The Biases of Pre-trained Language Models: An Empirical Study on Prompt-based Sentiment Analysis and Emotion Detection

Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria*, *Fellow, IEEE*

Abstract—Thanks to the breakthrough of large-scale pre-trained language model (PLM) technology, prompt-based classification tasks, e.g., sentiment analysis and emotion detection, have raised increasing attention. Such tasks are formalized as masked language prediction tasks which are in line with the pre-training objects of most language models. Thus, one can use a PLM to infer the masked words in a downstream task, then obtaining label predictions with manually defined label-word mapping templates. Prompt-based affective computing takes the advantages of both neural network modeling and explainable symbolic representations. However, there still remain many unclear issues related to the mechanisms of PLMs and prompt-based classification. We conduct a systematic empirical study on prompt-based sentiment analysis and emotion detection to study the biases of PLMs towards affective computing. We find that PLMs are biased in sentiment analysis and emotion detection tasks with respect to the number of label classes, emotional label-word selections, prompt templates and positions, and the word forms of emotion lexicons.

Index Terms—Sentiment Analysis, Emotion Detection, Prompt, Pre-trained Language Model

1 Introduction

Pre-trained language models (PLMs) have achieved remarkable performance on diverse natural language processing (NLP) tasks [1], [2], [3], [4]. These language models were trained with large-scale open-domain corpora, thus, embedding rich semantic and syntactic information. Since many PLMs introduced a masked word prediction object to learn the contextualized information, a new learning paradigm, prompting, becomes popular recently.

As a neuro-symbolic method whose decision is conditioned on the interaction between predefined emotion lexicon knowledge, label-word mapping templates, and neural network-based probability modeling, prompt-based affective computing has attracted widespread attention. Unlike conventional practices that fine-tune a PLM on a downstream classification task, a prompt-based method considers a classification task as a masked word prediction task. Taking a binary sentiment classification task as an example, given a sentence, "This movie is very interesting", a prompt-based method normally introduces an additional prompt with a [MASK] token upon the input sentence. Then, the sentence is formalized as "I feel [MASK]. This movie is very interesting". In this case, the language model can predict the probabilities of a selected positive emotion word, e.g., "joyful", and a negative emotion word, e.g., "sad", appearing in the [MASK] position.

The sentiment analysis labels, "positive" and "negative", are manually mapped to the emotion words, "joyful" and "sad", respectively, which is termed as label-word mapping. We define the words, e.g., "joyful" and "sad" that are used for predicting labels as label-words. If the probability of "joyful" is higher than the probability of "sad", the original sentence is classified as positive, otherwise, negative. In this way, a PLM can predict the sentiment polarity without fine-tuning, because masked word prediction was exactly used as one of the pre-training objects of many PLMs.

Compared with conventional fine-tuning-based classification, the advantage of prompt-based methods can exempt from annotating large-scale datasets and supervised learning with the datasets. Instead of forcing a PLM to fine-tune its parameters to fit the label distributions of a downstream task, prompting modifies the input to fit the task to the original language modeling [5]. Such an advantage is especially useful for low-resource few-shot and zero-shot learning tasks, e.g., metaphor interpretation [6], [7], text classification [8], [9] and natural language inference [10].

However, there are many variables that impact the performance of a prompt-based classification model, such as PLM selections, prompt templates and positions, the label-word and their Part-of-Speech (PoS) variations. The above issues motivate us to conduct a systematic empirical study with two related affective classification tasks, namely sentiment analysis and emotion detection. Both affective aspects reflect the feeling of a subjective, thus, the two tasks allow us to use the same sets of label-words. The difference is that emotion is a naturally aroused physical reaction towards a circumstance, while sentiment is the attitude towards the circumstance. NLP community likely uses positive and negative labels to distinguish the two extremes of sentiment status, and use an emotion model to categorize emotional intention, e.g., anger, fear, sadness, and joy *et al.*

* *Corresponding author: Erik Cambria.*

- *Rui Mao[1], Qian Liu[2], Wei Li[3] and Erik Cambria[4] are with the School of Computer Science and Engineering, Nanyang Technological University, Singapore, 639798. Email: {[1]rui.mao, [2]liu.qian, [4]cambria}@ntu.edu.sg, [3]wei008@e.ntu.edu.sg.*
- *Kai He is with the School of Computer Science and Technology, Xi'an Jiaotong University, Shanxi, China, 710049. Email: hk52025804@stu.xjtu.edu.cn.*

Whereas, it is still arguable towards different emotion taxonomies [11]. Thus, there are different emotion models with different hierarchically structured taxonomies [12], [13], [14], [15], [16], which presents different sets of widely used label-words for both sentiment analysis and emotion detection in our empirical study. We mean to examine the emotion models in prompt-based sentiment analysis and emotion detection, demonstrating the performance differences, based on different control experiments. The performance diversities illustrate the biases of PLMs in prompting. We hope our findings may deliver informative research opportunities for improving the performance of prompting methods in future work, such as better label-word selections, and prompt engineering.

We define the bias of PLMs as from a human perspective, one should produce the same or similar prediction results based on given variables, whereas, PLMs deliver different results. Studying the bias of PLMs is important, because many biases of PLMs are counter-intuitive. Thus, we come up with three research questions, based on sentiment analysis and emotion detection:

Q1. Does PLM-based prompting bias towards coarse-grained and fine-grained classification tasks?
Q2. Does PLM-based prompting bias towards frequent and infrequent label-words?
Q3. Does PLM-based prompting bias towards different prompt templates and label-word forms with similar meanings?

To explore these questions, we analyze different PLMs, e.g., BERT [1], RoBERTa [2], ALBERT [3] and BART [4], and emotion models with coarse-grained and fine-grained taxonomies [12], [13], [14], [15], [16]. In our experiments, we find that the fine-grained emotion taxonomy proposed by Susanto *et al.* [16] and RoBERTa large are more efficient for sentiment analysis and emotion detection tasks. We also find that prompt-based models yield weak performance on emotion detection (a fine-grained classification task), because the emotional label-words are less differentiable for emotion detection than sentiment analysis. We do not find large PLMs tend to yield higher probabilities for frequent label-words. Finally, the prompt templates and positions, the label-word selections such as different emotion model lexicons and PoS do impact the model performance somehow. PLM-based prompting prefers adjective label-words to nouns in our examined affective computing tasks, whereas using both nominal and adjective label-words yields higher accuracy. The position of prompts is a factor, impacting model performance. However, our testing results cannot achieve an agreement to determine whether prefixed prompting or suffixed prompting is better, because different prompt templates present different results.

The contribution of this work is twofold: (1) We conduct a systematic empirical study on prompting sentiment analysis and emotion detection tasks by comparing different PLMs, emotion taxonomies, prompt templates, prompt positions, different PoS of employed label-words; (2) We deliver an in-depth analysis about the biases of prompt-based classification, summarizing several research challenges in this area.

2 Related Work

2.1 Pre-trained language model

Due to the expensive costs associated with annotating large-scale datasets, attempts have been made to infuse deep neural networks with prior knowledge of semantics and syntax from large-scale open corpora through pre-training language models. The prior knowledge is learned via diverse objects, such as missing or masked word prediction, next sentence prediction, sentence-order prediction, corrupted text reconstruction, autoregressive language modeling in different PLMs [1], [2], [3], [4], [17], [18], [19]. Then, the PLMs are fine-tuned to learn domain-specific knowledge in different downstream tasks [20], [21], or using the hidden states of PLMs as features of a task-specific model [1], [17], [22], [23] in a supervised learning fashion.

