extracted from videos. When someone tends to express emotions in language, audio

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Robert Gordon University. Downloaded on May 28,2021 at 02:10:33 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3053275, IEEE Transactions on Affective Computing

data usually contains more emotional features in acoustic and linguistic information. Otherwise, if the tendency is to express emotions with facial expressions, emotional information is more concentrated on facial features.

Although the current research has made some achievements in visual and audio-based emotion recognition, there are still some challenges, such as obstructed facial expressions and noise interference. With the popularity of wearable devices, people pay more and more attention to detecting emotions through physiological signals, such as Electroencephalogram (EEG). Changes in physiological signals are closer to people's true inner emotional feelings than other modalities [169]. These signals reflect both various brain electrical activities and useful information about human emotional states. The generation or activity of emotions is highly correlated with the activity level of physiological signals [170], [171]. Therefore. EEG signals have been gradually studied for emotion recognition due to their advantages of strong objectivity [19], [172], [173].

Multi-modal emotion recognition system mainly includes two steps: feature extraction from each modal and multi-modal information fusion. The literature of multimodal feature fusion can be roughly distinguished in early and late fusion paradigms. Early fusion approaches work at the feature-level. Features extracted from each modality are fused and then fed into classifiers for final prediction. Late fusion approaches work at the decision-level. Some models are respectively pre-trained with each modality data, and the results from each modality are fused as a decision vector for final prediction. Previous works have demonstrated the effectiveness of multimodal features in creating rich feature representations [20], [21]. In [174], experimental results demonstrate that emotion prediction based on textual and audio features is more accurate than either. In [168], emotion is predicted by a multimodal emotional classification framework. In [94], an ensemble model based on the extended Bayesian model averaging method is proposed to fuse visual and textual information for emotion recognition. Experimental results show that the ensemble method based on a late decision-level fusion achieves higher classification performance than early feature level fusion.

Multimodal emotion recognition is conducive to realize knowledge supplements and strengthens among modalities. It provides more information for decision-making and improves the performance of emotion recognition. There are still some challenges of multimodal emotion recognition, as summarized in [167]. The first is how to extract features from each modality and effectively fuse them to achieve modal complementarity. Second, transformation (mapping) and alignment (synchronization) between modalities are still significant problems at present. Furthermore, how to transfer knowledge between modalities through effective co-training is particularly important when one modal has sparse features or limited resources.

4.4 Textual emotion recognition in dialogue

With the increasing amount of open dialogue data, TER in dialogue has received continuous attention in the NLP field. TER is an essential part of a successful and intelligent dialogue system. Most TER is conducted on sentence-level based on the assumption that the text is separated from the context and expresses static emotion. However, textual emotion in dialogue is dynamic and highly correlated with its context information, making dialogue TER more challenging.

4.4.1 Utterance Context Modeling

Directly applying the sentence-level deep TER algorithm to the dialogue, its performance is always not satisfactory. TER of each utterance highly depends on contextual cues, and context modeling is indispensable. It is essential to model the current utterance and contextual utterances during TER in dialogue, which helps to know the dialogue's overall emotional tendency.

To infer the underlying emotion in dialogue and bring more research on teaching machines to be empathetic, the shared task SemEval-2019 task 3 (EmoContext: Contextual Emotion Detection in Text) is organized to address the problem of TER in dialogue. This task aimed to recognize emotions (Happy, Sad, Angry, or Others) of the utterance based on its context. From submitted works, the methods for utterance context modeling can mainly be summarized into two kinds: (1) flatten context modeling and (2) hierarchical context modeling.

By flatten context modeling, context utterance and current utterance are concatenated, and all tokens are flattened into a word sequence. This word sequence is fed into neural networks for contextual learning and final prediction [1], [99]. However, emotions flow naturally in the dialogue, and the sequential nature of the dialogue is non-negligible. Flatten context processing makes the sequence of words too long and ignores the time step, destroying the hierarchical structure of the dialogue.

By hierarchical context modeling, each utterance in dialogue is embedded into a vector representation by utterance-level encoder, and context information is further extracted by hierarchy context-level encoder. Hierarchical LSTMs with attention mechanisms [175], [176] are the common choice for context modeling. In [77], plenty of superior models, such as LSTM and Universal Transformer, are modified into a hierarchical structure to extract more detailed emotional knowledge from context. Their contrast experimental results demonstrate that hierarchical architectures outperform simple flatten context modeling. In some other works, CNN network performs well in utterance-level embedding [79], while LSTM [177] and memory networks [178] perform well in context-level encoding. In [3], sequenced-based CNN (SCNN) with attention mechanism is proposed to facilitate sequential dependencies among utterances. They utilize CNN to extract features in both current and previous utterances, and then the sequence features are concatenated for further context modeling and final prediction. Their experimental results outperform the commonly used RNN-CNN [179]. They abandoned traditional sequence models, such as RNN, and demonstrate the capability of SCNN in context modeling.

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Robert Gordon University. Downloaded on May 28,2021 at 02:10:33 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3053275, IEEE Transactions on Affective Computing

Emotions expressed by humans often profoundly rely on commonsense knowledge, which is essential in context modeling. In [100], a Knowledge Enriched Transformer (KET) is proposed to address the above challenge and recognize real-time emotions in dialogues. Firstly, external commonsense knowledge is dynamically leveraged for concept-enriched word representation. Then, hierarchical self-attention in both utterance and context level is proposed to exploit the structural representation and learn contextual representation. Using lexicon resources is proven to provide significant improvements in identifying the emotion conveyed by a word, and this also applies to dialogue. In the semi-supervised multi-emotion detection model proposed in [155], they vectorize each utterance by combining the sum vector of all lexical items inside and its contextual information captured from the entire dialogue. The experimental results suggest that this method is effective and is only slightly less effective than human annotators.

4.4.2 Dynamic Emotion Modeling

The above methods mainly focus on tracking the contextual information and exploring the overall tone in dialogue. They mainly consider emotional inertia. They suppose that the participants tend to maintain a specific emotional state, and the context reflects a relatively consistent dialogue atmosphere. However, such an assumption ignores the real-time dynamic emotional changes of each party in a dialogue. The emotions of interlocutors are independent, but they influence each other during the dialogue. As the dialogue progresses, the emotions are in a dynamic shift because of the stimulation. It is affected by the emotional legacy of the party in the previous state and the stimulating emotion from other parties in the dialogue [177]. Dynamic emotion analysis in dialogue has a wide range of applications, such as emotional companionship of chatbots and healthcare dialogues of nursing robots. Dynamic emotion monitoring of both clinicians and patients during the dialogue is beneficial to better emotional management and guidance.

