Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/55/20241584

The future development direction of natural language processing from the perspective of text emotion analysis

Jie Chen[1], Zicheng Wang[2,3]

[1]School of Electronics and Information Engineering, Guangdong University of Petrochemical Technology, Maoming, 525000, China
[2]Department of Computer Science and Technology, Zhuhai College of Science and Technology, Zhuhai, 519040, China

[3]zichengwang@stu.zcst.edu.cn

Abstract. The development of natural language processing is significant for text emotion analysis because it helps to understand the expression of human emotions in different contexts and provides more accurate semantic understanding and emotion recognition capabilities for intelligent systems. In current natural language processing, sentiment analysis has become a key research field, and it is devoted to developing more accurate and efficient sentiment recognition models to adapt to the growing data scale and semantic complexity. This paper focuses on an overview of contemporary text emotion analysis technology and looks forward to the future development of natural language processing. This paper makes a detailed comparative analysis of the efficiency of different emotion analysis methods from the perspectives of key length, research content, research methods, and results. In the review, the advantages and limitations of various emotion analysis methods will be discussed in detail, including transformer-based and a series of the latest technologies. In addition, the performance differences of different methods of processing large-scale text data will be analyzed in-depth, and their performance in practical applications will be comprehensively evaluated. Finally, the research will discuss the possible future direction of natural language processing in emotion analysis in combination with current research trends and technology development trends to provide helpful enlightenment and guidance for researchers and practitioners in this field.

Keywords: Sentiment analysis, transformer, natural language processing.

1. Introduction

As an important branch of natural language processing, emotion analysis aims to dig deeply into the emotion and emotion information behind the text and conduct in-depth research on human emotion cognition, so it has crucial research significance and broad application prospects. Its origin can be traced back to the early 1990s. The burgeoning growth of social media and network information has catapulted emotion analysis into a forefront research domain. Research shows that sentiment analysis has been widely used in social media public opinion monitoring, product review analysis, public opinion trend prediction, and many other fields, with fruitful results. Internationally, scholars are committed to developing effective analysis models based on deep learning, such as Transformer, (Bidirectional Encoder Representations from Transformers) BERT, and have made remarkable research progress.

DOI: 10.54254/2755-2721/55/20241584

Researchers have also actively explored emotion analysis technology based on the Chinese context and carried out much proper empirical research. In general, emotional analysis research is of great significance for understanding human emotional cognition, social public opinion trends, and product market feedback and helps to improve the scientificity and accuracy of public opinion management and product marketing strategy formulation.

This paper uses the literature research method to systematically summarize the application of Transformer and its subsequent derivative models in emotion analysis. Through the compilation and examination of pertinent literature, this study evaluates the efficacy of various models in sentiment analysis, delves into the benefits and drawbacks of centralized models, and suggests avenues for future research and potential obstacles. The study will focus on each model's performance difference, application scenario adaptation, and technical innovation when dealing with emotion analysis tasks to provide comprehensive and in-depth reference and guidance for researchers in this field.

2. Theory elaboration

2.1. Transformer

The Transformer is a revolutionary deep learning model based on the self-attention mechanism, initially proposed by Vaswani et al. in 2017. Serving as a sequence-to-sequence learning model, it provides powerful tools for natural language processing tasks. In contrast to traditional recurrent neural networks (RNNs) and convolutional neural networks (CNNs), the Transformer introduces the self-attention mechanism, empowering the model to establish long-distance dependencies within sequences. The model's encoder-decoder structure, robust parallel computing capabilities, proficiency in handling long-distance dependencies, and excellent model interpretability have collectively contributed to its remarkable achievements in the field of natural language processing. The advent of the Transformer not only spurred advancements in natural language processing but also left an indelible mark on tasks including machine translation, text generation, and language modeling. Numerous subsequent models, including BERT and GPT, have drawn inspiration from the basic architecture and attention mechanism of the Transformer. This underscores that the success of the Transformer lies not only in its outstanding performance but also in its role as a influential reference and source of inspiration for subsequent models. Figure 1 vividly illustrates the fundamental structure of the Transformer, marking it as a milestone within the realm of natural language processing.