2.2 Prompt-based classification

Prompt-based classification has been widely explored in the contexts of zero- and few-short learning [9], [10], [24], [25]. These methods naturally inherit the learning paradigm of PLMs, e.g., masked word prediction, thus, making the best use of the learned knowledge of PLMs with zero or few learning samples from a downstream task. Puri and Catanzaro [24] fine-tuned a generative model (GPT-2 [26]) on a document-title pairing task with a questing-answer learning object. Then, they use the fine-tuned GPT-2 in zero-short and weak supervised learning classification tasks. Schick and Schütze [10] employed an ensemble strategy to generate soft-labels for unannotated data by using multiple prompting-fine-tuned PLMs. Then, they trained a classifier with the soft-labeled dataset. Although PLMs have achieved remarkable performance on prompt-tuning-based classification, the biases of the PLMs were rarely studies. We freeze the parameters of PLMs to study the biases, based on different setups of prompts and label-words.

2.3 Affective computing

Many textual affective computing tasks, e.g., sentiment analysis and emotion detection were processed as supervised machine learning tasks [27], [28], [29], [30], [31], [32], [33]. They set up learning objects to train the models to fit input text to label distributions. For example, Li *et al.* [33] proposed Bidirectional Emotional Recurrent Unit for supervised learning sentiment analysis in conversations. With PLMs, the fitting abilities of affective classifiers were significantly improved [34], [35], [36]. For example, Mao and Li [36] proposed a novel soft-parameter sharing mechanism and employed BERT as the sharing encoder in their multi-task learning-based affective computing model. On the other hand, the boundaries of interpretability of affective computing were expended with symbolic and sub-symbolic algorithms [37], [38], [39], [40], [41]. These methods likely utilized emotional lexicon knowledge. However, the debates in emotional taxonomies never stop [11], yielding different coarse-grained and fine-grained emotion models [12], [13], [14], [15], [16]. For example, Susanto *et al.* [16] revised an hour glass emotion model [42], proposing 8 basic emotion classes and their associated fine-grained emotions, achieving better performance on an emotion detection task.

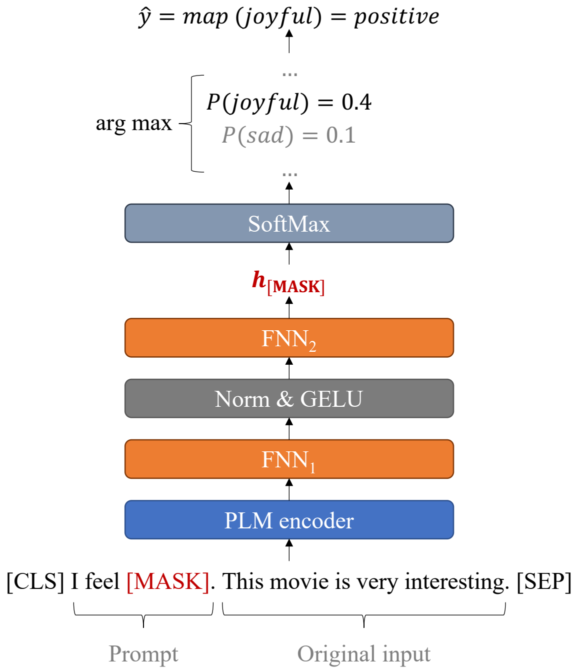


Fig. 1. The framework of an examined prompt-based sentiment analysis model. PLM denotes a pre-trained language model; FNN denotes a pre-trained feedforward neural network; GELU is an activation function; Norm denotes layer normalization; $h_{[\text{MASK}]}$ is the hidden state of the [MASK] token. "joyful" and "sad" are example label-words, sourced from an emotion model. $map(\cdot)$ is a manually defined mapping function that maps a label-word to a label class in sentiment analysis.

When neuro-symbolic affective computing comes to the era of prompting, Gao *et al.* [25] used manual prompts, "It was [MASK]", and label-words, "great" and "terrible", for prompting binary sentiment polarities in their zero-shot setup. They also explored an automatic prompt generation method with fine-tuning. Hambardzumyan *et al.* [43] introduced trainable parameters as virtual prompts for sentiment analysis. The performance was comparable to fine-tuning PLM-based sentiment analysis. However, to the best of our knowledge, there has not been a systematic empirical study to evaluate different emotion models on prompt-base affective computing.

Although previous works [9], [10] have studies automatic label-word identification, their findings also showed that manual label-word selections still performed better. Besides, we mean to analyze the biases of PLMs. Thus, we introduce manually developed emotion models as our label-words and zero-shot without fine-tuning.

3 Methodology

There are two components in a typical prompt-based classification model, namely masked word prediction (§ 3.1) and label-word mappings (§ 3.2). An example of a graphical illustration of an examined prompt-based sentiment analysis model can be viewed in Fig. 1. The descriptions of employed variables are summarized in Table 1.

3.1 Masked word prediction

Given a sequence of input text ($t = \{t_1, t_2, ..., t_e\}$, where t with a subscript denotes a token), we concatenate ($\oplus$) the sequence after a prompt ($s = \{s_1, s_2, ...,[\text{MASK}]_{g-1}, s_g\}$, where s with a subscript denotes a token; [MASK] is a special token, defined by a PLM during its pre-training) as the input sequence (seq =[CLS] $\oplus s \oplus t \oplus$ [SEP], where [CLS] and [SEP] are special tokens) of a PLM. The prompt sequence (s) is manually defined, according to the coherence of employed label-words, e.g., "The feeling is [MASK]". We use the hidden state of [MASK] to predict the probability distribution ($P(\cdot)$) of a word (w) appearing in the [MASK] position.

TABLE 1
Variable descriptions.

Variable	Description
t	An original text sequence from a dataset.
s	A prompt sequence.
seq	An input sequence to the model.
M	The hidden states of a PLM.
H	The hidden states of the last FNN layer.
$h_{[\text{MASK}]}$	The hidden state corresponds to the [MASK] position of seq in H.
$\mathcal{S}$	A label-word set.
w	An emotion word in $\mathcal{S}$.
w^*	An emotion word in $\mathcal{S}$ with the highest probability.
$\hat{y}$	A predicted label.

$$M = PLM(seq) \tag{1}$$

$$H = FNN_2(Norm(GELU(FNN_1(M)))) \tag{2}$$

$$P(w) = SoftMax(h_{[\text{MASK}]}), \tag{3}$$

where PLM denotes the encoder (or encoder-decoder, if applicable) of a PLM which is a stack of Transformer layers [44]; FNN denotes a pre-trained feedforward neural network of the PLM; $GELU$ is an activation function [45]; $Norm$ denotes layer normalization [46]. M and H denote the hidden states of the encoder (or decoder in an encoder-decoder-based PLM) and the last FNN layer, respectively. $h_{[\text{MASK}]}$ denotes the hidden state in H, corresponding to the [MASK] position. The coding of masked word prediction is based on Huggingface Transformers [47] Python package.

