IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 September 2023, accepted 28 September 2023, date of publication 6 October 2023, date of current version 11 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3322436

RESEARCH ARTICLE

Toward Transformer Fusions for Chinese Sentiment Intensity Prediction in Valence-Arousal Dimensions

YU-CHIH DENG[1], YIH-RU WANG[1], SIN-HORNG CHEN[1], AND LUNG-HAO LEE[2], (Member, IEEE)
[1]Department of Electrical Engineering, National Yang Ming Chiao Tung University, Hsinchu 300093, Taiwan
[2]Department of Electrical Engineering, National Central University, Taoyuan 320317, Taiwan

Corresponding author: Lung-Hao Lee (lhlee@ee.ncu.edu.tw)

This work was supported in part by the ASUS inc., with Taiwan Computing Cloud (TWCC) service; and in part by the National Science and Technology Council, Taiwan, under MOST 111-2628-E-008-005-MY3.

ABSTRACT BERT (Bidirectional Encoder Representations from Transformers) uses an encoder architecture with an attention mechanism to construct a transformer-based neural network. In this study, we develop a Chinese word-level BERT to learn contextual language representations and propose a transformer fusion framework for Chinese sentiment intensity prediction in the valence-arousal dimensions. Experimental results on the Chinese EmoBank indicate that our transformer-based fusion model outperforms other neural-network-based, regression-based and lexicon-based methods, reflecting the effectiveness of integrating semantic representations in different degrees of linguistic granularity. Our proposed transformer fusion framework is also simple and easy to fine-tune over different downstream tasks.

INDEX TERMS Transformer fusion, Chinese word-level BERT, pre-trained language models, dimensional sentiment analysis, affective computing.

I. INTRODUCTION

Sentiment analysis involving the use of natural language processing and computational linguistics to automatically identify affective information from texts has emerged as a leading technique for emotional AI applications [1], [2], [3], [4], [5], [6]. Representation of affective states is an essential issue in sentiment analysis and can be generally divided into category-based and dimension-based approaches. Category-based approaches represent affective states as several predefined discrete classes, such as positive, negative and neutral. Dimension-based approaches represent affective states as continuous numerical values, called intensity, in multiple dimensions to provide more fine-grained emotional information [7], [8], [9].

Figure 1 shows the two-dimensional valence-arousal (VA) space. Valence expresses the degree of pleasant and unpleasant (i.e., positive and negative) feelings, while arousal expresses the degree of excitement and calmness. Based on this representation, any affective expression can be mapped into the VA coordinate plane as a point by recognizing their valence-arousal ratings. For example, an affective word "無價" (priceless), with human-annotated VA ratings 6.2 and 4.8 in the Chinese EmoBank corpus [9], is located in the low-arousal and high-valence quadrant. A single sentence "是一個地方的無 價 之 寶" (It's a priceless treasure from somewhere) contains this affective word as a modifier to express an object with high value, with respective VA ratings of 7.5 and 6.5. A multi-word phrase "極為痛苦" (extremely painful) has a degree adverb to modify the affective word to express a negative-arousal and high-arousal feeling (with VA ratings of 1.65 and 7.993). The multi-sentence text "難以忍受、醫療無法治癒的身心痛苦" (physical and mental pains that are unbearable and medically incurable) contains multiple affective words to reflect negative complicated perceptions (valence 2.889 and arousal 4.286).

In general, sentiment intensity prediction methods can be summarized into four types: lexicon-based [4], [10], [11], [12], [13], [14], [15], [16], regression-based [17], [18], [19], [20], [21], [22], neural-network-based [23], [24], [25], [26],

The associate editor coordinating the review of this manuscript and approving it for publication was Alessandro Floris.

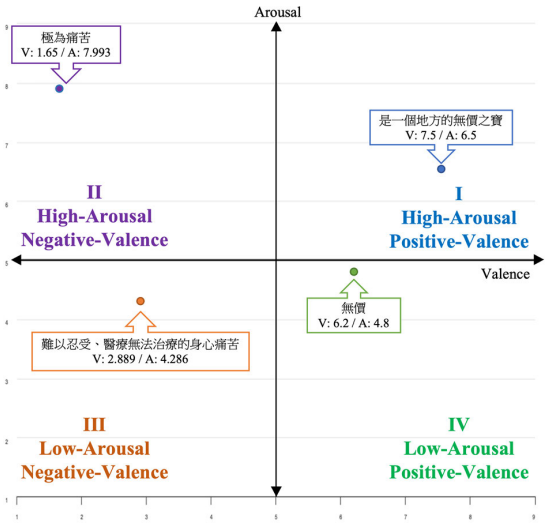

FIGURE 1. Two-Dimensional valence-arousal space. Based on this representation, any affective expression can be mapped into the VA coordinate plane as a point by recognizing their valence-arousal ratings.

[27], [28], [29], [30], [31], [32], [33], and transformer-based [34], [35], [36], [37], [38], [39], [40], [41], [42], [43]. The lexicon-based approaches provide baseline results for reference, while transformer-based models have usually achieved promising results in the valence-arousal dimensions [9]. Recently, BERT-like neural networks have provided state-of-the-art results in a wide variety of natural language processing tasks. BERT (Bidirectional Encoder Representations from Transformers) [44] uses an encoder architecture with an attention mechanism to construct a transformer-based neural network consisting of two main steps: 1) pre-training in which model is trained on unlabeled data over different pre-training tasks; and 2) fine-tuning where the BERT model is first initialized with the pre-trained parameters and then fine-tuned using data from the downstream tasks. Following the pre-training/fine-tuning fashions, several pre-trained language models (PLM) have been publicly released for parameter fine-tuning over different downstream tasks [45].