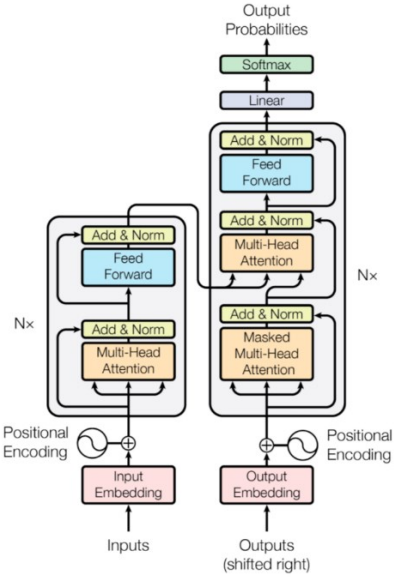

Figure 1. The Transformer - model architecture [1].

DOI: 10.54254/2755-2721/55/20241584

2.2. BERT

BERT, an innovation built upon the Transformer architecture introduced by Google in 2018[2], has brought about a paradigm shift in natural language processing (NLP). Departing from traditional unidirectional language models, BERT adopts a bidirectional training strategy, engaging in pre-training on extensive text corpora to grasp comprehensive language representations. Throughout the pre-training phase, BERT adeptly captures contextual nuances, generating deep-level representations for individual words. In practical applications, BERT proves invaluable as a fine-tuning model for specific tasks like text classification and named entity recognition, showcasing its versatility. This model's widespread adoption has played a pivotal role in propelling the field of NLP forward, establishing itself as a widely recognized benchmark.

The pre-training process entails unsupervised learning within a bidirectional Transformer structure, allowing simultaneous consideration of context within text sequences. BERT's innovation is further underscored by the introduction of the masked language model task, a mechanism aiding the model in contextual understanding through predicting masked words. The bidirectional context encoding, facilitated by the multi-layer encoder stack in Transformer, ensures a comprehensive grasp of both left and right context, thereby enhancing its capacity for understanding nuanced emotional expressions. The transformative impact of BERT reverberates across various NLP applications, marking it as a seminal advancement in the field.

2.3. A Lite BERT (ALBERT)

ALBERT conceived as a lightweight BERT model tailored for self-supervised language representation learning, integrates two innovative parameter reduction techniques to expedite training and economize memory usage [3]. Moreover, ALBERT employs a self-supervised loss function that underscores sentence coherence, consistently proving advantageous for tasks involving multi-sentence input.

Through strategic parameter sharing, including cross-layer parameter sharing, ALBERT effectively balances maintaining high performance while significantly reducing the overall number of parameters when compared to BERT. This inherent efficiency and scalability position ALBERT as an appealing choice, achieving performance levels comparable to or surpassing larger BERT models, all while demonstrating increased resource efficiency and mitigating the risk of overfitting.

2.4. Text-to-Text Transfer Transformer (T5)

Developed by the Google Research Institute, T5 leverages the transformer architecture to streamline transfer learning in natural language processing (NLP) tasks [4]. Trained on a substantial corpus of unlabeled text data extracted from the web, with particular emphasis on the general crawl dataset, the model adopts an encoder-decoder architecture.

Primarily tailored for NLP applications, T5 excels in tasks involving text generation and classification. Its training methodology encompasses the transformation of input text into target text, fostering expertise in tasks such as summarization, translation, and question-answering. Furthermore, T5 demonstrates efficacy in text classification applications, including emotion analysis and text matching.

With broad applicability across diverse NLP tasks, T5 proves valuable in text generation, classification, summarization, machine translation, and various other domains. Its versatility and robust performance position it as a prominent choice for a wide spectrum of language processing applications.