Next, we develop a set ($\mathcal{S}$) of label-words, where each word corresponds to a label in our task. The label-word with the highest probability appearing in the [MASK] position is given by

$$w^* = arg \max_{w \in \mathcal{S}} P(w). \tag{4}$$

The predicted label ($\hat{y}$) is given by a manually defined label-word mapping function ($map(\cdot)$, see § 3.2 for details)

$$\hat{y} = map(w^*). \tag{5}$$

3.2 Label-word mappings

Recall that the predicted label of a prompt-based sentiment analysis classifier is given by the predicted most likely emotion word that appears in a context, and a label-word mapping. For example, in Fig. 1, the given sentence "this movie is very interesting" is identified as positive, because the PLM predicted probability of "joyful" is higher than that of other emotion words within a label-word set; Besides, "joyful" can be considered as a positive sentiment word, intuitively. Thus, an effective label-word set and label-word mappings can improve model performance.

TABLE 2
Emotion models and label-word mappings. The first two letters of a model denote the model name abbreviations; The following chunks denote the number of instances in a class, where C denotes coarse-grained emotion classes; F denotes fine-grained emotion classes. * denotes label-word mappings are given by our alignments, while the rest are given by the original papers.

Models	Basic Emotions	Emotion Categories	Label-word Mappings	
EK-C6 [12]	anger, disgust, fear, joy, sadness, surprise	-	ANGER	anger
			FEAR	fear
			SADNESS	sadness
			JOY	joy
			POSITIVE	joy
			NEGATIVE	anger, disgust, fear, sadness
			AMBIGUOUS*	surprise
GE-C6-F27 [13]	anger, disgust, fear, joy, sadness, surprise	admiration, amusement, approval, caring, anger, annoyance, disappointment, disapproval, confusion, desire, excitement, gratitude, joy, disgust, embarrassment, fear, grief, curiosity, love, optimism, pride, relief, nervousness, remorse, sadness, realization, surprise	ANGER	anger, annoyance, disapproval
			FEAR	fear, nervousness
			SADNESS	disappointment, embarrassment, sadness, grief, remorse
			JOY	admiration, amusement, approval, caring, desire, excitement, gratitude, joy, love, optimism, pride, relief
			POSITIVE	admiration, amusement, approval, caring, desire, excitement, gratitude, joy, love, optimism, pride, relief
			NEGATIVE	anger, annoyance, disappointment, disapproval, disgust, sadness, fear, grief, nervousness, remorse, embarrassment
			AMBIGUOUS	confusion, curiosity, realization, surprise
WE-C8-F32 [14]	joy, trust, fear, surprise, sadness, anticipation, anger, disgust	ecstasy, joy, serenity, love, admiration, trust, acceptance, submission, terror, fear, grief, apprehension, awe, amazement, surprise, distraction, disapproval, sadness, pensiveness, remorse, loathing, disgust, boredom, rage, contempt, anger, annoyance, aggressiveness, vigilance, interest, optimism	ANGER	rage, anger, annoyance
			FEAR	terror, fear, apprehension
			SADNESS	grief, sadness, pensiveness
			JOY	ecstasy, joy, serenity
			POSITIVE*	ecstasy, joy, serenity, admiration, trust, acceptance, love, optimism
			NEGATIVE*	terror, fear, apprehension, grief, sadness, remorse, disapproval, loathing, disgust, boredom, contempt, rage, anger, annoyance, pensiveness
			AMBIGUOUS*	aggressiveness, submission, awe, amazement, surprise, distraction vigilance, anticipation, interest
SR-C8-F23 [15]	joy, trust, fear, surprise, sadness, anticipation, anger, disgust	anger, annoyance, hostility, fury, anticipation, expectancy, interest, disgust, dislike, fear, apprehension, panic, terror, joy, happiness, elation, sadness, gloominess, grief, sorrow, surprise, trust, like	ANGER	anger, annoyance, hostility, fury
			FEAR	fear, apprehension, panic, terror
			SADNESS	sadness, gloominess, grief, sorrow
			JOY	joy, happiness, elation
			POSITIVE*	joy, happiness, elation, trust, like
			NEGATIVE*	anger, annoyance, hostility, fury, disgust, dislike, sadness, gloominess, grief, sorrow, fear, panic, terror, apprehension
			AMBIGUOUS*	anticipation, expectancy, interest surprise
HG-C8-F24 [16]	joy, sadness, calmness, anger, pleasantness, disgust, eagerness, fear	ecstasy, joy, contentment, terror, melancholy, sadness, grief, bliss, calmness, serenity, annoyance, anger, rage, delight, pleasantness, acceptance, dislike, disgust, fear loathing, enthusiasm, eagerness, responsiveness, anxiety	ANGER	annoyance, anger, rage
			FEAR	anxiety, fear, terror
			SADNESS	melancholy, sadness, grief
			JOY	ecstasy, joy, contentment
			POSITIVE	enthusiasm, eagerness, joy, responsiveness, bliss, calmness, serenity, ecstasy, contentment, delight, pleasantness, acceptance
			NEGATIVE	dislike, disgust, loathing, melancholy, sadness, grief, annoyance, anger, rage, anxiety, fear, terror
			AMBIGUOUS	-

TABLE 3
Dataset statistics. Seq len denotes the average length of sequences. # seq denotes the number of sequences.

Task	Label	Seq len	# seq	%
Sentiment Analysis	POS	112.1	4000	50.0
	NEG	114.2	4000	50.0
	All	113.1	8000	-
Emotion Detection	ANGER	13.5	1701	24.0
	FEAR	12.3	2252	31.6
	SADNESS	13.9	1533	21.6
	JOY	12.9	1616	22.8
	All	13.1	7102	-

TABLE 4
Examined pre-trained language models.

PLM	Version	Param. size (M)	Data size (GB)
BERT	small	108.3	16
	large	334.6	
RoBERTa	small	125.7	160
	large	355.4	
ALBERT	small	11.2	16
	large	206.4	
BART	small	139.4	160
	large	406.3	

We develop several sets of label-words and label-word mappings, according to classical emotion models [12], [13], [14], [15], [16] to test their effectiveness. Our hypothesis is that the lexicons of the taxonomy of an emotion model are eligible candidates for representing an emotion class. Sentiment polarities can be inferred from positive and negative emotion lexicons. We classify positive and negative emotion lexicons according to the recommendations of the original papers, e.g., Hourglass (HG) [16], GoEmotions (GE) [13]. For those emotion models that do not have the recommended sentiment polarities, e.g., Ekman's model (EK) [12], Plutchik's wheel of emotions (WE) [14] and the emotion model of Mohammad *et al.* for semantic role labeling (SR) [15], we manually classify them. First, we classify "surprise" and "anticipation" as ambiguous terms, because their sentiment polarities highly depend on their contexts. Thus, "surprise", "anticipation" and their similar emotions with different intensities, e.g, "expectancy" and "interest" are neither positive nor negative. Second, the combinations of positive and negative emotions, e.g., "submission" and "aggressiveness" in Plutchik's wheel of emotions are also ambiguous. "awe" is classified as ambiguous, because it is context-dependent as well. Finally, the rest of emotion words are classified as either positive or negative, because the sentiment polarities of these words can be directly identified, based on their literal meanings.