However, existing Chinese PLMs were mainly trained on character-based sequences due to two main limitations. The first limitation is the need for word segmentation preprocessing due to a lack of delimiters between Chinese characters, and incorrectly segmenting word boundaries will cause error propagation, affecting the language representation in different contexts [46]. Nevertheless, word semantics can be exploited to enrich the character representation of Chinese PLM [47]. Taking the sentence "每次看到梵谷兄弟相關的文章就很感動" (I am very touched every time when I read articles related to the Van Gogh brothers) from the Chinese EmoBank corpus [9] as an example, this sentence can be correctly segmented as 每(every)"次(time) 看到(read) 梵谷(Van Gogh) 兄弟(brothers) 相關(related) 的(pronounced as De) 文章(articles) 就(then) 很(very) 感動(touched)". After word segmentation [48], we can find the affective word "感動" (touched) is modified by a degree adverb "很" (very) to express a positive-valence and high-arousal feelings. This is a helpful clue to predict the sentiment intensity of this sentence with VA ratings of 7.0 and 6.75. The second limitation is the need for huge pre-training data sets. Because Chinese words usually contain multiple characters, more data is needed to sufficiently reflect contexts for training a word-level Chinese PLM. For example, the above-mentioned sentence has 17 tokens in terms of characters, but only 11 tokens in terms of words. This shows the need for greater amounts of data to train a word-level Chinese PLM as opposed to character-level.

Recently, fusion-based methods have been used for categorical sentiment analysis [49], [50], [51] to classify sentiments as predefined discrete classes on multimedia or multimodal targets. To our best knowledge, there is no transformer-based fusion model for dimensional sentiment analysis to identify the intensity in continuous numerical values, especially for Chinese texts. We are thus motivated to develop a Chinese word-level BERT to address the above limitations for latent language representations and propose a transformer fusion framework based on different linguistic granularities for Chinese sentiment intensity prediction in the valence-arousal dimensions. The main contributions are summarized as follows:

(1) We develop a Chinese word-level BERT for contextual language representation.

We use the NCTU word segmentation tool [48] to process collected text corpora, with a total of 2.8 billion words. We pre-train a Chinese word-level BERT model () [44] over the same Masked Language Model (MLM) task based on a dynamic masking strategy [52]. We plan to release our word-level BERT as a pre-trained language model for further research.

(2) We explore transformer fusion methods for Chinese sentiment intensity prediction.

We propose a transformer fusion framework to integrate word-level and character-level transformers for Chinese sentiment intensity prediction in the valence-arousal dimensions. Chinese Valence-Arousal Sentences (CVAS) and Chinese Valence-Arousal Texts (CVAT) from the Chinese EmoBank corpus [9] were used to evaluate performance. In experiments, our proposed fusion model outperformed other neural-network-based, regression-based, and lexicon-based models, confirming the effectiveness of our transformer fusion framework.

The rest of this paper is organized as follows. Section II describes related studies for dimensional sentiment intensity prediction. Section III introduces details of our transformer fusion model for valence-arousal rating prediction. Section IV presents the experimental results and evaluation analysis. Conclusions are finally drawn in Section V.

II. RELATED WORK

This section describes the existing methods for sentiment intensity prediction, including lexicon-based [4], [10], [11], [12], [13], [14], [15], [16], regression-based [17], [18], [19],

[20], [21], [22], neural-network-based [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33] and transformer-based [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [49], [50], [51] approaches.

A. LEXICON-BASED METHODS

A number of one-dimensional sentiment lexicons provide word-level sentiment intensity, including SentiWordNet [10], SO-CAL [11], SentiStrength [12], and VADER [13]. Affective Norms of English Words (ANEW) [14], [16] and Extended ANEW [15] are three-dimensional lexicons which provide real-valued scores for the valence, arousal and dominance dimensions. Lexicon-based methods typically determine the sentiment intensity of a given text by averaging the sentiment scores of words matched in the lexicon [4]. These approaches are simple and easy to implement, but do not capture real sentiment expressions due to complex linguistic usages in the texts. For example, two phrases ''完全不同意'' (totally not agree) and ''不完全同意'' (not totally agree) have the same words with different ordering, and express meanings with almost opposite affective states. Hence, lexicon-based methods are usually used to provide baseline results for reference.

B. REGRESSION-BASED METHODS

Regression-based methods have been intensively studied to predict valence-arousal ratings. A cross-lingual approach was used to train a linear regression model for valence-arousal score prediction, in which the dimension scores of English seed words were regarded as the source language and their translated Chinese seed words were viewed as the target language [17]. The valence ratings of new words were estimated based on semantic similarity scores and a kernel model which was trained using least mean squares estimation [18]. A locally weighted regression method was proposed to improve linear regression to predict the valence-arousal values of affective words [19]. A community-based weighted graph model that performs the regression task on a graph was developed to predict the dimension scores of words [20]. A linear regression model was built to predict sentence-level affective ratings based on combinations of partial affective ratings of word n-grams [21]. The support vector regression was used to predict the sentiment intensity of words and phrases [22].

C. NEURAL-NETWORK-BASED METHODS

In recent years, neural network models with sentiment embeddings that capture contextual and emotional information of words have been applied to dimensional score prediction [23]. To learn sentiment embeddings, a word vector refinement model was proposed to refine existing pretrained word vectors using real-valued intensity scores provided by affective lexicons [24]. A boosted neural network trained on character-enhanced word embeddings was used to predict valence-arousal ratings of words [25]. A convolutional neural network (CNN) was trained on Twitter word embeddings to exploit neural activation values for Twitter sentiment classification and quantification [26]. A densely connected long short-term memory (LSTM) network was used to concatenate features at different levels to predict dimension scores of Chinese affective words and phrases [27]. An ensemble of different neural networks was developed to determine the intensity level for different emotion categories such as anger, fear, joy and sadness [28]. Bi-directional LSTM and CNN were combined to consider global and local information to predict emotional intensity of tweets [29]. A neural-network-based architecture that combines convolutional layers, fully-connected layers, linguistic features, and pretrained CNN activations in a non-sequential fashion was used for emotion intensity prediction in tweets [30]. An adversarial attention network was presented to predict the dimension scores of short texts [31]. A pipelined neural network model was used to sequentially learn word intensity and modifier weights for phrase-level sentiment intensity prediction [32]. A weighted-sum tree GRU model was developed to include dependency features for predicting Chinese phrase-level sentiment intensity in valence-arousal dimensions [33].