Dynamic emotional transition is a kind of non-deterministic pattern. In traditional methods, the emotion state transition matrix is the basis for emotional interactive learning [180]. Markov model is often applied in describing dynamic emotions in a continuous dialogue [181]. Morkov hypothesis assumes that the current emotion is only dependent on the previous states, which is similar to context modeling. In [182], a dynamic emotion model is proposed to integrate both initial and stimulating emotions into the Markov model. It combines CNN-LSTM network and MCMC Markov chain Monte Carlo). In their model, emotional transition characteristics are confirmed by emotional sampling sequences, which are generated by the emotional transition probability distribution. In [183], to recognize the user's emotions during real-time chatting, the emotion trajectory of each user is derived from consecutive chatting messages. The emotion trajectory of each user is represented as two separate sequences of valence-time and arousal-time, respectively. In this way, the messages sent by a pair of chatting users can be compared, and the similarity of their emotions can be measured.

Party interaction plays an essential role in TER in dialogue. Self and inter-speaker emotional influence are two prime factors in dynamic emotion modeling. To model the speaker-based emotion, CMN (Conversational Memory Network) [184] incorporated the above two factors. In CMN, the context information of each party is stored in separate memory cells, which realized speaker-specific modeling. Utterance related context is filtered from both memory histories by a multiple hop scheme along with attention mechanism. However, there is no connection between each speaker's memory state in CMN. Later, an interactive conversational memory network (ICON) is proposed in [178], which obtained better experimental results than CMN. In ICON, not only two GRUs are employed for self-party modeling, but also DGIM (Dynamic Global Influence Module) is proposed to model dynamic interaction and maintain a global representation of the conversation. Both CMN and ICON aim to discover the speaker's information for predicting the current utterance. However, they do not go in a meaningful way to model the connection between utterance and the referred speaker. Given an utterance, it is unknown who is the corresponding speaker. To alleviant this limitation, DialogueRNN is proposed in [177]. DialogueRNN models each party and global conversation with three GRUs. The speaker GRU aims to track each party's emotional dynamics and ensures that the model is aware of the corresponding speaker of each utterance. Besides, global GRU and emotion GRU are employed to model context and emotion dynamics throughout the conversation. Three GRUs in DialogurRNN plays a pivotal role in inter-party relation modeling.

5 Conclusion

As emerging research in NLP, TER has a broad application prospect and has attracted widespread attention from researchers. In this paper, we made a systematic review of deep TER technologies. We first introduced some publicly available databases which contain emotion knowledge and are commonly accepted in this field. Then, we investigated the latest progress of TER in recent years and summarized deep TER approaches around the following three aspects: 1) acquiring word embedding with rich semantic and emotional knowledge, 2) integrating emotional knowledge into the deep learning architecture, and 3) training the TER model effectively even with the limited emotional resource. Besides, we conduct further discussions on existing open challenges associate with this research and point out some potential opportunities. We surmised that exploring the associations among existing emotional resources and then conducting data integration can significantly contribute to high-quality emotional datasets. Additionally, using multi-label emotions may more accurately describe emotional states and alleviate the challenge caused by fuzzy emotion boundaries. Referring to the one-sided limitation of textual data, incomplete emotional information may be supplemented by

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Robert Gordon University. Downloaded on May 28,2021 at 02:10:33 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3053275, IEEE Transactions on Affective Computing

interacting with other modalities. These research directions are worthy of further attention. We also discussed TER in dialogue and believe that multi-party emotional interaction, personality modeling, and dynamic emotional tracking can form new research directions. We believe that the latest TER technology systematically reviewed in this article will provide new insights for further research in this field.

Acknowledgment

This work has been partially supported by the Research Clusters program of Tokushima University (Grant No. 2003002), and NSFC-Shenzhen Joint Foundation (Key Project) (Grant No. U1613217).

References

[1] A. Chatterjee, K.N. Narahari, M. Joshi, and P. Agrawal, “SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text,” in *Proc. 13th Int. Workshop on Semantic Eval. (SemEval-2019),* 2019, pp. 39–48.

[2] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," in *Proc. 8th Int. Joint Conf. on Natural Lang. Process.*, Nov. 2017, pp. 986–995.

[3] S.M. Zahiri and J.D. Choi, "Emotion detection on TV show transcripts with sequence-based convolutional neural networks," in *Workshops at the 32th AAAI Conf. on Artif. Intell.*, 2018, pp. 44-51.

[4] K. Sailunaz, M. Dhaliwal, J. Rokne, and R. Alhajj, “Emotion detection from text and speech: a survey,” *Soc. Netw. Anal. Min.*, vol. 8, no. 1, pp.28, 2018.

[5] F. Ren and Y. Wu, “Predicting User-Topic Opinions in Twitter with Social and Topical Context,” *IEEE Trans. Affect. Comput.*, vol.4, no.4, pp. 412-424, 2013.

[6] X. Kang, F. Ren, and Y. Wu, “Exploring Latent Semantic Information for Textual Emotion Recognition in Blog Articles,” *IEEE/CAA Journal of Automtics Sinica,* vol.5, no.1, pp. 204-216, 2018.

[7] S.A. Golder and M.W. Macy, "Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures," *Science*, vol. 333, no. 6051, pp. 1878-1881, Sep. 2011.

[8] S. Kim, J. Lee, G. Lebanon, and H. Park, "Estimating temporal dynamics of human emotions," in *29th AAAI Conf. on Artif. Intell.*, Jan. 2015, pp. 168-174.

[9] M.D. Munezero, C.S. Montero, E. Sutinen, and J. Pajunen, “Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text,” *IEEE Trans. Affect. Comput.*, vol. 5, no. 2, pp. 101–111, Apr. 2014.

[10] A. Yadollahi, A.G. Shahraki, and O.R. Zaiane, “Current state of text sentiment analysis from opinion to emotion mining,” *ACM Comput. Surv.*, vol. 50, no. 2, pp. 1–33, Jan. 2017.

[11] Y. Shi, L. Zhu, W. Li, K. Guo, and Y. Zheng, “Survey on classic and latest textual sentiment analysis articles and techniques,” *Int. J. Inf. Tech. Decis. Mak.*, vol. 18, no. 04, pp.1243-1287, 2019.

[12] R. Liu, Y. Shi, C. Ji, and M. Jia, “A survey of sentiment analysis based on transfer learning”, *IEEE Access*, vol. 7, pp. 85401-85412, 2019.

[13] N. Boudad, R. Faizi, R. O. H. Thami, and R. Chiheb, “Sentiment analysis in Arabic: A review of the literature,” *Ain Shams Eng. J.*, vol.9, no. 4, pp. 2479-2490, 2018.