For emotion detection tasks, we employ a dataset, containing four label classes, namely anger, fear, sadness and joy. The emotion lexicons can precisely map to the emotion label classes according to the original papers. The final rules about the sentiment analysis and emotion detection label-word mappings ($map(\cdot)$ in Eq. 5) are shown in Table 2.

4 Experiment

4.1 Datasets

Sentiment Analysis. We employ a sentiment analysis dataset that was proposed by Blitzer, Dredze, and Pereira [48]. The data were sourced from Amazon product reviews about books, DVDs, electronics and kitchen appliances. The dataset contains two label classes, namely positive and negative. Since the maximum length of input sequences of our employed PLMs is equal to 512, we set up the maximum length of original text as 270[1], so that the tokenized sequence length given by different PLM tokenizers are still less than 512. We setup up a unique maximum length of the original text for fair performance comparison between different PLMs.

Emotion Detection. We employ an emotion detection dataset that was proposed on WASSA-2017 Shared Task on Emotion Intensity [49]. The data were sourced from Twitter, containing four emotion classes, e.g., anger, fear, sadness and joy. We remove emoji, hash tag and Twitter ID with "@" symbols from the original dataset during the pre-processing.

The detailed dataset statistics can be viewed in Table 3. Unlike the original emotion detection dataset that measures the Pearson correlation with gold intensity scores, we use accuracy as the evaluation measure in our sentiment analysis and emotion detection tasks to conduct performance variant comparison between the two tasks. Accuracy is employed because it is an intuitive measure that has been widely employed in previous affective computing works [41], [50]. Besides, the employed datasets are generally balanced.

4.2 Pre-trained Language Models

We examine the following PLMs, because they have different architectures, learning objects, and tokenizers. The detailed statistics about the number of parameters and the sizes of pre-training data are shown in Table 4. We examine both small and larger versions of each PLM in our experiments.

BERT [1] is a bi-directional encoder-based PLM, learning masked word prediction and next-sentence prediction as the pre-training objects. Input sequences are tokenized with WordPiece embeddings [51]. In our experiments, BERT small denotes BERT-base-cased. BERT large denotes BERT-large-cased.

RoBERTa [2] is also a bi-directional encoder-based PLM. It excludes the next-sentence prediction task of BERT, training with dynamic masking and large mini-batches. It uses Byte-Pair Encoding as a tokenizer [26]. In our experiments RoBERTa small denotes RoBERTa-base. RoBERTa large denotes RoBERTa-large.

ALBERT [3] is a lite BERT-like PLM. It employs factorized embedding parameterization and cross-layer parameter sharing techniques to significantly reduce the size of parameters. It takes masked word prediction and sentence-order prediction as the learning objects. The tokenizer is based on SentencePiece [52]. In our experiments ALBERT small denotes ALBERT-base-v2. ALBERT large denotes ALBERT-xxlarge-v2.

BART [4] is an encoder-decoder structured sequence-to-sequence PLM. The pre-training means to reconstruct the

[1] We test different tokenizers of PLMs with the employed data and find that the maximum of 270 tokens (before PLM tokenization) of original text can make sure that all tokenizers can yield tokenized sequences with no more than 512 tokens for our employed datasets.

TABLE 5
The comparison between different PLMs and emotion models. The performance is measured by accuracy. The performance of EK-C6 and GE-C6 and the performance of WE-C8 and SR-C8 are reported together because their label-word mappings (basic emotion classes) are the same. C4$^\sharp$ denotes that there are four basic emotion classes in the dataset, where the emotion classes are the subset of all coarse-grained emotion models.

Task	Emotion model	BERT		RoBERTa		ALBERT		BART		Avg.
		Small	Large	Small	Large	Small	Large	Small	Large	
Sentiment Analysis	EK-C6/GE-C6	50.0	50.0	67.1	64.1	62.0	62.7	53.5	62.5	59.0
	WE-C8/SR-C8	50.0	50.0	67.1	64.1	62.0	62.7	53.5	62.5	59.0
	HG-C8	64.2	72.2	68.8	69.2	63.0	62.9	61.4	69.8	66.4
	GE-F27	62.8	75.1	81.1	80.2	72.1	83.3	77.9	75.4	76.0
	WE-F32	62.2	76.0	82.0	84.3	76.5	70.8	73.6	70.9	74.6
	SR-F23	53.1	61.0	70.0	67.6	57.2	77.3	70.6	73.3	66.3
	HG-F24	63.0	72.0	79.8	84.7	77.3	71.2	63.9	75.2	73.4
Emotion Detection	C4$^\sharp$	31.9	33.3	38.5	39.6	31.2	33.7	33.5	30.0	34.0
	GE-F27	29.5	27.8	30.6	31.8	32.8	33.5	30.1	30.3	30.8
	WE-F32	33.4	34.4	36.9	35.9	32.7	35.5	35.3	30.9	34.4
	SR-F23	28.2	30.8	39.9	39.8	28.7	36.4	35.9	31.3	33.9
	HG-F24	32.7	34.3	38.2	40.9	34.3	37.4	35.5	35.6	36.1
Avg.		46.7	51.4	58.3	58.5	52.5	55.6	52.1	54.0	-

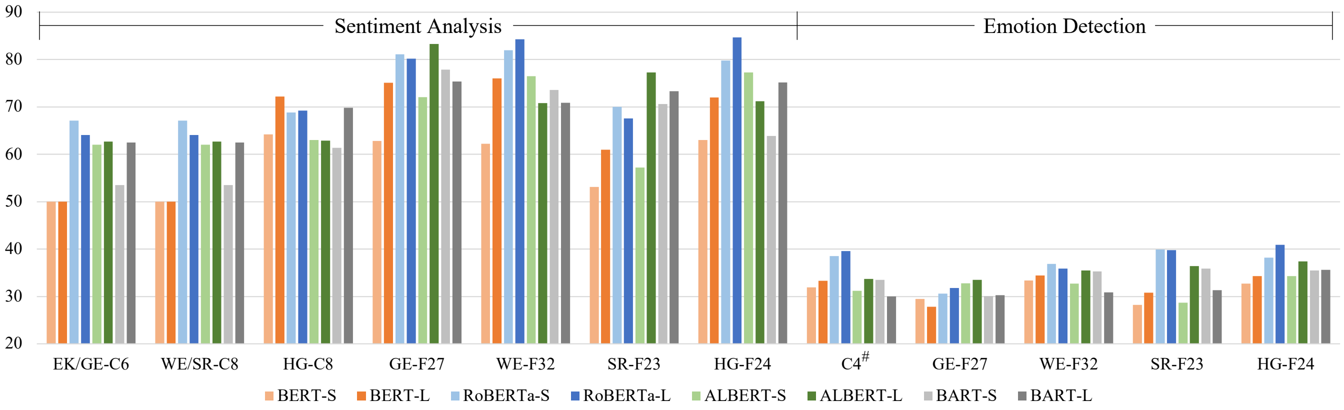


Fig. 2. The comparison between different PLMs and emotion models. -S denotes small. -L denotes large.

original text from corrupting text, including the learning objects such as token masking, token deletion, text infilling, sentence permutation, and document rotation. The tokenizer is also Byte-Pair Encoding. In our experiments, BART small denotes BART-base. BART large denotes BART-large.