D. TRANSFORMER-BASED METHODS

Recently, BERT-like transformer architectures have been widely used for dimensional sentiment analysis. The pre-trained and case-sensitive BERT-base model was fine-tuned to predict the degree of sentiment intensity associated with multiple entities for aspect-based sentiment analysis [34]. A multi-task architecture based on the RoBERTa transformer was proposed to predict empathy and distress scores [35]. The RoBERTa multi-task model and the vanilla ELECTRA model was combined to predict empathy scores [36]. A demographic-aware EmpathBERT architecture was presented to infuse demographic information for empathy prediction [37]. The BERT transformer was used to recognize the emotion intensity scores of Japanese tweets on the topics of vaccinations [38]. Pre-trained BERTweet was used as the shared text encoder between a multi-label emotion classifier and a multi-dimension emotion regressor in a multi-task learning framework [39]. The pre-trained MacBERT transformers were used to fine-tune valence-arousal score prediction shared task for educational texts [40]. The BERT model was combined with specific sentiment word masking to improve sentence-level valence-arousal prediction [41]. The pre-trained RoBERTa-Large model was fine-tuned with categorical emotion labels to predict the continuous dimensions of valence, arousal, and dominance scores [42]. The domain-distilled BERT model was proposed to learn domain-invariant features on scarce language resources for dimensional sentiment score prediction [43].

Recently, transformer-based fusion methods have also been used for sentiment analysis, usually with promising results. BECMER combines a CNN model on audio signals and a BERT transformer on the lyrics for music emotion recognition [49]. The HFU-BERT framework improves the

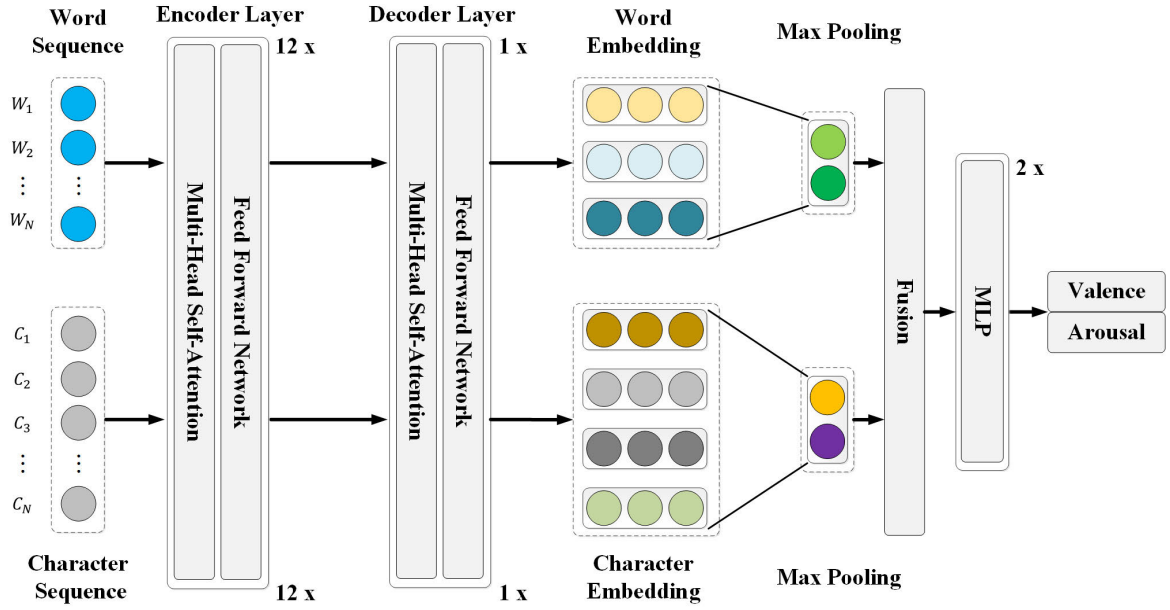

FIGURE 2. Our proposed transformer fusion framework. We propose word-level BERT to fuse the existing character-level BERT for Chinese dimensional sentiment intensity prediction. Two transformers in different granularities are separately pretrained and fine-tuned, and then jointly optimized to predict valence-arousal values.

BERT transformer by integrating heterogeneous language, audio, and visual features for multimodal emotion recognition [50]. A stacking method was used to fine-tune BERT to generate metadata for each emotion type separately and then assemble them to train a meta-classifier for emotion category prediction [51].

Different from the above fusion methods used for categorical sentiment analysis on multimedia and modal targets, we aim to develop a transformer-based fusion framework for dimensional sentiment analysis for Chinese texts. This paper reports the pre-training of a Chinese word-level BERT for contextualized language representations and propose a transformer fusion framework to combine word- and character-level BERT transformers for sentiment intensity prediction in the valence-arousal dimensions.

III. TRANSFORMER FUSION MODEL

Figure 2 shows our proposed network architecture for Chinese dimensional sentiment intensity prediction, comprised of two parts: 1) Chinese word-level BERT; and 2) word-/character-level transformer fusion.

A. CHINESE WORD-LEVEL BERT

BERT (Bidirectional Encoder Representations from Transformers) [44] is a pre-trained language model proposed by Google Research that uses a multi-layer transformer architecture as its network architecture. BERT uses an encoder architecture with an attention mechanism [53] to construct a transformer-based neural network architecture, providing state-of-the-art results in a wide variety of natural language processing tasks. There are two steps in the framework: 1) pre-training, in which the model is trained on unlabeled data over predefined tasks and 2) fine-tuning, in which the BERT model is first initialized with the pre-trained parameters and then fine-tuned using labeled data from the downstream tasks.