2.5. A Robustly Optimized BERT Pretraining Approach (RoBERTa)

RoBERTa, an advanced iteration of the BERT model introduced by Facebook AI in 2019 [5], is dedicated to enhancing the model's generalization ability and overall performance through meticulous optimization of pre-training strategies and training parameters. By employing a larger batch size, extended training duration, and a dynamic mask strategy, RoBERTa maximizes the utilization of extensive unsupervised text data. Noteworthy is RoBERTa's decision to discard the NSP (Next Sentence Prediction) task from the original BERT model, opting instead for a larger batch size and extended

DOI: 10.54254/2755-2721/55/20241584

training time. These strategic optimizations synergistically result in improved performance across a spectrum of natural language processing tasks, encompassing text classification, question-answering systems, and language inference. The introduction of RoBERTa serves as a catalyst for the progress of pre-training models in natural language processing, solidifying its position as an indispensable language representation learning model in contemporary research.

3. Comparison of experimental theories and analysis of advantages and disadvantages

This part will summarize the papers with higher impact factors in the three directions of binary classification, fine-grained affective analysis, and affective analysis on IMDB.

3.1. Emotional analysis of binary classification

In the paper "Muppet: Large-Scale Multitask Representation with Pre-Fine-Tuning," a pioneering stage known as "pre-fine-tuning" is introduced, serving as a crucial link between the fine-tuning and pre-training phases. This groundbreaking methodology is designed to elevate the model's generalization performance by integrating extensive multitask learning steps.[6]. Pre-fine-tuning includes multi-task learning on about 50 different tasks, with a total of 4.8 million training examples to encourage the model to better generalize on different tasks [6]. The author points out that the standard multi-task learning scheme is usually unstable, and proposes a new training scheme, which uses loss scaling and task heterogeneous batches to balance the gradient steps between different tasks, to improve the stability and performance of training. The author calls the model at this stage "MUPPET", and shows that combining pre-tuning with RoBERTA and BART models can achieve consistent improvement on multiple tasks without specific intermediate transmission tasks. The experimental results show that pre-tuning has apparent advantages, especially in the case of resource shortage. The research also emphasized the importance of optimization technology, task number, and scale for effective multi-task learning, providing strong experimental evidence for the in-depth understanding of critical factors in the pre-fine-tuning stage. Table 1 shows the results of the GLUE benchmark task and MRC dataset.

Table 1. Results of the GLUE benchmark task and MRC dataset [6].

	GLUE						MRC
	MNLI	QQP	RTE	QNLI	MRPC	SST-2	SQuAD
RoBERTa-B	87.6	91.9	78.7	92.8	90.2	94.8	82.6
+ MUPPET	88.1	91.9	87.8	93.3	91.7	96.7	86.6
RoBERTa-L	90.2	92.2	88.1	94.7	90.9	96.4	88.7
+MUPPET	90.8	92.2	92.8	94.9	91.4	97.4	89.4
BART	89.9	92.5	87.0	94.9	90.4	96.6	
+MUPPET	89.9	92.7	92.4	94.6	92.2	96.9	
ELECTRA-B	88.8	91.5	82.7	93.2	89.5	95	80.5
ELECTRA-L	90.9	92.4	88.0	95.0	90.8	96.9	88.1
MT-DNN	87.1	91.9/89.2	83.4	92.9	91.0/87.5	94.3	

As shown in Table 1, on the GLUE benchmark task and MRC dataset, the pre-training representation of pre-fine tuning is significantly better than the standard pre-training representation. For MRC tasks, the author reports the exact match (EM) and F1 indicators. For SQuAD, the research uses the pre-fine-tuning task head to repeat the experiment. The experimental results show that the improvement is relatively gentle on larger datasets, while it shows significant improvement on smaller datasets. For example, on RTE tasks, the pre-fine-tuned RoBERTA BASE model has improved by nearly 9 points, reaching the accuracy level equivalent to RoBERTA Large; The RoBERTA Large model with pre-fine-tuning can be comparable to the model with a scale of one order of magnitude in RTE tasks.