5 Results

5.1 Overall performance

We first test the utilities of different PLMs, based on an intuitive prompt, “I feel [MASK]”. We add this prompt in front of an original input. Since the most coherent word form appearing in the [MASK] position of the prompt is an adjective, we manually translate the nominal emotion lexicons in Table 2 to their adjective forms.

As seen in Table 5, RoBERTa large with HG-F24 yields the best performance on sentiment analysis and emotion detection tasks. For the small versions of PLMs, RoBERTa small with WE-F32 and RoBERTa small with SR-F23 achieve better performance on sentiment analysis and emotion detection tasks, respectively. It shows that RoBERTa is more effective than other PLMs. ALBERT small achieves comparable performance to other small PLMs apart from RoBERTa small, whereas the number of parameters of ALBERT small is about $1/10$ of RoBERTa small.

We also observe that fine-grained emotion taxonomies likely outperform coarse-grained ones on sentiment analysis. Whereas, the advantage of fine-grained taxonomies is shallow on emotion detection. Overall, GE-F27 exceeds other emotion models on sentiment analysis, measured by the average accuracy over different PLMs, while HG-F24 performs better on emotion detection. This shows that fine-grained emotion models and the associated lexicons likely have higher chances to yield more accurate sentiment and emotion classification results.

However, we do not find an optimal emotion model that significantly surpasses other emotion models on the two examined tasks. Different PLMs likely have different emotion taxonomy preference.

5.2 The bias towards classification

By comparing the two examined tasks in Fig. 2, we find that the performance of emotion detection is much worse than the performance of sentiment analysis, although the two tasks are inferred by using the same emotion models, PLMs, and the prompt template. This implies that the emotion lexicons are more supportive on binary sentiment

TABLE 6
Breakdown analysis. Δ denotes the performance gap between the best and the worst classes.

PLM-large	Emotion model	Sentiment Analysis			Emotion Detection				
		POS	NEG	Δ	ANGER	FEAR	SADNESS	JOY	Δ
BERT	GE-F27	88.1	62.1	26.0	0.7	9.1	30.5	80.0	79.2
	WE-F32	63.8	88.3	24.5	29.5	79.7	9.6	0.0	79.7
	SR-F23	25.9	96.1	70.2	90.9	7.5	3.7	25.6	87.2
	HG-F24	58.8	85.2	26.4	3.0	84.5	31.4	0.0	84.5
	Avg.			36.8					82.7
RoBERTa	GE-F27	67.3	93.1	25.9	0.8	17.9	47.1	69.4	68.6
	WE-F32	83.5	85.2	1.6	18.0	32.9	10.0	83.6	73.6
	SR-F23	37.6	97.7	60.1	22.2	31.3	51.3	59.4	37.2
	HG-F24	73.9	95.5	21.6	3.1	76.9	21.3	49.1	73.8
	Avg.			27.3					63.3
ALBERT	GE-F27	76.7	89.8	13.1	15.0	7.0	63.3	61.5	56.3
	WE-F32	43.2	98.4	55.3	41.1	5.5	71.8	37.0	66.3
	SR-F23	79.2	75.5	3.7	38.2	5.2	64.1	51.6	58.9
	HG-F24	43.4	98.9	55.5	38.6	15.7	69.3	36.0	53.6
	Avg.			31.9					58.7
BART	GE-F27	54.5	96.4	41.9	4.5	29.4	48.3	41.7	43.8
	WE-F32	44.3	97.5	53.2	22.6	5.9	50.9	55.5	49.6
	SR-F23	68.3	78.3	10.0	22.3	5.5	48.8	60.1	54.6
	HG-F24	51.9	98.5	46.6	13.5	52.5	39.1	32.2	39.0
	Avg.			37.9					46.8

classification than emotion classification with four label classes. We conduct breakdown analysis to analyze such a phenomenon.

The breakdown analysis is based on the large versions of PLMs and fine-grained emotion taxonomies, because they achieve better performance than their alternatives in the previous subsection. In Table 6, we find that PLMs likely yield very weak performance to one or two emotion classes, based on different emotion taxonomies. However, the performance gap between positive and negative sentiments is much small. We use Δ to measure the gap between a label class with the highest accuracy and a label class with the lowest accuracy. As seen, the averaged Δ of sentiment analysis is lower than 38.0% across different PLMs, while the averaged Δ of emotion detection is higher than 46.0%. The averaged emotion detection Δ of BERT large even reaches 82.7%, whereas its sentiment analysis Δ is just 36.8%. The accuracy of many weak emotion classes (11 out of 16) is lower than 6.0%. The weak emotion classes are inconsistent in different PLMs and emotion models. Such a trend highlights that the masked word predictions of PLMs are biased for a fine-grained classification task.

Emotion detection is a fine-grained classification task. The reasons for yielding weak performance on emotion detection can be summarized as three aspects: (1) The emotion label space is divided into multiple classes, and each class may have less label-words. Emotion detection including sufficient label-words favored by PLMs becomes more accidental. Thus, its performance could be weaker than sentiment analysis. (2) The label space of emotion detection is not symmetrically divided. Thus, there are more non-opposing labels in a non-symmetrical label space. For example, a non-sadness emotion class can be one of "anger", "fear", and "joy" classes, whereas "anger" and "fear" are not the opposite emotion of "sadness". It increases the difficulty of distinguishing labels in vector space. (3) Most of the employed basic emotion taxonomies [12], [13], [14], [15] were developed according to facial expressions, rather than text. Their taxonomies are still arguable [11]. When the emotion detection comes to text, the textual emotion expressions could be ambiguous. People may show sadness or anger emotions towards the same context, e.g., the betrayal of a friend. PLMs do not particularly model the difference between different textual emotions during their pre-training. The co-occurrences of emotion words and contexts are subjective to individual feelings. Then, it is more difficult to infer the nuance between, e.g., "anger" and "sadness" from text in downstream prompting. In light of the above reasons, PLMs yield biased performance on emotion detection with fine-grained label classes.

On the other hand, the examined sentiment analysis task is a coarse-grained classification task, because its label space is symmetrically divided into two opposite classes, say either "positive" or "negative". A non-positive label is a negative label in the binary label classes, and vice versa. Thus, the label-words are more distinguishable between label classes in sentiment analysis. Besides, each label-class likely has more label-words in a coarse-grained classification task, which means the selected label-words more likely cover the preference of different PLMs.

Unbalanced label-words also result in biased predictions between different classes. For example, GE-F27 achieves better performance on the "sadness" and "joy" classes than other classes across all PLMs, while such consistent status does not appear in other emotion models. This is because "sadness" and "joy" classes of GE-F27 have more label-words (see Table 2) than other classes. Thus, the number of label-words triggers the biased predictions of PLMs.

5.3 The bias towards word frequencies

Intuitively, humans likely use coarse-grained emotion words more often than fine-grained ones, because the emotion intensity towards a specific circumstance can be more subjective and different than the basic emotion intention between individuals, resulting the fact that some emotional lexicons, e.g., "sad" and "angry" more frequently appear

TABLE 7
The Pearson's correlation coefficients (r) between the average predicted probabilities of PLMs and word frequencies.