BERT proposes two pre-trained tasks: 1) Masked Language Model (MLM): a fixed ratio of tokens is masked to train BERT and the model then predicts the original value of the masked words based on the context and 2) Next Sentence Prediction (NSP): BERT is trained to predict whether the following sentence is probable or not based on the previous sentence. Through pre-training, BERT learns contextual embeddings for representations from large-scale data sets. After pre-training, BERT can be fine-tuned on smaller data sets to optimize its performance on specific tasks.

While a character-level BERT pre-trained model is publicly released [52], a Chinese word-level BERT is lacking due to the need for pre-processing in Chinese word segmentation over huge data sets. Therefore, we collected a huge set of text corpora and segmented the texts into words using the NCTU word segmentation tool [48] to train the word-level BERT model. We only trained the MLM task using the dynamic masking strategy [54] for language model training. We use the SentencePiece that uses Byte-Pair Encoding (BPE) as the subword detection mechanism.

B. WORD-/CHAR- LEVEL TRANSFORMER FUSION

We further propose a transformer fusion framework to combine our developed word-level BERT with the existing character-level BERT for sentiment intensity prediction in the valence-arousal dimensions. In the encoding layer, the word/character-level token embedding X_{emb} at a given position is obtained by looking up the embedding vector and adding up the word vectors that correspond to that position, as shown in Eq. (1). The positional encoding uses sine and cosine functions to encode the positional information,

ensuring a consistent relative relationship among different positions. For the self-attention mechanism, three parameter matrices W^q, W^k and W^v are used to respectively map the input vector X_{emb} to three new vectors $Q = W^q X_{emb}$, $K = W^k X_{emb}$, and $V = W^v X_{emb}$. The residual convergence of our multi-head vector is accelerated by the layer normalization calculation shown in Eq. (2). Finally, the multi-headed embedding vector X_{enc} is computed using two linear transformations and the activation function GeLU, as shown in Eq. (3).

In the decoding layer, we obtain the different granularity embedding X_{word} and X_{char} via 1-layer transformer that is identical to the encoding layer using a 2-head multi-head attention with 256 hidden dimensions, while using max-pooling to retain the important features for each dimension [55], as shown in Eq. (4). Eventually, we concatenate different granularity embeddings P_{word} and P_{char} together, as shown in Eq. (5), which are used to obtain prediction scores using a 2-layer Multi-Layer Perceptron (MLP) with the activation function hyperbolic tangent (tanh).

$$X_{emb} = EmbLookup(X) + PosEncoding \quad (1)$$

$$X_{att} = LayerNorm(X_{emb} + SelfAtt(Q, K, V)) \quad (2)$$

$$X_{enc} = GeLU(Linear(Linear(X_{att}))) \quad (3)$$

$$P_{word}, P_{char} = max(X_{word}, X_{char}) \quad (4)$$

$$r = \tanh(\tanh(P_{word} + P_{char})) \quad (5)$$

For sentiment intensity prediction in the valence-arousal dimensions, we use the downstream task datasets to fine-tune pre-trained word/character-level BERT model in our transformer fusion framework to obtain the valence-arousal ratings.

Take the following sentence "為什麼自己現在可以這麼毅然的捨棄?" (Why can I give up so resolutely now?) with VA ratings of 4.333 and 4.000 as an example. We can obtain a 17-tokens character sequence as "為 什 麼 自 … 的 捨 棄 ?" and a 9-tokens word sequence "為什麼(why) 自己(I) 現 在(now) 可 以(can) 這 麼(so) 毅 然(resolutely) 的(pronounced as De) 捨 棄(give up)" to generate the embeddings at both the character- and word-levels based on Eq. (1). Both embedding sequences at different levels of linguistic granularity are fed into the encoder layer of the 12-layer word-/character-level BERT model, using the process described in Eq. (2) and using the GeLU activation function specified in Eq. (3). Then, the outputs are passed to the decoder layer of a 1-layer transformer using 2-head multi-head attention. Consequently, through the max pooling operation described in Eq. (4), we can respectively obtain sampled word/character embeddings for fusion. Finally, following Eq. (5), we merge the word/character representations through 2-layer MLPs to predict the VA ratings. Comparing the predicted results of this example sentence, standalone word-level BERT predicted a valence of 4.969 and an arousal of 5.314, while the standalone character-level BERT model predicted VA ratings of 5.048 and 4.801. Our word/character-level BERT fusion model can obtain improved valence (3.916) and arousal (4.003) results, relatively close to human-annotated VA ratings of 4.333 and 4.000.

IV. EVALUATION

A. DATASETS

Chinese valence-arousal sentences (CVAS) and Chinese valence-arousal texts (CVAT) from the Chinese EmoBank corpus [9] were used to evaluate sentiment intensity prediction performance. Valence-Arousal (VA) ratings were annotated through crowdsourcing with each instance randomly assigned to 10 annotators. Both the valence and arousal dimension use a nine-degree scale. A value of 1 on the valence and arousal dimensions respectively denotes extremely high-negative and low-arousal sentiment, while a 9 denotes extremely high-positive and high-arousal sentiment, and 5 denotes a neutral and medium-arousal sentiment. Outlier ratings were identified and excluded from the calculation of the average VA ratings.

CVAS was collected from Chinese tweets, including 2,852 single sentences with an average of 11.7 characters or 7.3 words. CVAT collects web texts crawled across six different categories: news articles, political discussion forums, car discussion forums, hotel reviews, book reviews, and laptop reviews. A total of 2,969 multi-sentence texts were included in the CVAT each with an average of 55.1 characters or 35.5 words. Each instance in the CVAT is about five times comparing with CVAS in terms of character or word lengths. In addition, the ratios between the number of characters divided by the number of words are respectively near 1.6 in the CVAS and 1.55 in the CVAT.