[14] A. Seyeditabari, N. Tabari, and W. Zadrozny, “Emotion detection in text: a review,” 2018. [Online]. Available: https://arxiv.org/pdf/1806.00674

[15] N.Alswaidan, M. E. B. Menai, "A survey of state-of-the-art approaches for emotion recognition in text." *Knowl. Inf. Syst.*, vol. 62, pp. 1-51, Mar. 2020.

[16] M. Swain, A. Routray, and P. Kabisatpathy, “Databases, features and classifiers for speech emotion recognition: a review,” *Int. J. Speech Technol.*, vol. 21, no. 1, pp. 93-120, 2018.

[17] B. Ko, “A brief review of facial emotion recognition based on visual information,” *Sensors*, vol. 18, no. 2, pp. 401, 2018.

[18] P. V. Rouast, M. Adam, and R. Chiong, “Deep learning for human affect recognition: insights and new developments”, *IEEE Trans. Affect. Comput.*, early access, doi: 10.1109/TAFFC.2018.2890471.

[19] B.G. Martínez, A.M. Rodrigo, R. Alcaraz, and A.F. Caballero, "A review on nonlinear methods using electroencephalographic recordings for emotion recognition," *IEEE Trans. Affect. Comput.*, early access, doi: 10.1109/TAFFC.2018.2890636.

[20] N. J. Shoumy, L. M. Ang, K. P. Seng, D. M. Rahaman, and T. Zia, “Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals,” *J. Netw. Comput. Appl.*, vol. 149, pp.102447, Jan. 2020.

[21] M. Imani, and G. A. Montazer, “A survey of emotion recognition methods with emphasis on E-Learning environments,” *J. Netw. Comput. Appl.*, vol.147, pp. 102423, Dec. 2019.

[22] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy. “Emotion recognition in conversation: Research challenges, datasets, and recent advances,” *IEEE Access*, vol. 7, pp. 100943-100953, Jul.2019.

[23] C. Strapparava, R. Mihalcea, "Affect Detection in Texts 13," *The Oxford handbook of affective computing*, pp. 184-216, 2015.

[24] E. Cambria, “Affective computing and sentiment analysis,” *IEEE Intell. Syst.*, vol. 31, no. 2, pp. 102–107, Mar. 2016.

[25] F. Ren, X. Kang, and C. Quan, “Examining Accumulated Emotional Traits in Suicide Blogs with an Emotion Topic Model,” *IEEE J. Biomed. Health Inform.,* vol.20, no.5, pp. 1384-1396, 2016.

[26] Z. Ahmad, R. Jindal, A. Ekbal, and P. Bhattachharyya, "Borrow from rich cousin: transfer learning for emotion detection using cross lingual embedding," *Expert Syst. Appl.*, vol. 139, pp. 112851, Jan. 2020.

[27] M. Anjaria and R.M.R. Guddeti, “Influence factor based opinion mining of Twitter data using supervised learning,” in *2014 6th Int. Conf. on Communication Systems and Networks (COMSNETS),* IEEE, 2014, pp. 1–8.

[28] F. Ren, Y. Wang, and C. Quan, “A novel factored POMDP model for affective dialogue management,” *J. Intell. Fuzzy Syst.*, vol.31, no.1, pp. 127-136, 2016.

[29] C.S. Montero and J. Suhonen, “Emotion analysis meets learning analytics: online learner profiling beyond numerical data”, *in Proc. 14th Koli Calling Int. Conf. on Comput. Education Research,* ACM, 2014, pp. 165–169.

[30] P. Ekman, "An argument for basic emotions," *Cogn. Emot.*, vol. 6. no. 3-4, pp. 169-200, 1992.

[31] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *Am. Sci.*, vol. 89, no. 4, pp. 344-350, 2001.

[32] C. Strapparava and R. Mihalcea, "Semeval-2007 task 14: Affective text," in *Proc. 4th Int. Workshop on Semantic Eval. (SemEval-2007),* 2007, pp. 70-74.

[33] Z. Wang, S. Li, F. Wu, Q. Sun, and G. Zhou, "Overview of NLPCC 2018 Shared Task 1: Emotion Detection in Code-Switching Text," *CCF Int. Conf. on Natural Lang.Process. and Chinese Comput.*, Springer, Cham, 2018, pp. 429-433.

[34] C.O. Alm, D. Roth, and R. Sproat, "Emotions from text: machine learning for text-based emotion prediction," in *Proc. Conf. on human Lang. Technol. and empirical methods in natural Lang. Process.*, ACL, 2005, pp. 579-586.

[35] S.Y. Chen, C.C. Hsu, C.C. Kuo, and L.W. Ku, "Emotionlines: An emotion corpus of multi-party conversations," 2018. [Online]. Available: https://arxiv.org/pdf/1802.08379

[36] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. 57th Annu. Meeting of the Assoc. for Comput. Linguistics*, 2019, pp. 527–536

[37] C. Strapparava and A. Valitutti, “Wordnet-affect: an affective extension of wordnet,” in *Proc. 4th Int. Conf. Lang. Resour. Eval.*, 2004, vol.4, pp.1083-1086.

[38] J.C.D. Albornoz, L. Plaza, and P. Gervás, “SentiSense: An easily scalable concept-based affective lexicon for Sentiment Analysis,” in *Proc. 8th Int. Conf. Lang. Resour. Eval.*, 2012, pp. 3562-

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Robert Gordon University. Downloaded on May 28,2021 at 02:10:33 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3053275, IEEE Transactions on Affective Computing

3567.

[39] S.M. Mohammad and P.D. Turney, “Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon,” in *Proc. NAACL HLT 2010 workshop on comput. approaches to anal. and gener. of emotion in text.*,2010, pp. 26-34.

[40] J.W. Pennebaker, R.L. Boyd, K. Jordan, and K. Blackburn, ”The development and psychometric properties of LIWC2015”, Austin, TX: University of Texas at Austin, 2015. doi: 10.15781/T29G6Z.

[41] S. Kiritchenko and S.M. Mohammad, “Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems,” in *Proc. 7th Joint Conf. on Lexical and Comput. Semantics*, Jun. 2018, pp. 43-53.

[42] S.M. Mohammad and F. Bravo-Marquez, “WASSA-2017 shared task on emotion intensity,” in *Proc. 8th Workshop on Comput. Approaches to Subjectivity, Sentiment and Social Media Analysis*, Sep. 2017, pp. 34–49.

[43] C. Quan and F. Ren, "A blog emotion corpus for emotional expression analysis in Chinese,” *Comput. Speech Lang.*, vol. 24, no. 4, pp. 726-749, 2010.

[44] C. Quan and F. Ren, "Sentence emotion analysis and recognition based on emotion words using Ren-CECps,” *Int. Journal of Advanced Intell.*, vol. 2, no.1, pp. 105-117, 2010.