PLM	Version	r	p-value
BERT	Small	-0.016	0.908
	Large	-0.028	0.844
RoBERTa	Small	-0.039	0.780
	Large	-0.033	0.814
ALBERT	Small	0.583	0.000
	Large	0.337	0.013
BART	Small	0.043	0.760
	Large	0.198	0.156

in corpora than others. Thus, we verify if PLMs are biased to frequent label-words, yielding higher probabilities for them. We test the correlation co-efficient between label-word frequencies in a corpus and the average probability of a word appearing the [MASK] position. Whereas, the results are negative for large PLMs.

We use Google Books Ngram corpus[2] to obtain the statistics[3] of word frequencies. Totally, there are 894.3 billion tokens and 7.9 million unique tokens in the corpus, which is big enough to represent word distributions in everyday language. We use f_j to denote the frequency of a word (w_j) and Pearson's correlation coefficient (r) as a measure to evaluate, if PLMs bias to frequent words. We average the predicted probability distribution ($P(\cdot)$) of our examined emotion lexicons ($w_j, j \in \{1, ..., m\}$, including course-grained and fine-grained taxonomies; m denotes the total number of emotion lexicons) over all n input sequences in the combination of sentiment and emotion datasets, then computing Pearson's correlation coefficient (r_i) of a PLM ($PLM_i, i \in \{BERT_s$, $BERT_l$, $RoBERTa_s$, $RoBERTa_l$, $ALBERT_s$, $ALBERT_l$, $BART_s$, $BART_l\}$) by

$$p_{i,j} = \frac{1}{n}\sum_{k=1}^{n} P(w_j \mid seq_k, PLM_i) \tag{6}$$

$$r_i = \frac{\sum_{j=1}^{m}(p_{i,j} - \frac{1}{m}\sum_{j=1}^{m} p_{i,j})(f_j - \frac{1}{m}\sum_{j=1}^{m} f_j)}{\sqrt{\sum_{j=1}^{m}(p_{i,j} - \frac{1}{m}\sum_{j=1}^{m} p_{i,j})^2}\sqrt{\sum_{j=1}^{m}(f_j - \frac{1}{m}\sum_{j=1}^{m} f_j)^2}}. \tag{7}$$

$p_{i,j}$ denotes the probability of w_j appearing in the [MASK] position, given by PLM_i over n sequences, where each sequence (seq_k) contains an original text and a prompt.

The correlation coefficients (r) are shown in Table 7. The highest r appears in ALBERT small ($r = 0.58, p-value < 0.005$). For other PLMs, we cannot find clear evidence to prove that there is a strong correlation between the probability of a word appearing in a context and the word frequency in a typical corpus. Thus, large PLMs show the capacity of modeling the contextual information, mitigating the impact of word frequencies.

5.4 The bias towards prompt and label-word forms

We examine the biases of prompt-based models by modifying prompt templates, positions, the PoS of label-words. Apart from "I feel [MASK]" (Prompt 1), we develop another prompt template "The emotion is [MASK]" (Prompt 2). Both nouns and adjectives can appear in the [MASK] position of Prompt 2. Thus, we also test nominal and adjective emotion words based on the two prompt templates, respectively. In the previous subsections, Prompt 1 was placed in front of an original input. We define this prompt position as "head". Here, we additionally examine the performance difference by connecting a prompt after an original input, say "tail".

As seen in Table 8, different prompt templates yield different performance, although other variables are controlled and the two prompts are semantically similar. For example, given adjective label-words, the gap of average performance between Head 1 and Head 2 can be larger than 2.1% on sentiment analysis and emotion detection tasks. In the comparison between nominal and adjective label-words, it is unsurprising to see that Head 1 yields better performance on adjectives than nouns (+6.7% in sentiment analysis and +7.8% in emotion detection), because it is incoherent that nouns appear in the context of "I feel [MASK]". Thus, the coherence of label-words and contexts does impact the prompting significantly.

However, for Prompt 2, we also observe adjectives likely yield better performance in Head 2 (+4.6% in sentiment analysis and +2.3% in emotion detection). This is biased, because both a noun and its corresponding adjective should semantically represent the same emotion in a coherent context, whereas a PLM has different performance towards label-words with different PoS. In practice, the nuance in word forms results in different performance, which leads more challenges in label-word selections and evaluations.

If we use both adjective and nominal forms of label-words, the models likely achieve more accurate results. It seems that increasing the vocabulary size of label-words by using fine-grained emotion taxonomies and different PoS of emotion words is a way of boosting the performance of a none-tuning prompt-based model. It also presents a challenge that for a prompt-tuning-based classification model, how the model can achieve global optimum with dynamic label-word selections in the context that there are many label-word candidates.

Finally, we observe that different positions of prompts deliver different results, although other variables are controlled. This is also biased, because cognitively, different prompt positions do not change the sentiment polarities and emotion intention of an original input. Thus, one may take these variables into account in order to achieve the optimal performance in prompt-based classification.

6 Future work

Given the above findings about the biases of PLMs, we leave several research questions about prompt-based classification and affective computing for future work:

(1) How to achieve an effective label-word selection? Since PLMs likely yield unbalanced performance in many-class classification, it is important to find a way to identify efficient label-words for each class. We did not find strong correlation coefficient between predicted label-word probabilities and word frequencies, which makes a label-word selection particularly challenging. Sometimes, one cannot find an appropriate label-word for a label class simply

[2] https://books.google.com/ngrams
[3] https://github.com/hackerb9/gwordlist

TABLE 8
The comparison between different prompt templates, positions, label-word PoS, across different PLMs. The performance is based on HG-F24 emotion model. "Both" denotes that both nominal and adjective word forms are included in a label-word mapping set.

Task	PLM	Head 1			Head 2			Tail 1			Tail 2		
		Noun	Adj	Both	Noun	Adj	Both	Noun	Adj	Both	Noun	Adj	Both
Sentiment Analysis	BERT-S	56.0	63.0	63.1	54.8	59.7	60.8	51.7	62.6	62.6	53.6	61.9	63.3
	BERT-L	53.1	72.0	71.9	58.3	66.3	68.1	52.4	69.3	69.6	59.3	68.1	69.1
	RoBERTa-S	73.6	79.8	79.4	80.7	83.4	83.5	75.6	78.4	78.2	80.1	84.4	84.5
	RoBERTa-L	71.2	84.7	84.9	83.6	84.8	84.9	73.7	81.0	80.9	83.3	83.7	83.8
	ALBERT-S	66.4	77.3	77.3	74.4	77.6	77.6	70.4	77.2	77.2	75.6	80.1	80.1
	ALBERT-L	69.3	71.2	71.3	65.9	77.7	74.6	76.5	77.6	77.7	70.5	76.3	75.0
	BART-S	67.9	63.9	64.3	75.6	75.1	78.1	71.1	63.0	63.7	79.4	77.9	79.6
	BART-L	75.6	75.2	75.5	73.7	79.3	78.9	76.2	67.2	68.0	76.5	80.7	80.7
	Avg.	66.6	73.4	73.5	70.9	75.5	75.8	68.5	72.0	72.2	72.3	76.6	77.0
Emotion Detection	BERT-S	25.0	32.7	32.9	24.2	28.3	27.9	25.7	32.4	32.8	26.6	26.3	27.6
	BERT-L	24.6	34.3	34.0	27.2	29.7	30.5	25.7	33.8	33.1	28.7	27.3	28.3
	RoBERTa-S	27.5	38.2	38.6	28.6	32.0	32.5	29.0	39.2	39.5	29.6	32.9	33.7
	RoBERTa-L	27.7	40.9	41.5	31.9	30.1	31.6	29.5	40.1	40.3	33.6	29.6	31.7
	ALBERT-S	33.2	34.3	34.3	28.8	28.5	28.9	35.6	35.4	35.5	31.2	29.1	29.6
	ALBERT-L	32.2	37.4	37.3	30.9	38.5	35.0	36.1	37.1	37.9	31.5	36.8	34.6
	BART-S	28.0	35.5	36.8	27.3	35.2	33.9	29.8	32.3	33.8	28.8	33.7	33.9
	BART-L	28.4	35.6	35.4	33.1	27.9	28.9	30.1	36.7	35.9	34.3	28.0	28.8
	Avg.	28.3	36.1	36.3	29.0	31.3	31.1	30.2	35.9	36.1	30.5	30.5	31.0