B. SETTINGS

To train Chinese word-level BERT, we collected the following text resources: LDC Chinese Gigaword (Version 2.0),[1] Sinica Balance Corpus (Version 4.0),[2] Chinese Information Retrieval Benchmark (Version 3.03),[3] Taiwan Panorama Magazine,[4] Mandarin Conversation Dialogue Corpus (MCDC),[5] National Educational Radio Corpus,[6] Microphone Speech Database (TCC300),[7] and NYCU text corpus (collected from Chinese Wikipedia[8] and other web pages). After preprocessing based on NCTU word segmentation tool [48] and text normalization, we obtained about 2.8 billion words to train Chinese word-level BERT model. Our quantity scale is huge, but it still has a clear gap comparing with English BERT released by Google[9] that was trained on 3.3 billion words.

[1] https://catalog.ldc.upenn.edu/LDC2005T14
[2] http://www.aclclp.org.tw/use_asbc.php
[3] http://www.aclclp.org.tw/use_cir.php
[4] https://www.taiwan-panorama.com/
[5] http://shachi.org/resources/4037
[6] http://www.aclclp.org.tw/use_mat_c.php#ner
[7] http://www.aclclp.org.tw/use_mat.php#tcc300edu
[8] https://zh.wikipedia.org/wiki
[9] https://web.stanford.edu/class/cs224n/slides/Jacob_Devlin_BERT.pdf

The experimental implementations were carried out using the ASUS Taiwan Computing Cloud (TWCC)[10] computing resource. The hyper-parameters of our transformer fusion framework were set up as follows: batch size 16; max pooling style; decoder used 1-layer transformer; and compared character-level BERT.[11] Our developed word-level BERT both had 12-layers, 768-hidden and 12-heads; 2-layer MLP dimensions of 768; the optimizer was AdamW; and the number of epochs were restricted to 20.

C. METRICS

We used five-fold cross-validation evaluation, identical to that used for the Chinese EmoBank corpus [9]. The sentiment intensity predication performance is evaluated by examining the difference between machine-predicted ratings and human-annotated ratings using two metrics to independently evaluate the valence and arousal predictions: Mean Absolute Error (MAE) and Pearson Correlation Coefficient (PCC), defined as follows

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|a_i - p_i| \tag{6}$$

$$PCC = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{a_i - \mu_A}{\sigma_A})(\frac{p_i - \mu_P}{\sigma_A}) \tag{7}$$

where $a_i \in A$ and $p_i \in P$ respectively denote the i-th actual value and predicted value, n is the number of test samples, and σ_A respectively represent the mean value and the standard deviation of A, while μ_P and σ_P respectively represent the mean value and the standard deviation of P.

The actual and predicted real values range from 1 to 9, so MAE measures the error rate in a range where the lowest value is 1 and the highest value is 9. A lower MAE indicates more accurate prediction performance. The PCC is a value between -1 and 1 that measures the linear correlation between the actual value and the predicated value. A lower MAE and a higher PCC indicate more accurate prediction performance. Each metric for the valence and arousal dimensions is ranked independently. A model's overall ranking is computed based on the cumulative rank across the four metrics. The lower the cumulative rank, the better the system performance.

D. RESULTS

In the first set of experiments, the following four model types were compared to demonstrate their performance. Experimental results of the first three types were obtained from the Chinese EmoBank corpus evaluation [9] for reference, whereas the last one was conducted by this study.

- Lexicon-based method [5], [15]: Chinese Valence-Arousal Words (CVAW) and Chinese Valence-Arousal Phrases (CVAP) from the Chinese EmoBank corpus were used to predict the valence (or arousal) ratings of a given instance in CVAS (or CVAT) by averaging the valence (or arousal) ratings of words/phrases in the CVAW and CVAP.
- Regression-based methods: including the Linear Regression (LR) [17] and Support Vector Regression (SVR) [22].
- Neural-Network-based methods: including Convolutional Neural Network (CNN) [26], Recurrent Neural Network (RNN) [56], Long Short-Term Memory (LSTM) [57], Attention LSTM [58].
- Transformer-based methods: including character-level BERT model () [44] released by the Google Research, our developed word-level BERT and transformer fusion model.

Table 1 shows the prediction results of CVAS. For both the lexicon-based and regression-based methods, the SVR approach outperformed the others in both the valence and arousal dimensions. The character-level BERT outperformed the other neural-network-based methods in both dimensions. Comparing the results achieved by our character-level BERT, our word-level BERT had a slightly lower cumulative rank. In our observations, short sentences with an average of 7.3 words (or 11.7 characters) do not provide sufficient information for valence-arousal rating prediction using complicated neural networks, especially for those word-level based models. Our fusion model ranked first for valence MAE (0.494) and valence PCC (0.891), while the character-level BERT ranked first for arousal MAE (0.700). Finally, both methods tied first for overall performance with the same cumulative rank.

Table 2 shows the prediction results of CVAT. For lexicon-based, regression-based, and neural-network-based methods, we obtained nearly consistent findings. For transformer-based methods, the overall performance of our word-level BERT was close to that of character-level BERT in terms of overall cumulative rank. Based on our observations, the average word length of a given text in CVAT is about five times that of a short sentence in CVAS. These characteristic benefits the word-level based models. Our fusion model ranked first for valence MAE (0.519), arousal MAE (0.494) and arousal PCC (0.695) and second for valence PCC (0.831). Overall, our fusion model ranked first in terms of cumulative rank.

In summary, almost all models on the CVAS clearly underperformed the corresponding model results on the CVAT. The valence-arousal ratings for CVAS data containing single sentences from Twitter were more difficult to predict than for multi-sentences texts that provide more information in CVAT. Comparing results achieved by word-level BERT on CVAS and CVAT, we find that performance improve with increased sentence length. Character-level BERT outperformed word-level, possibly because the insufficient size of pre-trained data sets, with a difference of about 500 million words. However, our fusion model combining word- and character-level BERT provided the best overall performance by including features in different linguistic granularities.