In ALBERT: a model of self-supervised learning of language representation, ALBERT integrates two parameter reduction technologies, namely cross-layer parameter sharing and decomposition

DOI: 10.54254/2755-2721/55/20241584

embedding parameterization, which effectively overcomes the main obstacles to expanding the pre-training model [3]. By implementing cross-layer parameter sharing and decomposing the vocabulary embedding matrix, ALBERT significantly reduces the number of parameters, improves parameter efficiency, and does not seriously affect performance. Compared with the BERT target, the number of parameters configured by ALBERT is reduced by 18 times, and the training speed is increased by about 1.7 times [3]. In addition, to further improve the performance of ALBERT, the research introduces sentence order prediction (SOP) self-monitoring loss, focusing on maintaining the coherence between sentences. The experimental results show that after these design decisions, ALBERT has achieved new and advanced results in several natural language understanding benchmark tests.

Table 2. The most advanced results of GLUE [3].

Models	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS	WNLI	Avg
Single-task single models on dev										
BERT-large	86.6	92.3	91.3	70.4	93.2	88	60.6	90	- -	- -
XLNet-large	89.8	93.9	91.8	83.8	95.6	89.2	63.6	91.8	- -	- -
RoBERTa-large	90.2	94.7	92.2	86.6	96.4	90.9	68	92.4	- -	- -
ALBERT (1M)	90.4	95.2	92	88.1	96.8	90.2	68.7	92.7	- -	- -
ALBERT (1.5M)	90.8	95.3	92.2	89.2	96.9	90.9	71.4	93	- -	- -
Ensembles on test (from leaderboard as of Sept. 16, 2019)										
ALICE	88.2	95.7	90.7	83.5	95.2	92.6	69.2	91.1	80.8	87
MT-DNN	87.9	96	89.9	86.3	96.5	92.7	68.4	91.1	89	87.6
XLNet	90.2	98.6	90.3	86.3	96.8	93	67.8	91.6	90.4	88.4
RoBERTa	90.8	98.9	90.2	88.2	96.7	92.3	67.8	92.2	89	88.5
Adv-RoBERTa	91.1	98.8	90.3	88.7	96.8	93.1	68	92.4	89	88.8
ALBERT	91.3	99.2	90.5	89.2	97.1	93.4	69.1	92.5	91.8	89.4

Table 2 shows the latest scores of ALBERT in the GLUE benchmark test. Both integration results and single model show that ALBERT has significantly improved in GLUE score, SQuAD 2.0 test F1 score and RACE test accuracy. Specifically, the accuracy of ALBERT in the RACE test was 89.4%, 17.4% higher than that of BERT, 7.6% higher than that of XLNet, 6.2% higher than that of RoBERTa, and 5.3% higher than that of DCMI+[3]. Moreover, the accuracy of a single model of ALBERT reached 86.5%, still 2.4% higher than the most advanced integrated model. These data show the excellent performance of ALBERT in several benchmark tests, proving its excellent performance in natural language processing tasks.

Table 3. The most advanced results of SQuAD and RACE [3]

Models	SQuAD1.1 dev	SQuAD2.0 dev	SQuAD2.0 test	RACE test (Middle/High)
Single model (from leaderboard as of Sept. 23, 2019)				
BERT-large	90.9/84.1	81.8/79.0	89.1/86.3	72.0 (76.6/70.1)
XLNet	94.5/89.0	88.8/86.1	89.1/86.3	81.8 (85.5/80.2)
RoBERTa	94.6/88.9	89.4/86.5	89.8/86.8	83.2 (86.5/81.3)
UPM	-	-	89.9/87.2	-
XLNet + SG-Net Verifier++	-	-	90.1/87.2	-
ALBERT (1M)	94.8/89.2	89.9/87.2	-	86.0 (88.2/85.1)
ALBERT (1.5M)	94.8/89.3	90.2/87.4	90.9/88.1	86.5 (89.0/85.5)

DOI: 10.54254/2755-2721/55/20241584

Table 3. (continued).