[45] S.M. Mohammad, “Word Affect Intensities,” Presented at Proc. 11th Int. Conf. Lang. Resour. Eval. (LREC2018). [Online]. Available: https://www.aclweb.org/anthology/L18-1027.pdf.

[46] O. Araque, L. Gatti, J. Staiano, and M. Guerini, "DepecheMood++: a Bilingual Emotion Lexicon Built Through Simple Yet Powerful Techniques," *IEEE Trans. Affect. Comput.*, early access, doi: 10.1109/TAFFC.2019.2934444.

[47] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament,” *Curr. Psychol.*, vol. 14, no. 4, pp. 261-292, 1996.

[48] S. Arifin and P.Y.K. Cheung, "Affective level video segmentation by utilizing the pleasure-arousal-dominance information,” *IEEE Trans. Multimedia*, vol. 10, no. 7, pp. 1325-1341, 2008.

[49] J.A. Russell, "A circumplex model of affect,” *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161-1178, 1980.

[50] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent,” *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 5-17, 2011.

[51] M.M. Bradley and P.J. Lang, “Affective Norms for English Words, ANEW: Instruction Manual and Affective Ratings,” *Technical report C-1, the center for research in psychophysiology*, University of Florida, vol. 30, no. 1, pp. 25-36, 1999.

[52] S. Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words,” in *Proc. 56th Annu. Meeting of the Assoc. for Comput. Linguistics*, vol.1, 2018.

[53] E. Cambria, Y. Li, F. Z. Xing, S. Poria, and K. Kwok. “SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis,” in *Proc. 29th ACM Int. Conf. on Inf. Knowl. Manage.*, Oct. 2020, pp. 105-114.

[54] Y. Susanto, A. G. Livingstone, B. C. Ng, and E. Cambria. “The hourglass model *revisited,” IEEE Intell. Syst., vol. 35*, no.5, pp.96-102, 2020.

[55] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S.Lee, and S.S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database,” *Lang. Resources and Eval.*, vol. 42, no. 4, pp. 335-359, 2008.

[56] S. Buechel and U. Hahn, “Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis,” in *Proc. 15th Conf. Eur. Chapter of the Assoc. for Comput. Linguistics*, Apr. 2017, vol. 2, pp. 578–585.

[57] X. Li, G.M. Lu, J.J. Yan, and Z.Y. Zhang, “A Survey of Dimensional Emotion Prediction by Multimodal Cues,” *Acta Automaica Sinica*, vol. 44, no. 12, pp. 2142-2159, Dec. 2018.

[58] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J.Deam, "Distributed representations of words and phrases and their compositionality,” *Adv. Neural Inf. Process. Syst.*, pp. 3111-3119, 2013.

[59] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient estimation of word representations in vector space,” 2013. [Online]. Available: https://arxiv.org/pdf/1301.3781.pdf

[60] J. Pennington, R. Socher and C. Manning, "Glove: Global vectors for word representation,” in *Proc. 2014 Conf. on empirical methods in natural Lang. Process. (EMNLP)*, 2014, pp. 1532-1543.

[61] 1 J. Camacho-Collados and M.T. Pilehvar, "From word to sense embeddings: A survey on vector representations of meaning,” *J. Artif. Intell. Res.*, vol. 63, pp. 743-788, 2018.

[62] B. McCann, J. Bradbury, C. Xiong and R.Socher, "Learned in translation: Contextualized word vectors,” *Adv. Neural Inf. Process. Syst.*, pp. 6294-6305, 2017.

[63] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations,” in *Proc. 2018 Conf. of the North Amer. Chapter of the Assoc. for Comput. Linguistics: Human Lang. Technol.*, Vol. 1, Jun. 2018, pp. 2227–2237.

[64] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, "Improving language understanding by generative pre-training,” 2018. [Online]. Available: https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf

[65] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proc. of NAACL-HLT 2019*, Jun. 2019, pp. 4171–4186.

[66] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners,” OpenAI Blog, vol. 1, no. 8, 2019. [Online]. Available: http://www.persagen.com/files/misc/radford2019language.pdf

[67] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, and S. Agarwal. “Language models are few-shot learners,” 2020. [Online]. Available: https://arxiv.org/pdf/2005.14165

[68] Z. Dai, Z. Yang, Y. Yang, W.W. Cohen, J. Carbonell Q.V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context,” 2019. [Online]. Available: https://arxiv.org/abs/1901.02860

[69] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q.V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding,” Presented at 33rd Conf. on Neural Inf. Process. Syst. (NeurIPS 2019). [Online]. Available: https://arxiv.org/abs/1906.08237

[70] K. Song, X. Tan, T. Qin, J. Lu, and T.Y. Liu "Mass: Masked sequence to sequence pre-training for language generation,” 2019. [Online]. Available: https://arxiv.org/pdf/1905.02450

[71] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H.W. Hon, "Unified Language Model Pre-training for Natural Language Understanding and Generation,” presented at *33rd Conf. on Neural Inf. Process. Syst. (NeurIPS 2019)*. [Online]. Available: https://arxiv.org/abs/1905.03197

[72] P. Xu, A. Madotto, C.S. Wu, J.H. Park, and P. Fung "Emo2vec: Learning generalized emotion representation by multi-task training,” in *Proc. 9th Workshop on Comput. Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2018, pp. 292–298.

[73] B. Eisner, T. Rocktäschel, I. Augenstein, M. Bošnjak, and S. Riedel, "emoji2vec: Learning emoji representations from their description,” in *Proc. 4th Int. Workshop on Natural Lang. Process. for Social Media at EMNLP 2016*, Nov. 2016, pp. 48-54.

[74] 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. 2017 Conf. on Empirical Methods in Natural Lang. Process.*, Sep. 2017 pp.1615–1625.

[75] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for twitter sentiment classification,” in *Proc. 52nd Annu. Meeting of the Assoc. for Comput. Linguistics*, 2014, vol. 1, pp. 1555–1565.

[76] B. Shi, Z. Fu, L. Bing, and W. Lam, "Learning domain-sensitive and sentiment-aware word embeddings,” in *Proc. 56th Annu. Meeting of the Assoc. for Comput. Linguistics*, Jul. 2018, pp. 2494–2504.

[77] G.I. Winata, A. Madotto, Z. Lin, J. Shin, Y. Xu, P. Xu, and P.

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Robert Gordon University. Downloaded on May 28,2021 at 02:10:33 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3053275, IEEE Transactions on Affective Computing

Fung, "CAiRE_HKUST at SemEval-2019 Task 3: Hierarchical Attention for Dialogue Emotion Classification," in *Proc. 13th Int. Workshop on Semantic Eval. (SemEval-2019)*, Jun. 2019, pp. 142-147.