[10] ASUSTWCC computing resources: https://www.twcc.ai/
[11] Multilingual BERT: https://github.com/google-research/bert

TABLE 1. Results of sentiment intensity prediction on CVAS.

CVAS						
Method		Valence		Arousal		Cumulative Rank
		MAE (rank)	PCC (rank)	MAE (rank)	PCC (rank)	
Lexicon	CVAW/CVAP	0.940 (9)	0.593 (7)	1.214 (10)	0.266 (9)	35
Regression	LR	1.079 (10)	0.493 (10)	1.183 (9)	0.264 (10)	39
	SVR	0.886 (6)	0.612 (5)	0.954 (6)	0.414 (6)	23
Neural Network	CNN	0.920 (8)	0.564 (9)	0.975 (8)	0.364 (8)	33
	RNN	0.909 (7)	0.573 (8)	0.964 (7)	0.373 (7)	29
	LSTM	0.871 (5)	0.602 (6)	0.946 (5)	0.429 (5)	21
	Attention	0.857 (4)	0.621 (4)	0.943 (4)	0.432 (4)	16
Transformer	char-BERT	0.531 (2)	0.792 (2)	**0.700 (1)**	0.490 (2)	**7**
	word-BERT (Our word-level)	0.534 (3)	0.782 (3)	0.766 (3)	**0.555 (1)**	10
	word-BERT + char-BERT (Our fusion)	**0.494 (1)**	**0.891 (1)**	0.706 (2)	0.459 (3)	**7**

TABLE 2. Results of sentiment intensity prediction on CVAT.

CVAT						
Method		Valence		Arousal		Cumulative Rank
		MAE (rank)	PCC (rank)	MAE (rank)	PCC (rank)	
Lexicon	CVAW/CVAP	0.928 (10)	0.621 (10)	0.871 (10)	0.279 (10)	40
Regression	LR	0.791 (8)	0.701 (8)	0.833 (9)	0.423 (8)	33
	SVR	0.710 (6)	0.760 (6)	0.716 (6)	0.530 (6)	24
Neural Network	CNN	0.814 (9)	0.665 (9)	0.799 (8)	0.396 (9)	35
	RNN	0.716 (7)	0.740 (7)	0.738 (7)	0.493 (7)	28
	LSTM	0.657 (5)	0.777 (5)	0.715 (5)	0.534 (5)	20
	Attention	0.621 (4)	0.802 (3)	0.696 (4)	0.553 (4)	15
Transformer	char-BERT	0.531 (2)	**0.874 (1)**	0.615 (3)	0.609 (3)	9
	word-BERT (Our word-level)	0.573 (3)	0.802 (3)	0.582 (2)	0.663 (2)	10
	word-BERT + char-BERT (Our fusion)	**0.519 (1)**	0.831 (2)	**0.494 (1)**	**0.695 (1)**	**5**

V. CONCLUSION

We propose a transformer fusion framework for Chinese sentiment intensity prediction in the valence-arousal dimensions, making the following contributions:

(1) We develop a Chinese word-level BERT model based on huge collected data sets to obtain contextual language representations. We plan to release the pre-trained language model for further research.

(2) We propose a transformer fusion framework to predict valence-arousal ratings for dimensional sentiment analysis. Experimental results on the Chinese EmoBank indicate that our fusion model integrating word- and character-level BERT outperformed other neural-network-based, regression-based and lexicon-based methods.

Future work will exploit other semantic features and develop other pre-training tasks to further improve performance for Chinese dimensional sentiment analysis.

ACKNOWLEDGMENT

The authors sincerely appreciate ASUS TWCC for providing computing resources.

REFERENCES

[1] G. Mishne and M. de Rijke, "MoodViews: Tools for blog mood analysis," presented at the AAAI Spring Symp., Comput. Approaches Analyzing Weblogs, 2006.

[2] A. Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," *Comput. Intell.*, vol. 22, no. 2, pp. 110–125, May 2006.

[3] M.-C. De Marneffe, C. D. Manning, and C. Potts, "Was it good? It was provocative.' Learning the meaning of scalar adjectives," in *Proc. 48th Annu. Meeting Assoc. Comput. Linguistics*, 2010, pp. 167–176.

[4] G. Paltoglou and M. Thelwall, "Seeing stars of valence and arousal in blog posts," *IEEE Trans. Affect. Comput.*, vol. 4, no. 1, pp. 116–123, Jan. 2013.

[5] G. Paltoglou, M. Theunis, A. Kappas, and M. Thelwall, "Predicting emotional responses to long informal text," *IEEE Trans. Affect. Comput.*, vol. 4, no. 1, pp. 106–115, Jan. 2013.

[6] L.-C. Yu, J.-L. Wu, P.-C. Chang, and H.-S. Chu, "Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news," *Knowledge-Based Syst.*, vol. 41, pp. 89–97, Mar. 2013.

[7] L.-C. Yu, L.-H. Lee, S. Hao, J. Wang, Y. He, J. Hu, K. R. Lai, and X. Zhang, "Building Chinese affective resources in valence-arousal dimensions," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 540–545.

[8] 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 Assoc. Comput. Linguistics, Short Papers*, 2017, pp. 578–585.

[9] L.-H. Lee, J.-H. Li, and L.-C. Yu, "Chinese EmoBank: Building valence-arousal resources for dimensional sentiment analysis," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 21, no. 4, pp. 1–18, Jul. 2022.

[10] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in *Proc. 7th Int. Conf. Lang. Resour. Eval.*, 2010, pp. 2200–2204.

[11] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Comput. Linguistics*, vol. 37, no. 2, pp. 267–307, Jun. 2011.