Ensembles (from leaderboard as of Sept. 23, 2019)				
BERT-large	92.2/86.2	0	-	-
XLNet + SG-Net Verifier	-	-	90.7/88.2	-
UPM	-	-	90.7/88.2	
XLNet + DAAF + Verifier	-	-	90.9/88.6	-
DCMN+	-	-	-	84.1 (88.5/82.3)
ALBERT	95.5/90.1	91.4/88.9	92.2/89.7	89.4 (91.2/88.6)

As shown in Table 3, although the parameters of the ALBERT xx large model are less than those of the BERT large model, it has achieved significantly better results. However, the computing cost is also higher due to its large structure. Therefore, the key to the next step is to accelerate ALBERT's training and reasoning speed through sparse attention and block attention to cope with the computational burden of its complex structure [3]. Besides, orthogonal research can provide additional representation capabilities, including more efficient language modeling training and complex example mining. Although there is credible evidence that sentence order prediction is a more valuable and consistent learning task that can produce better language representation, the current self-monitoring training loss may not fully capture more dimensions to create richer representations. These prospects provide essential guidance and direction for further improving the performance and efficiency of the ALBERT model.

In addition, in ELECTRA: Pre-training Text Encoders as Discriminators rather than Generators, a new strategy of alternative mask language modeling (MLM) pre-training method is proposed, called alternative token detection [7]. Unlike the traditional MLM method, this method destroys the input by replacing some tokens as reasonable substitutes sampled from the small generator network. The model no longer predicts the original identity of the damaged token but trains a discriminant model to predict whether the generator sample replaces each token in the input. The results show that, compared with MLM, this new pre-training task is practical on all input tags, not only on the masked subset. Therefore, even under the same model size, data, and computing conditions, the context representation learned by this method is significantly better than BERT. Especially for small models, the effect is particularly significant; With the same amount of training computation as GPT, the performance of this method in the GLUE natural language understanding benchmark is better than GPT. In addition, this method has also achieved good performance in terms of scale. Its performance is comparable to that of RoBERTA and XLNet. Still, the amount of computation used is less than a quarter of that of RoBERTA and XLNet, and it outperforms them under the same amount of computation. This new method provides a feasible solution for improving the efficiency and performance of the pre-training model.

Table 4. Some small models developed on GLUE [7]

Model Train / Infer FLOPs	Speedup	Params	Train Time + Hardware	GLUE
ELMo 3.3e18 / 2.6e10	19x / 1.2x	96M	14d on 3 GTX 1080 GPUs	71.2
GPT 4.0e19 / 3.0e10	1.6x / 0.97x	117M	25d on 8 P6000 GPUs	78.8
BERT-Small 1.4e18 / 3.7e9	45x / 8x	14M	4d on 1 V100 GPU	75.1
BERT-Base 6.4e19 / 2.9e10	1x / 1x	110M	4d on 16 TPUv3s	82.2
ELECTRA-Small 1.4e18 / 3.7e9	45x / 8x	14M	4d on 1 V100 GPU	79.9
50% trained 7.1e17 / 3.7e9	90x / 8x	14M	2d on 1 V100 GPU	79
25% trained 3.6e17 / 3.7e9	181x / 8x	14M	1d on 1 V100 GPU	77.7
12.5% trained 1.8e17 / 3.7e9	361x / 8x	14M	12h on 1 V100 GPU	76
6.25% trained 8.9e16 / 3.7e9	722x / 8x	14M	6h on 1 V100 GPU	74.1
ELECTRA-Base 6.4e19 / 2.9e10	1x / 1x	110M	4d on 16 TPUv3s	85.1

DOI: 10.54254/2755-2721/55/20241584

As shown in Table 4, the comparison results of trim models on the GLUE development set are shown in the table. When extrapolating FLOP and assuming a single input with a length of 128, ELECTRA Small/Base shows good performance, and its score is 5 GLUE points higher than that of the BERT Small model of the same kind, even more than that of the larger GPT model. Although the training time of ELECTRA Small is short (only 6 hours), it can still achieve good performance. Although small models extracted from larger pre-training transformers can also get perfect GLUE scores, these models need to spend a lot of computing resources to pre-train larger teacher models [7]. The results further prove the strong performance of ELECTRA on the medium scale; The primary size of the ELECTRA model in this study is better than BERT Base, and even better than BERT Large. This shows that ELECTRA can achieve substantial performance results with relatively few computing resources, which expands the possibility of applying pre-training models in natural language processing.