[78] D. Shen, G. Wang, W. Wang, M.R. Min, Q. Su, Y. Zhang, C. Li, R. Henao, and L. Carin, "Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms," in *Proc. 56th Annu. Meeting of the Assoc. for Comput. Linguistics*, 2018, pp. 440-450.

[79] S. Poria, E. Cambria, D. Hazarika, and N. Majumder, "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. Meet. of the Assoc. for Comput. Linguist.*, 2017, vol. 1, pp. 873-883.

[80] D. Tang, B. Qin, and T. Liu, "Aspect level sentiment classification with deep memory network," in *Proc. 2016 Conf. on Empirical Methods in Natural Lang. Process.*, Nov. 2016, pp. 214–224.

[81] J.A. Balazs, E. Marrese-Taylor, Y. Matsuo, "IIIDYT at IEST 2018: Implicit Emotion Classification With Deep Contextualized Word Representations," in *Proc. 9th Workshop on Comput. Approaches to Subjectivity, Sentiment and Social Media Analysis*, Oct. 2018, pp. 50–56.

[82] M.A. Mageed and L. Ungar, "Emonet: Fine-grained emotion detection with gated recurrent neural networks," in *Proc. 55th Annu. Meeting of the Assoc. for Comput. Linguistics*, 2017, vol. 1, pp. 718–728.

[83] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. 2014 Conf. on EMNLP*, Oct. 2014, pp. 1746–1751.

[84] T. Rao, X. Li, H. Zhang, and M. Xu, "Multi-level region-based Convolutional Neural Network for image emotion classification," *Neurocomputing*, vol. 333, no. 14, pp. 429-439, Mar. 2019.

[85] A Vaswani, N Shazeer, N Parmar, J. Uszkoreit , L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, pp. 5998-6008, 2017.

[86] K. Kowsari, D.E. Brown, M. Heidarysafa, K.J. Meimandi, M.S. Gerber, and L.E. Barnes "Hdltex: Hierarchical deep learning for text classification," in *Proc. 2017 16th IEEE Int. Conf. on Machine Learning and Applications (ICMLA)*, 2017, pp. 364-371.

[87] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. 2016 Conf. of the North Amer. Chapter of the Assoc. for Comput. Linguistics: Human Lang. Technol.*, 2016, pp. 1480–1489.

[88] K.S. Tai, R. Socher, and C.D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," in *Proc. 53rd Annu. Meeting of the Assoc. for Comput. Linguistics and the 7th Int. Joint Conf. on Natural Lang. Process.*, Jul. 2015, pp. 1556–1566.

[89] C. Chen, R. Zhuo, and J. Ren, "Gated recurrent neural network with sentimental relations for sentiment classification," *Inf. Sci.*, vol. 502, pp. 268-278, 2019.

[90] W. Ying, R. Xiang, and Q. Lu, "Improving Multi-label Emotion Classification by Integrating both General and Domain Knowledge," in *Proc. 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, 2019, pp. 316–321.

[91] H. Meisheri and L. Dey, "TCS Research at SemEval-2018 Task 1: Learning Robust Representations using Multi-Attention Architecture," in *Proc. 12th Int. Workshop on Semantic Eval. (SemEval-2018)*, 2018, pp. 291-299.

[92] P. Goel, D. Kulshreshtha, P. Jain, and K.K. Shukla, "Prayas at emoint 2017: An ensemble of deep neural architectures for emotion intensity prediction in tweets," in *Proc. 8th Workshop on Comput. Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2017, pp. 58-65.

[93] I. Perikos and I. Hatzilygeroudis, "Recognizing emotions in text using ensemble of classifiers," *Eng. Appl. Artif. Intell.*, vol. 51, pp. 191–201, May 2016.

[94] S. Corchs, E. Fersini and F. Gasparini, "Ensemble learning on visual and textual data for social image emotion classification," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 8, pp. 2057-2070, 2019.

[95] E. Fersini, E. Messina, and F.A. Pozzi, "Sentiment analysis: Bayesian ensemble learning," *Decis. Support Syst.*, vol. 68, pp. 26–38, Dec. 2014.

[96] F. Ren, and J. Deng. "Background knowledge based multi-stream neural network for text classification," *Applied Sciences*, vol. 8, no. 12, pp. 2472, 2018.

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

[98] M.S. Akhtar, D. Ghosal, and A. Ekbal, "A Multi-task Ensemble Framework for Emotion, Sentiment and Intensity Prediction," 2018. [Online]. Available: https://arxiv.org/pdf/1808.01216

[99] P. Agrawal and A. Suri, "NELEC at SemEval-2019 Task 3: Think Twice Before Going Deep," in *Proc. 13th Int. Workshop on Semantic Eval. (SemEval -2019)*, Jun. 2019, pp. 266-271.

[100] P. Zhong, D. Wang, and C. Miao, "Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations," in *Proc. 2019 Conf. on Empirical Methods in Natural Lang. Process. and the 9th Int. Joint Conf. on Natural Lang. Process.*, Nov. 2019, pp. 165–176.

[101] A. Kumar, D. Kawahara, and S. Kurohashi, "Knowledge-enriched two-layered attention network for sentiment analysis," in *Proc. 2018 Conf. of the North Amer. Chapter of the Assoc. for Comput. Linguistics: Human Lang. Technol.*, Jun. 2018, pp. 253–258.

[102] C Baziotis, N Athanasiou, A Chronopoulou, A. Kolovou, G. Paraskevopoulos, N. Ellinas, S. Narayanan, and A. Potamianos, "Ntua-slp at semeval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning," in *Proc. 12th Int. Workshop on Semantic Eval. (SemEval-2018)*, Jun. 2018, pp. 245–255.

[103] A Chatterjee, U Gupta, MK Chinnakotla, R. Srikanth, M. Galley, and P. Agrawal, "Understanding emotions in text using deep learning and big data," *Comput. Human Behav.*, vol. 93, pp. 309-317, Apr. 2019.

[104] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efcient text classification," in *Proc. 15th Conf. of the Eur. Chapter of the Assoc. for Comput. Linguistics*, 2017, pp. 427–431.

[105] H. Wang, "ReNN: Rule-embedded Neural Networks," in *Proc. 24th Int. Conf. on Pattern Recognit. (ICPR)*. IEEE, 2018, pp. 824-829.

[106] Q. Xie, X. Ma, Z. Dai and E. Hovy, "An interpretable knowledge transfer model for knowledge base completion," in *Proc. 55th Annu. Meeting of the Assoc. for Comput. Linguistics*, Jul. 2017, pp. 950–962.