[12] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment strength detection for the social web," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 63, no. 1, pp. 163–173, Jan. 2012.

[13] C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. 8th Int. AAAI Conf. Weblogs Social Media*, 2014, pp. 216–225.

[14] M. M. Bradley and P. J. Lang, "Affective norms for english words (ANEW): Instruction manual and affective ratings," Center Res. Psychophysiol., Univ. Florida, Gainesville, FL, USA, Tech. Rep. C-1, 1999.

[15] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behav. Res. Methods*, vol. 45, no. 4, pp. 1191–1207, Dec. 2013.

[16] D. Gökçay, E. Isbilir, and G. Yildirim, "Predicting the sentiment in sentences based on words: An exploratory study on ANEW and ANET," in *Proc. IEEE 3rd Int. Conf. Cognit. Infocommun. (CogInfoCom)*, Dec. 2012, pp. 715–718.

[17] W.-L. Wei, C.-H. Wu, and J.-C. Lin, "A regression approach to affective rating of Chinese words from ANEW," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.* Memphis, TN, USA: Springer, Oct. 2011, pp. 121–131.

[18] N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan, "Kernel models for affective lexicon creation," in *Proc. Interspeech*, Aug. 2011, pp. 1–4.

[19] J. Wang, L.-C. Yu, K. R. Lai, and X. Zhang, "Locally weighted linear regression for cross-lingual valence-arousal prediction of affective words," *Neurocomputing*, vol. 194, pp. 271–278, Jun. 2016.

[20] J. Wang, L.-C. Yu, K. R. Lai, and X. Zhang, "Community-based weighted graph model for valence-arousal prediction of affective words," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 1957–1968, Nov. 2016.

[21] N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan, "Distributional semantic models for affective text analysis," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 11, pp. 2379–2392, Nov. 2013.

[22] S. Amir, R. Astudillo, W. Ling, B. Martins, M. J. Silva, and I. Trancoso, "INESC-ID: A regression model for large scale Twitter sentiment lexicon induction," in *Proc. 9th Int. Workshop Semantic Eval. (SemEval)*, 2015, pp. 613–618.

[23] D. Tang, F. Wei, B. Qin, N. Yang, T. Liu, and M. Zhou, "Sentiment embeddings with applications to sentiment analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 2, pp. 496–509, Feb. 2016.

[24] L.-C. Yu, J. Wang, K. R. Lai, and X. Zhang, "Refining word embeddings using intensity scores for sentiment analysis," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 3, pp. 671–681, Mar. 2018.

[25] S. Du and X. Zhang, "Aicyber's system for IALP 2016 shared task: Character-enhanced word vectors and boosted neural networks," in *Proc. Int. Conf. Asian Lang. Process. (IALP)*, Nov. 2016, pp. 161–163.

[26] D. Vilares, Y. Doval, M. A. Alonso, and C. Gómez-Rodríguez, "LyS at SemEval-2016 task 4: Exploiting neural activation values for Twitter sentiment classification and quantification," in *Proc. 10th Int. Workshop Semantic Eval. (SemEval-)*, 2016, pp. 79–84.

[27] C. Wu, F. Wu, Y. Huang, S. Wu, and Z. Yuan, "THU_NGN at IJCNLP-2017 task 2: Dimensional sentiment analysis for Chinese phrases with deep LSTM," in *Proc. 8th Int. Joint Conf. Natural Lang. Process. (IJCNLP)*, 2017, pp. 47–52.

[28] 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 Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, 2017, pp. 58–65.

[29] Y. He, L.-C. Yu, K. R. Lai, and W. Liu, "YZU-NLP at EmoInt-2017: Determining emotion intensity using a bi-directional LSTM-CNN model," in *Proc. 8th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, 2017, pp. 238–242.

[30] D. Kulshreshtha, P. Goel, and A. K. Singh, "How emotional are you? Neural architectures for emotion intensity prediction in microblogs," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 2914–2926.

[31] S. Zhu, S. Li, and G. Zhou, "Adversarial attention modeling for multi-dimensional emotion regression," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 471–480.

[32] L.-C. Yu, J. Wang, K. R. Lai, and X. Zhang, "Pipelined neural networks for phrase-level sentiment intensity prediction," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 447–458, Jul. 2020.

[33] Y.-C. Deng, C.-Y. Tsai, Y.-R. Wang, S.-H. Chen, and L.-H. Lee, "Predicting Chinese phrase-level sentiment intensity in valence-arousal dimensions with linguistic dependency features," *IEEE Access*, vol. 10, pp. 126612–126620, 2022.

[34] J. Zheng, S. Friedman, S. Schmer-Galunder, I. Magnusson, R. Wheelock, J. Gottlieb, D. Gomez, and C. Miller, "Towards a multi-entity aspect-based sentiment analysis for characterizing directed social regard in online messaging," in *Proc. 6th Workshop Online Abuse Harms (WOAH)*, 2022, pp. 203–208.

[35] A. Kulkarni, S. Somwase, S. Rajput, and M. Marathe, "PVG at WASSA 2021: A multi-input, multi-task, transformer-based architecture for empathy and distress prediction," *Proc. 11th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, 2021, pp. 105–111.

[36] J. Mundra, R. Gupta, and S. Mukherjee, "WASSA@IIK at WASSA 2021: Multi-task learning and transformer finetuning for emotion classification and empathy prediction," in *Proc. 11th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal., Sentiment Social Media Anal.*, 2021, pp. 112–116.

[37] B. P. R. Guda, A. Garimella, and N. Chhaya, "EmpathBERT: A BERT-based framework for demographic-aware empathy prediction," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics, Main Volume*, 2021, pp. 3072–3079.