based on the semantic similarities, because the label class may have complex semantic information. For example, in biomedical relation extraction [53], a label class may represent a complex interaction description between two genes. It would be difficult to select effective label-words in this type of classification tasks. Since the nuance of label-words, e.g., PoS and the coherence in the context does impact prompting results, a dynamic label-word selection method would be valuable for a learning class that has many label-word candidates achieving global optimum. For emotion detection tasks, using existing emotion models and their associated emotion lexicons does not present satisfied performance, based on our examined PLMs. It would be valuable to develop an efficient emotion model whose emotion lexicons are distinguishable between different emotion classes in text domain to balance the performance on different classes. This may also benefit sentiment analysis tasks, because one can use these emotion lexicons as the label-words for positive and negative sentiment polarity predictions.

(2) How to generate an effective prompt accordingly? There are many variables that may effect the utility of a prompt template, e.g., positions, the selected PLM, label-words, and the PoS of label-words. It is valuable to explore a method to generate prompts, according to these variables. In this work, we demonstrate the performance gap, based on large annotated emotion and sentiment classification data. However, for few- and zero-learning tasks, it is hard to evaluate, whether a prompt is better than its competitors. It is necessary to find a method to evaluate the utility of a prompt, according to the interaction with input sequences on a PLM, rather than testing on a target downstream task.

7 Conclusion

In this work, we conduct an empirical study on prompt-based sentiment analysis and emotion detection to analyze the biases of pre-trained language models (PLMs). Our experimental results show that RoBERTa surpasses other examined PLMs (BERT, ALBERT and Bart) in sentiment analysis and emotion detection. Using fine-grained emotion taxonomies as label-words is more effective than coarse-grained ones in the examined affective computing tasks. Using multiple word forms of the label-words also yields better performance, although adjective emotion words are more effective than their nominal counterparts. On the other hand, we find that PLMs are biased to some label classes in fine-grained classification tasks, e.g., four-class emotion detection, yielding very weak performance. The PLM biases are also manifested in the variables, such as label-word selections, the PoS of label-words, prompt templates and positions. These variables seemingly have no influence on human sentiment and emotion predictions, whereas for PLMs, they yield different classification results. One may develop better label-word selection and prompt engineering methods to mitigate the biases to achieve better results. Ensemble learning is also a possible way to take complementary advantages of different PLMs [54], [55], [56].

Acknowledgments

This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019.

[2] 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," *arXiv preprint arXiv:1907.11692*, 2019.

[3] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=H1eA7AEtvS

[4] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880.

[5] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *arXiv preprint arXiv:2107.13586*, 2021.

[6] R. Mao, C. Lin, and F. Guerin, "Word embedding and WordNet based metaphor identification and interpretation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1222–1231.

[7] R. Mao, X. Li, M. Ge, and E. Cambria, "MetaPro: A computational metaphor processing model for text pre-processing," *Information Fusion*, vol. 86-87, pp. 30–43, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253522000550

[8] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 3816–3830. [Online]. Available: https://aclanthology.org/2021.acl-long.295

[9] T. Schick, H. Schmid, and H. Schütze, "Automatically identifying words that can serve as labels for few-shot text classification," in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 5569–5578. [Online]. Available: https://aclanthology.org/2020.coling-main.488

[10] T. Schick and H. Schütze, "Exploiting cloze-questions for few-shot text classification and natural language inference," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr. 2021, pp. 255–269. [Online]. Available: https://aclanthology.org/2021.eacl-main.20

[11] K. Crawford, *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, 2021. [Online]. Available: http://www.jstor.org/stable/j.ctv1ghv45t

[12] P. Ekman, "Expression and the nature of emotion," *Approaches to emotion*, vol. 3, no. 19, p. 344, 1984.

[13] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A dataset of fine-grained emotions," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4040–4054.

[14] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of emotion*. Elsevier, 1980, pp. 3–33.

[15] S. Mohammad, X. Zhu, and J. Martin, "Semantic role labeling of emotions in tweets," in *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2014, pp. 32–41.

[16] Y. Susanto, A. G. Livingstone, B. C. Ng, and E. Cambria, "The hourglass model revisited," *IEEE Intelligent Systems*, vol. 35, no. 5, pp. 96–102, 2020.

[17] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237. [Online]. Available: https://aclanthology.org/N18-1202

[18] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[19] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[20] Q. Liu, X. Geng, H. Huang, T. Qin, J. Lu, and D. Jiang, "Mgrc: An end-to-end multigranularity reading comprehension model for question answering," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[21] J. Cai, Z. Zhu, P. Nie, and Q. Liu, "A pairwise probe for understanding bert fine-tuning on machine reading comprehension," in *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 1665–1668.

[22] R. Mao, C. Lin, and F. Guerin, "End-to-end sequential metaphor identification inspired by linguistic theories," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3888–3898.

[23] ——, "Combining pre-trained word embeddings and linguistic features for sequential metaphor identification," *arXiv preprint arXiv:2104.03285*, 2021.

[24] R. Puri and B. Catanzaro, "Zero-shot text classification with generative language models," *arXiv preprint arXiv:1912.10165*, 2019.

[25] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," *arXiv preprint arXiv:2012.15723*, 2020.

[26] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[27] Q. Ye, Z. Zhang, and R. Law, "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches," *Expert systems with applications*, vol. 36, no. 3, pp. 6527–6535, 2009.

[28] V. K. Jain, S. Kumar, and S. L. Fernandes, "Extraction of emotions from multilingual text using intelligent text processing and computational linguistics," *Journal of computational science*, vol. 21, pp. 316–326, 2017.

[29] W. Li, K. Guo, Y. Shi, L. Zhu, and Y. Zheng, "DWWP: Domain-specific new words detection and word propagation system for sentiment analysis in the tourism domain," *Knowledge-Based Systems*, vol. 146, pp. 203–214, 2018.

[30] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for twitter sentiment classification," in *Proceedings of Annual Meeting of the Association for Computational Linguistics*, 2014, pp. 1555–1565.