3.2. Emotional analysis of SST-5 fine-grained classification

This study primarily enhances the performance of the capsule network [8] by refining the dynamic routing mechanism mentioned earlier. This paper attempts to solve traditional capsule networks' computational complexity and training difficulties by introducing weighted kernel density estimation and fast dynamic routing methods. At the same time, this paper also studies credit allocation under different routing combinations to further optimize the capsule network's performance.

Table 5. Classification Accuracy [8]

Classification Benchmark	Accuracy (%)
Natural Language	
IMDB	96.2
SST-5*	59.8
SST-2	96.0
Vision	
ImageNet-1K @ 224×224 Top1	86.7
ImageNet-1K @ 224×224 Top5	98.1
CIFAR-100	93.8
CIFAR-10	99.2

This research has been tested on six classification benchmarks of natural language and has obtained the precision that competes with the most advanced level and is superior to the most advanced level in one case, as shown in Table 5. For natural language tasks, this study used RoBERTA large as a pre-trained transformer. Because RoBERTA large has some advantages over gpt-2:

First of all, RoBERTA large uses a more significant model size and more training data, which makes it perform better on multiple natural language processing tasks. Its training data includes many unlabeled Internet texts, which helps improve the model's language understanding ability.

Secondly, RoBERTA large model uses longer training time and more training steps. This empowers the model to more effectively grasp the semantic nuances of the context, thereby enhancing its proficiency in comprehending and generating context.

Moreover, RoBERTA large incorporates various training techniques, including dynamic masking, continuous text blocks, independent training, among others, to enhance the overall performance of the model. These strategies contribute to the model's efficacy in tasks such as semantic reasoning, emotion analysis, and named entity recognition.

In general, RoBERTA large performs better on multiple natural language processing tasks than gpt-2 model. Its advantages are mainly reflected in the larger model size, more training data, and longer training time. However, the specific performance differences need further experiments and evaluation to be determined.

DOI: 10.54254/2755-2721/55/20241584

For each benchmark in our study, a classification header is integrated into the pre-trained Transformer. This specialized header processes token embeddings from every layer of the Transformer, consolidating them into a unified sequence. Subsequently, three routes are sequentially applied, as illustrated in Table 6, with the input being the flattened, unified sequence of token embeddings computed from each layer of the Transformer.

Table 6. R1, R2 and R3 routes [8].

	$n^{(inp)}$	$n^{(out)}$	$d^{(inp)}$	$d^{(out)}$
R1	-	$n^{(hid)}$	$d^{(emb)}$	$d^{(hid)}$
R2	$n^{(hid)}$	$n^{(hid)}$	$d^{(hid)}$	$d^{(hid)}$
R3	$n^{(hid)}$	$n^{(cls)}$	$d^{(hid)}$	1

The number of R1 input vectors is not specified, because the length of the flat sequence is variable, nis (h id) is the number of hidden interpretation vectors selected in this study, dis (e m b) is the embedding size of the pre-trained Transformer, dis (h id) is the size of hidden interpretation vectors, and nis (c l s) is the number of specific classes for each task [8].

The paper presents LM-CPPF, a pioneering methodology named "Paraphrasing-Guided Data Augmentation for Contrastive Prompt-Based Few-Shot Fine-Tuning." It underscores the efficacy of Large Language Models (LLMs) in mastering tasks with limited datasets, and utilizing prompts for multitasking. LM-CPPF integrates insights from LLMs such as GPT-3 and OPT-175, employing retelling techniques to create diverse viewpoints. The models proficiently generate paraphrased sentences with diverse grammatical structures, surpassing mere lexical alterations.

Table 7. Performance of the benchmark model and LM-CPPF on six data sets [9].