[38] P. J. Ramos, K. Ferawati, K. Liew, E. Aramaki, and S. Wakamiya, "Emotion analysis of writers and readers of Japanese tweets on vaccinations," in *Proc. 12th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, 2022, pp. 95–103.
[39] R. Mukherjee, A. Naik, S. Poddar, S. Dasgupta, and N. Ganguly, "Understanding the role of affect dimensions in detecting emotions from tweets: A multi-task approach," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 2303–2307.
[40] M.-C. Hung, C.-Y. Chen, P.-J. Chen, and L.-H. Lee, "NCU-NLP at ROCLING-2021 shared task: Using MacBERT transformers for dimensional sentiment analysis," in *Proc. 33rd Conf. Comput. Linguistics Speech Process. (ROCLING)*, 2021, pp. 380–384.
[41] J.-L. Wu and W.-Y. Chung, "Sentiment-based masked language modeling for improving sentence-level valence–arousal prediction," *Int. J. Speech Technol.*, vol. 52, no. 14, pp. 16353–16369, Nov. 2022.
[42] S. Park, J. Kim, S. Ye, J. Jeon, H. Y. Park, and A. Oh, "Dimensional emotion detection from categorical emotion," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 4367–4380.
[43] W. Lin and L.-C. Yu, "Scarce resource dimensional sentiment analysis using domain-distilled BERT," *J. Inf. Sci. Eng.*, vol. 39, no. 2, pp. 305–321, 2023.
[44] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL), (Long Short Papers)*, vol. 1, 2019, pp. 1–16.
[45] T. Wolf, "HuggingFace's transformers: State-of-the-art natural language processing," 2019, *arXiv:1910.03771*.
[46] X. Tan, G. Deng, and X. Hu, "Multi-granularity context semantic fusion model for Chinese event detection," in *Proc. 10th Int. Conf. Internet Comput. Sci. Eng.*, Jul. 2021, pp. 1–7.
[47] W. Li, R. Sun, and Y. Wu, "Exploiting word semantics to enrich character representations of Chinese pre-trained models," in *Proc. Int. Conf. Natural Lang. Process. Chin. Comput.*, 2022, pp. 3–15.
[48] Y.-R. Wang and Y.-F. Liao, "Word vector/conditional random field-based Chinese spelling error detection for SIGHAN-2015 evaluation," in *Proc. 8th SIGHAN Workshop Chin. Lang. Process.*, 2015, pp. 46–49.
[49] B.-H. Sung and S.-C. Wei, "BECMER: A fusion model using BERT and CNN for music emotion recognition," in *Proc. IEEE 22nd Int. Conf. Inf. Reuse Integr. Data Sci. (IRI)*, Aug. 2021, pp. 437–444.
[50] S. Lee, D. K. Han, and H. Ko, "Multimodal emotion recognition fusion analysis adapting BERT with heterogeneous feature unification," *IEEE Access*, vol. 9, pp. 94557–94572, 2021.
[51] S.-Y. Lin, Y.-C. Kung, and F.-Y. Leu, "Predictive intelligence in harmful news identification by BERT-based ensemble learning model with text sentiment analysis," *Inf. Process. Manage.*, vol. 59, no. 2, Mar. 2022, Art. no. 102872.
[52] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-training with whole word masking for Chinese BERT," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3504–3514, 2021.
[53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6000–6010.
[54] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
[55] K. Yang, D. Lee, T. Whang, S. Lee, and H. Lim, "EmotionX-KU: BERT-max based contextual emotion classifier," 2019, *arXiv:1906.11565*.
[56] O. Irsoy and C. Cardie, "Opinion mining with deep recurrent neural networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 720–728.
[57] X. Wang, Y. Liu, C. Sun, B. Wang, and X. Wang, "Predicting polarities of tweets by composing word embeddings with long short-term memory," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process. (Long Papers)*, vol. 1, 2015, pp. 1343–1353.
[58] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 1480–1489.

YU-CHIH DENG received the M.S. degree in communication engineering from National Taipei University, Taiwan, in 2017. He is currently pursuing the Ph.D. degree with the Speech Processing Laboratory, Institute of Electrical Engineering, National Yang Ming Chiao Tung University, Taiwan, under the guidance of Prof. Yih-Ru Wang and Sin-Horng Chen. His research interests include natural language processing and automatic speech recognition.

YIH-RU WANG received the B.S. and M.S. degrees from the Department of Communication Engineering, National Chiao Tung University, Taiwan, in 1982 and 1987, respectively, and the Ph.D. degree from the Institute of Electronic Engineering, National Chiao Tung University, in 1995. He was an Associate Professor with National Yang Ming Chiao Tung University, until 2021. He is currently a part-time researcher. His general research interests include automatic speech recognition and natural language processing.

SIN-HORNG CHEN received the B.S. degree in communications engineering and the M.S. degree in electrical engineering from National Chiao Tung University, Taiwan, in 1976 and 1978, respectively, and the Ph.D. degree in electrical engineering from Texas Tech University, USA, in 1983. He was appointed as an Associate Professor and a Professor with the Department of Communications Engineering, NCTU, in 1983 and 1990, respectively, where he was the Dean of the ECE College and the acting President of NCTU. Currently, he is a Chair Professor with National Yang Ming Chiao Tung University. His major research interests include speech signal processing, particularly Mandarin speech recognition, text-to-speech, and speech prosody.

LUNG-HAO LEE (Member, IEEE) received the B.S. degree in statistics from National Taipei University, Taiwan, in 2003, the M.S. degree in information management from Yuan Ze University, Taoyuan, Taiwan, in 2005, and the Ph.D. degree in computer science and information engineering from National Taiwan University, in 2015. From 2015 to 2018, he was a Postdoctoral Fellow with National Taiwan Normal University. He joined the Department of Electrical Engineering, National Central University, Taiwan, in 2018, as an Assistant Professor. He is currently an associate professor. His research interests include natural language processing, information retrieval and extraction, biomedical and health informatics, and artificial intelligence technologies.