[31] Q. Liu, H.-Y. Huang, Y. Gao, X. Wei, Y. Tian, and L. Liu, "Task-oriented word embedding for text classification," in *Proceedings of the 27th international conference on computational linguistics*, 2018, pp. 2023–2032.

[32] Q. Liu, J. Lu, G. Zhang, T. Shen, Z. Zhang, and H. Huang, "Domain-specific meta-embedding with latent semantic structures," *Information Sciences*, vol. 555, pp. 410–423, 2021.

[33] W. Li, W. Shao, S. Ji, and E. Cambria, "BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis," *Neurocomputing*, vol. 467, pp. 73–82, 2022.

[34] H. Xu, B. Liu, L. Shu, and S. Y. Philip, "BERT post-training for review reading comprehension and aspect-based sentiment analysis," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 2324–2335.

[35] A. F. Adoma, N.-M. Henry, and W. Chen, "Comparative analyses of BERT, RoBERTa, DistilBERT, and XLNet for text-based emotion recognition," in *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*. IEEE, 2020, pp. 117–121.

[36] R. Mao and X. Li, "Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 15, 2021, pp. 13 534–13 542.

[37] C. Strapparava and A. Valitutti, "WordNet affect: an affective extension of WordNet," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, 2004.

[38] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010.

[39] C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, 2014, pp. 216–225.

[40] B. Liang, H. Su, L. Gui, E. Cambria, and R. Xu, "Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks," *Knowledge-Based Systems*, vol. 235, p. 107643, 2022.

[41] E. Cambria, Q. Liu, S. Decherchi, F. Xing, and K. Kwok, "SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis," *Proceedings of LREC 2022*, pp. 3829–3839, 2022.

[42] E. Cambria, A. Livingstone, and A. Hussain, "The hourglass of emotions," in *Cognitive behavioural systems*. Springer, 2012, pp. 144–157.

[43] K. Hambardzumyan, H. Khachatrian, and J. May, "WARP: Word-

level adversarial reprogramming," *arXiv preprint arXiv:2101.00121*, 2021.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[45] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[46] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[47] T. Wolf, J. Chaumond, L. Debut, V. Sanh, C. Delangue, A. Moi, P. Cistac, M. Funtowicz, J. Davison, S. Shleifer *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.

[48] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," in *Proceedings of the 45th annual meeting of the association of computational linguistics*, 2007, pp. 440–447.

[49] S. M. Mohammad and F. Bravo-Márquez, "WASSA-2017 shared task on emotion intensity," in *8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis WASSA 2017: Proceedings of the Workshop*. The Association for Computational Linguistics, 2017, pp. 34–49.

[50] 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 *Proceedings of the 29th ACM international conference on information & knowledge management*, 2020, pp. 105–114.

[51] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

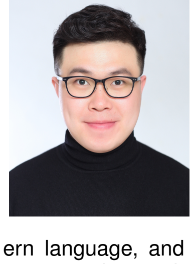
[52] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2018, pp. 66–71.

[53] T. Gao, X. Han, H. Zhu, Z. Liu, P. Li, M. Sun, and J. Zhou, "FewRel 2.0: Towards more challenging few-shot relation classification," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6250–6255. [Online]. Available: https://aclanthology.org/D19-1649

[54] D. Guru, M. Suhil, S. Pavithra, and G. Priya, "Ensemble of feature selection methods for text classification: An analytical study," in *International Conference on Intelligent Systems Design and Applications*. Springer, 2017, pp. 337–349.

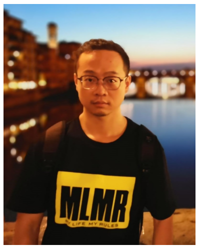
[55] B. Ray, A. Garain, and R. Sarkar, "An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews," *Applied Soft Computing*, vol. 98, p. 106935, 2021.

[56] R. Xiang and S. Yin, "Semi-supervised text classification with temporal ensembling," in *2021 International Conference on Computer Communication and Artificial Intelligence (CCAI)*. IEEE, 2021, pp. 204–208.

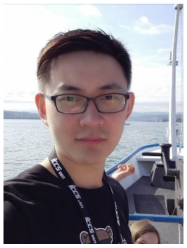


Rui Mao received his Ph.D. in Computing Science from the University of Aberdeen in 2020. Currently, he is a Research Fellow in Nanyang Technological University, studying conversational AI. His research interests include computational metaphor processing, token-level and sequence-level semantic understanding, and affective computing. He and his founded company have developed the first neural network search engine (www.wensousou.com) for searching ancient Chinese poems by using modern language, and a system (metapro.ruimao.tech) for linguistic and conceptual metaphor understanding. He has published several papers as the first author in top-tier international conferences and journals, e.g., ACL, AAAI, and Information Fusion. He served as Area Chair in COLING and EMNLP, and Reviewer in KBS, COGN COMPUT, and INFFUS.

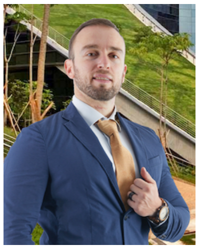
Qian Liu is a PostDoc Research Fellow in Nanyang Technological University, Singapore. She got her Ph.D. degree in computer science from Beijing Institute of Technology (in 2020) and University of Technology Sydney (in 2021). Her research interests include natural language processing and information retrieval. She has published several papers in international conferences such as WWW, AAAI, COLING, ect, and top-tier journals such as IEEE Transaction on Knowledge and Data Engineering (TKDE), IEEE Transactions on Neural Networks and Learning Systems (TNNLS), and IEEE Transaction on Fuzzy Systems (TFS).



Kai He is a PhD candidate at the school of Computer Science and Technology, Xi'an Jiaotong University, China, under the supervision of Prof. Chen Li. Currently, he is on an academic visit to the school of Computer Science and Engineering in Nanyang Technological University under the supervision of Prof. Erik Cambria. His research interests include information extraction and sentiment analysis in natural language processing field.



Wei Li received his Bachelor of Industrial Engineering from Shanghai Jiao Tong University in 2015. After that, he obtained his Master of Management Science and Engineering from University of Chinese Academy of Sciences in 2018. He enrolled as a PhD student under the supervision of Prof. Erik Cambria at NTU SCSE in 2019. His main research interests include sentiment analysis, natural language processing and deep learning. In particular, he is working on the construction of advanced deep learning models for sentiment analysis.



Erik Cambria (Fellow, IEEE) is the Founder of SenticNet, a Singapore-based company offering B2B sentiment analysis services, and a Professor at NTU, where he also holds the appointment of Provost Chair in Computer Science and Engineering. Prior to joining NTU, he worked at Microsoft Research Asia (Beijing) and HP Labs India (Bangalore) and earned his PhD through a joint programme between the University of Stirling and MIT Media Lab in 2012. His research focuses on neurosymbolic AI for explainable natural language processing in domains like sentiment analysis, dialogue systems, and financial forecasting. He is recipient of several awards, e.g., IEEE Outstanding Career Award, was listed among the AI's 10 to Watch, and was featured in Forbes as one of the 5 People Building Our AI Future. He is Associate Editor of many top-tier AI journals, e.g., INFFUS and IEEE TAFFC, and is involved in various international conferences as program chair, SPC member and invited speaker.