[107] K. Vo, D. Pham, M. Nguyen, T. Mai, and T. Quan, "Combination of domain knowledge and deep learning for sentiment analysis," in *Int. Workshop on Multi-disciplinary Trends in Artif. Intell.*, 2017, pp. 162-173.

[108] K. Vo, T. Nguyen, D. Pham, M. Nguyen, M. Truong, T. Mai, and T. Quan, "Combination of Domain Knowledge and Deep Learning for Sentiment Analysis of Short and Informal Messages on Social Media," 2019. [Online]. Available: https://arxiv.org/abs/1902.06050

[109] Q. Qian, M. Huang, J. Lei, and X. Zhu, "Linguistically regularized lstms for sentiment classification," in *Proc. 55th Annu. Meeting of the Assoc. for Comput. Linguistics*, Jul. 2017, pp. 1679–1689.

[110] R. Gupta, S. Sahu, C. Espy-Wilson, "Semi-supervised and transfer learning approaches for low resource sentiment classification," in *2018 IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, Jun. 2018, pp. 5109-5113.

[111] S.J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345-1359, 2009.

[112] G Wiedemann, E Ruppert, R Jindal, and C. Biemann, "Transfer learning from lda to bilstm-cnn for offensive language detection in twitter," 2018. [Online]. Available: https://arxiv.org/abs/1811.02906

[113] B. Kratzwald, S. Ilić, M. Kraus, S. Feuerriegel, and H. Prendinger, "Deep learning for affective computing: Text-based emotion recognition in decision support," *Decis. Support Syst.*, vol. 115, no. 2018, pp. 24-35, 2018.

[114] A. Chronopoulou, A. Margatina, C. Baziotis, and A. Potamianos, "NTUA-SLP at IEST 2018: Ensemble of neural transfer methods for implicit emotion classification," Presented at *Proc. 9th Workshop on Comput. Approaches to Subjectivity, Sentiment and Social Media Analysis*, Jan. 2018. [Online]. Available:

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Robert Gordon University. Downloaded on May 28,2021 at 02:10:33 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3053275, IEEE Transactions on Affective Computing

https://arxiv.org/abs/1809.00717

[115] J. Yu, L. Marujo, J. Jiang, P. Karuturi, and W. Brendel "Improving multi-label emotion classification via sentiment classification with dual attention transfer network," *in Proc. 2018 Conf. on Empirical Methods in Natural Lang. Process. ACL,* 2018, pp. 1097-1102.

[116] S.L. Smith, D.H. Turban, S. Hamblin, and N.Y. Hammerla, "Offline bilingual word vectors, orthogonal transformations and the inverted softmax," Presented at *5th Int. Conf. on Learning Representations (ICLR)*, Apr. 2017. [Online]. Available: https://arxiv.org/abs/1702.03859

[117] M. Yang, W. Yin, Q. Qu, W. Tu, Y. Shen, and X. Chen, "Neural Attentive Network for Cross-Domain Aspect-level Sentiment Classification," *IEEE Trans. Affect. Comput.,* early access, Feb. 2019, doi: 10.1109/TAFFC.2019.2897093.

[118] J. Gideon, M. McInnis, and E.M. Provost, "Improving Cross-Corpus Speech Emotion Recognition with Adversarial Discriminative Domain Generalization (ADDoG)," *IEEE Trans. Affect. Comput.,* early access, May 2019, doi: 10.1109/TAFFC.2019.2916092.

[119] Y Zhang, N Zhang, L Si, Y Lu, Q Wang, and X. Yuan, "Cross-Domain and Cross-Category Emotion Tagging for Comments of Online News," in *Proc. 37th Int. ACM SIGIR Conf. on Research & development in Inf. retrieval. ACM*, 2014, pp. 627-636.

[120] S. Ruder, "An overview of multi-task learning in deep neural networks," 2017. [Online]. Available: https://arxiv.org/abs/1706.05098

[121] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning." in *Proc. 25th Int. Conf. on Machine learning, ACM,* 2008, pp. 160–167.

[122] B. McCann, N.S. Keskar, C. Xiong, and R. Socher, "The natural language decathlon: Multitask learning as question answering," 2018. [Online]. Available: https://arxiv.org/pdf/1806.08730.pdf

[123] G. Balikas, S. Moura, and M.R. Amini, "Multitask Learning for Fine-Grained Twitter Sentiment Analysis," in *Proc. 40th Int. ACM SIGIR Conf. on Research and Development in Inf. Retrieval (SIGIR '17),* Aug. 2017, pp. 1005-1008.

[124] X Wang, M Peng, L Pan, M Hu, C Jin, F Ren, "Two-level attention with two-stage multi-task learning for facial emotion recognition," 2018. [Online]. Available: *https://arxiv.org/abs/1811.12139*

[125] W. Gao, S. Li, S.Y.M. Lee, G. Zhou, and C.R. Huang, "Joint learning on sentiment and emotion classification," in *Proc. 22nd ACM Int. Conf. on Inf. & Knowl. Manage. ACM*, 2013, pp. 1505-1508.

[126] D. Wu and J. Huang, "Affect estimation in 3D space using multi-task active learning for regression," *IEEE Trans. Affect. Comput.*, early access, May 2019, doi: 10.1109/TAFFC.2019.2916040.

[127] C. Huang, A. Trabelsi, O.R. Zaïane, "ANA at SemEval-2019 Task 3: Contextual Emotion detection in Conversations through hierarchical LSTMs and BERT," in *Proc. 13th Int. Workshop on Semantic Eval. (SemEval-2019)*, Jun. 2019, pp. 49–53.

[128] A. Basile, M.F. Salvador, N. Pawar, S. Stajner, M. C. Rios, and Y. Benajiba, "SymantoResearch at SemEval-2019 Task 3: Combined Neural Models for Emotion Classification in Human-Chatbot Conversations," in *Proc. 13th Int. Workshop on Semantic Eval. (SemEval-2019)*, 2019, pp. 330–334.

[129] 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. Twenty-Seventh Int. Joint Conf. on Artif. Intell. (IJCAI 18)*, 2018, pp. 4595-4601.

[130] S. Zhu, S. Li, and G. Zhou, "Adversarial Attention Modeling for Multi-dimensional Emotion Regression," in *Proc. 57th Annu. Meeting of the Assoc. for Comput. Linguistics*, 2019, pp. 471-480.

[131] I. Augenstein, S. Ruder, and A. Søgaard "Multi-task learning of pairwise sequence classification tasks over disparate label spaces," in *Proc. 2018 Conf. of the North Amer. Chapter of the Assoc. for Comput. Linguistics: Human Lang. Technol.*, 2018, pp. 1896–1906.

[132] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Prog. in Artif. Intell.*, vol. 5, no. 4, pp. 221-232, Apr. 2016.

[133] M. Buda, A. Maki, and M.A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Netw.*, vol. 106, pp. 249-259, 2018.

[134] X. Zhu, Y. Liu, J. Li, T. Wan, and Z. Qin, "Emotion classification with data augmentation using generative adversarial networks," in *Pacific-Asia Conf. on Knowl. Discovery and Data Mining. Springer, Cham*, 2018, pp. 349-360.

[135] J.Y. Zhu, T. Park, P. Isola, and A.A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. on Comput. Vis.*, 2017, pp. 2223-2232.

[136] J.W. Wei and K. Zou, "Eda: Easy data augmentation techniques for boosting performance on text classification tasks," in *Proc. 2019 Conf. on Empirical Methods in Natural Lang. Process. and the 9th Int. Joint Conf. on Natural Lang. Process.*, 2019, pp. 6382–6388.

[137] N.V. Chawla, "Data mining for imbalanced datasets: An overview," *Data mining and Knowl. discovery handbook. Springer, Boston, MA*, 2009, pp. 875-886.

[138] J. S. Y. Liew, H. R. Turtle, and E. D. Liddy, "EmoTweet-28: A FineGrained Emotion Corpus for Sentiment Analysis," in *Proc. Tenth Int. Conf. on Lang. Resources and Eval. (LREC 2016),* 2016, pp. 1149-1156.

[139] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "Semeval-2018 task 1: Affect in tweets," in *Proc. 12th Int. Workshop on Semantic Eval.*, pp.1-17, Jun. 2018.

[140] F. Ren and K. Matsumoto, "Semi-Automatic Creation of Youth Slang Corpus and Its Application to Affective Computing", *IEEE Trans. Affect. Comput.*, vol.7, no.2, pp. 176-189, 2016.

[141] S.M. Mohammad, "# Emotional tweets," in *Proc. First Joint Conf. on Lexical and Comput. Semantics, ACL,* 2012, pp. 246-255.

[142] K. Fort and B. Sagot, "Influence of pre-annotation on POS-tagged corpus development," in *Proc. fourth linguistic annotation workshop, ACL,* 2010, pp. 56-63.

[143] T. Lingren, L. Deleger, K. Molnar, H. Zhai, J.M. Derr, M. Kaiser, L. Stoutenborough, Q. Li, and I. Solti, "Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements," *J. Am. Med. Inform. Assoc.*, vol. 21, no. 3, pp. 406-413, 2013.

[144] L. Canales, W. Daelemans, E. Boldrini, and P.M. Barco, "EmoLabel: Semi-Automatic Methodology for Emotion Annotation of Social Media Text," *IEEE Trans. Affect. Comput.*, early access, Jul. 2019, doi: 10.1109/TAFFC.2019.2927564.

[145] J. Deng, and F. Ren. "Multi-label Emotion Detection via Emotion-Specified Feature Extraction and Emotion Correlation Learning," *IEEE Trans. Affect. Comput.*, early access, Oct. 2020, doi: 10.1109/TAFFC.2020.3034215.

[146] S.M. Liu, and J.H. Chen, "A multi-label classification based approach for sentiment classification," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1083-1093, Feb. 2015.

[147] L. Buitinck, J.V. Amerongen, E. Tan , and M.D. Rijke, "Multi-emotion detection in user-generated reviews," in *Eur. Conf. on Inf. Retrieval*, Mar. 2015, pp. 43-48.

[148] G. Tsoumakas and I. Vlahavas, "Random k-labelsets: An ensemble method for multilabel classification," in *Eur. Conf. on machine learning (ECML)*, 2007, pp. 406–417.

[149] M. Jabreel and A. Moreno, "A Deep Learning-Based Approach for Multi-Label Emotion Classification in Tweets," *Appl. Sci.*, vol. 9, no. 6, pp. 1123, 2019.

[150] Y. Kim, H. Lee, and K. Jung "AttnConvnet at SemEval-2018 Task 1: attention-based convolutional neural networks for multi-label emotion classification," in *Proc. 12th Int. Workshop on Semantic Eval.*, Jun. 2018, pp. 141–145.

[151] N. Colneriĉ, and J. Demsar, "Emotion Recognition on Twitter: Comparative Study and Training a Unison Model," *IEEE Trans. Affect. Comput.*, pp. 1949-3045, Feb. 2018.

[152] D. Pan, and J. Nie, "Mutux at SemEval-2018 Task 1: Exploring Impacts of Context Information On Emotion Detection," in *Proc. 12th Int. Workshop on Semant. Eval.*, Jun. 2018, pp. 345-349.

[153] H. Huihui, and R. Xia, "Joint Binary Neural Network for Multi-label Learning with Applications to Emotion Classification," in *Proc. Nat. Lang. Process. and Chin. Compu. (NLPCC 2018)*, Aug. 2018, pp. 250-259.

[154] S. Lian, J. Liu, R. Lu, and X. Luo, "Captured multi-label relations via joint deep supervised autoencod-

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Robert Gordon University. Downloaded on May 28,2021 at 02:10:33 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3053275, IEEE Transactions on Affective Computing

er," *Appl. Soft Comput.*, vol. 74, pp. 709-728, Jan. 2019.

[155] D.A. Phan, Y. Matsumoto, and H. Shindo, "Autoencoder for Semisupervised Multiple Emotion Detection of Conversation Transcripts," *IEEE Trans. Affect. Comput.*, early access, Dec. 2018, doi: 10.1109/TAFFC. 2018.2885304.

[156] M.L. Zhang, and Z.H. Zhou, "Multi-label neural networks with applications to functional genomics and text categorization," *IEEE Trans. Knowl. Data Eng.*, vol. 18, pp. 1338–1351, 2006.

[157] C.K. Yeh, W.C. Wu, W.J. Ko, and Y.C.F. Wang, "Learning deep latent space for multi-label classification," in *Proc. 31th AAAI Conf. on Artif. Intell.*, 2017, pp. 2838–2844.

[158] K. Wang, M. Yang, and W. Yang, "Deep Correlation Structure Preserved Label Space Embedding for Multi-label Classification," in *Proc. Asian Conf. on Mach. Learn.*, 2018, vol. 95, pp.1-16.

[159] Z.F. He, M. Yang, Y. Gao, H.D. Liu, and Y. Yin, "Joint multi-label classification and label correlations with missing labels and feature selection," *Knowl. Based Syst.*, vol. 163, pp. 145-158, Jan. 2019.

[160] M. Rei, and A. Søgaard, "Jointly Learning to Label Sentences and Tokens," in *Proc. 33th AAAI Conf. on Artif. Intell. (AAAI-19)*, Jul. 2019, pp. 6916-6923.

[161] D. Zhou, Y. Yang, and Y. He, "Relevant emotion ranking from text constrained with emotion relationships," in *Proc. 2018 Conf. of the North Am. Chapter of the Assoc. for Comput. Linguist.: Hum. Lang. Technol.*, vol. 1, pp. 561-571, Jun. 2018.

[162] P. Yang, X. Sun, W. Li, S. Ma, W. Wu, and H. Wang, "SGM: sequence generation model for multi-label classification," in *Proc. 27th Int. Conf. on Comput. Linguistics*, 2018, pp. 3915–3926.

[163] W. Liao, Y. Wang, Y. Yin, X. Zhang, and P. Ma, "Improved Sequence Generation Model for Multi-label Classification via CNN and Initialized Fully Connection," *Neurocomputing*, vol. 32, no. 21, pp. 188-195, Dec. 2019.

[164] M.L. Zhang, Y.K. Li, X.Y. Liu, and X. Geng, "Binary relevance for multi-label learning: an overview," *Front. Compt. Sci.*, vol. 12, no. 2, pp. 191-202, 2018.

[165] B.W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90-99, 2018.

[166] T. Xu, Y. Zhou, Z. Wang, and Y. Peng, "Learning Emotions EEG-based Recognition and Brain Activity: A Survey Study on BCI for Intelligent Tutoring System," *Procedia computer science*, vol. 130, pp. 376-382, 2018.

[167] T. Baltrušaitis, C. Ahuja, and L.P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423-443, 2018.

[168] G. Xu, W. Li, and J. Liu, "A social emotion classification approach using multi-model fusion," *Future Gener. Comp. Syst.*, vol. 102, pp. 347-356, 2020.

[169] L.C. Quiros, E. Gedik, and H. Hung, "Multimodal self-assessed personality estimation during crowded mingle scenarios using wearables devices and cameras," *IEEE Trans. Affect. Comput.*, early access, Jul. 2019, doi: 10.1109/TAFFC.2019.2930605.

[170] Z. Li, X. Wu, X. Xu, H. Wang, Z. Guo, Z. Zhan, and L. Yao, "The Recognition of Multiple Anxiety Levels Based on Electroencephalograph," *IEEE Trans. Affect. Comput.*, early access, Aug. 2019, doi: 10.1109/TAFFC.2019.2936198.

[171] H. Huang, Q. Xie, J. Pan, Y. He, Z. Wen, R. Yu and Y. Li, "An EEG-Based Brain Computer Interface for Emotion Recognition and Its Application in Patients with Disorder of Consciousness," *IEEE Trans. Affect. Comput.*, early access, 2019, doi: 10.1109/TAFFC.2019.2901456.

[172] W.L. Zheng, J.Y. Zhu, and B.L. Lu "Identifying stable patterns over time for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 417 – 429, 2019.

[173] Y. Li, W. Zheng, L. Wang, Y. Zong and Z. Cui, "From Regional to Global Brain: A Novel Hierarchical Spatial-Temporal Neural Network Model for EEG Emotion Recognition," *IEEE Trans. Affect. Comput.*, early access, Jun. 2019, doi: 10.1109/TAFFC.2019.2922912.

[174] J. Bhaskar, K. Sruthi, and P. Nedungadi "Hybrid approach for emotion classification of audio conversation based on text and speech mining," *Procedia Comput. Sci.*, vol. 46, pp. 635-643, 2015.

[175] A. Sordoni, Y. Bengio, H. Vahabi, C. Lioma, and J.G. Simonsen, "A hierarchical recurrent encoder-decoder for generative context-aware query suggestion," in *Proc. 24th ACM Int. on Conf. on Inf. and Knowl. Manage.*, ACM, 2015, pp. 553–562.

[176] B. Sanghwan, J. Choi, and S.G. Lee, "SNU_IDS at SemEval-2019 Task 3: Addressing Training-Test Class Distribution Mismatch in Conversational Classification," in *Proc. 13th Int. Workshop on Semantic Eval. (SemEval-2019)*, 2019, pp. 312-317.

[177] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," in *Proc. AAAI Conf. on Artif. Intell.*, vol. 33. 2019, pp. 6818-6825.

[178] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "Icon: Interactive conversational memory network for multimodal emotion detection," in *Proc. 2018 Conf. on Empirical Methods in Natural Lang. Process.*, 2018, pp. 2594–2604.

[179] J. Donahue and L.A. Hendricks, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. on comput. vision and pattern recognit.*, 2015, pp. 2625-2634.

[180] L. Yang, H.F. Lin, and W. Guo, "Text-based emotion transformation analysis," *Comput. Eng. Sci.*, vol. 33 no. 9, pp. 123–129, 2011.

[181] T.T. Teoh and S.Y. Cho, "Human emotional states modeling by Hidden Markov Model," in *Proc. 7th Int. Conf. on Natural Comput.*, Jul. 2011, doi: 10.1109/ICNC.2011.6022189

[182] X. Sun, C. Zhang and L. Li, "Dynamic emotion modelling and anomaly detection in conversation based on emotional transition tensor," *Inf. Fusion*, vol. 46, pp. 11-22, 2019.

[183] C.H. Chen, W.P. Lee, and J.Y. Huang, "Tracking and recognizing emotions in short text messages from online chatting services," *Inf. Process. Manage.*, vol. 54, no.6, pp. 1325-1344, 2018.

[184] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proc. 2018 Conf. of the North Amer. Chapter of the Assoc. for Comput. Linguistics: Human Lang. Technol., (NAACL)*, 2018, pp. 2122-2132.

Jiawen Deng received a double degree of master in Advanced Technology and Science from Tokushima University, Japan, and Mechanical Engineering from Nantong University, China. She currently is a Ph.D. student at Tokushima University. Her research interests include Natural Language Processing, Artificial Intelligence, and Affective Computing.

Fuji Ren received his Ph.D. degree in 1991 from the Faculty of Engineering, Hokkaido University, Japan. From 1991 to1994, he worked at CSK as a chief researcher. In 1994, he joined the Faculty of Information Sciences, Hiroshima City University, as an Associate Professor. Since 2001, he has been a Professor of the Faculty of Engineering, Tokushima University. His current research interests include Natural Language Processing, Artificial Intelligence, Affective Computing, Emotional Robot. He is the Academician of The Engineering Academy of Japan and EU Academy of Sciences. He is a senior member of IEEE, Editor-in-Chief of International Journal of Advanced Intelligence, a vice president of CAAI, and a Fellow of The Japan Federation of Engineering Societies, a Fellow of IEICE, a Fellow of CAAI. He is the President of International Advanced Information Institute, Japan.

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Robert Gordon University. Downloaded on May 28,2021 at 02:10:33 UTC from IEEE Xplore. Restrictions apply.