Task	LM-BFF	LM-BFF+ SupConLoss	LM-BFF+ Multi-templates	LM-CPPF GPT-3	LM-CPPF OPT	LM-CPPF GTP-2	LM-CPPF FT GPT-2
SST-2	89.5	90.3	91.0	92.3	91.8	91.1	91.4
SST-5	48.5	49.6	50.3	52.8	52.2	51.4	51.6
MNLI	62.3	63.2	64.8	68.4	66.2	65.6	65.8
CoLA	6.9	9.6	11.6	14.1	13.3	10.7	11.8
QNLI	61.2	65.4	67.2	69.2	68.5	67.5	67.8
CR	89.7	89.9	90.2	91.4	91.1	90.2	90.7

Table 7 showcases the performance comparison between LM-CPPF and the baseline across six datasets in this study. Included in the evaluation are LM-BFF+multi template, a method proposed by Jian et al., and LMBFF+SupConLoss, which shares the same architecture but foregoes data enhancement, focusing solely on integrating monitoring comparison and the MLM loss function. The study also examines two GPT-2 scenarios: the training model and GPT-2 fine-tuning (FT) on the ParaNMT-50M dataset.

Table 8. Few shot Paraphrasing method and reverse translation (BT) and simple data enhancement (EDA) methods [9]

Task	Few-shot Paraphrasing	Back Translation					SR	RI	RS	RD	EDA
		AR	FR	DE	ZH	HI					
SST-2	91.8	90.8	90.6	90.4	90.7	90.3	90.5	89.5	90.8	91.3	90.4
SST-5	52.2	49.2	49.3	49.1	49.6	48.3	47.9	49.3	49.3	48.2	48.2
MNLI	66.2	64.3	63.1	63.8	65.4	62.2	62.9	63.2	61.7	60.2	60.3
CoLA	13.3	6.7	6.8	6.4	7.1	5.9	6.3	5.8	5.8	5.1	5.1

DOI: 10.54254/2755-2721/55/20241584

Table 8. (continued).

QNLI	68.5	66.5	66.2	65.8	66.6	64.3	66.1	65.9	66.3	65.6	63.3
CR	91.1	88.5	88.6	88.4	88.7	87.9	89.8	89.1	89.3	89.6	89.7

Table 8 presents an evaluation of the performance of Few-shot Paraphrasing, Back Translation, and Easy Data Augmentation, as detailed in this paper. The assessment encompasses five measures for Back Translation and four measures for Easy Data Augmentation [9].

The implementation of LM-CPPF consists of two steps. Initially, the labels are matched using the target sentence in the template and specific demonstration in the prompt. The matched phrases are then employed to calculate the masked language modeling loss. In the second step, a supervised contrastive loss is computed by comparing the target cue with another example sharing the same template but with a different presentation.

Experimental findings underscore the efficacy of Few-shot Paraphrasing as a data augmentation method for fine-tuning pretrained language models based on contrast cues. It outperforms other data augmentation methods, including Back Translation and Easy Data Augmentation, particularly in text classification tasks. In summary, LM-CPPF significantly boosts the performance of LM-BFF through contrastive learning on paraphrases generated by large language models.

3.3. IMDb emotion analysis

3.3.1. XLNet. In the paper titled "XLNet: Generalized Autoregressive Pretraining for Language Understanding," the authors propose XLNet, an innovative autoregressive pre-training method designed for language understanding [10]. By optimizing the expectation of all permutations, XLNet proficiently captures bidirectional contexts, mitigating the constraints observed in BERT. Experimental findings showcase the superiority of XLNet over BERT across a spectrum of 20 tasks, spanning question answering, natural language reasoning, emotion analysis, and document sorting.

Past solutions include autoregressive (AR) language modeling and self-encoder (AE) based pre-training, but they all have limitations. Therefore, this paper aims to propose a pre-training method that can combine both advantages and avoid their weaknesses. XLNet is realized by optimizing the expected logarithmic likelihood of sequences by considering all conceivable permutations of factorization order. This approach ensures that each position can leverage contextual information from both the left and right sides. It does not rely on data corruption and avoids pre-training fine-tuning differences. XLNet also integrates the architecture design of Transformer-XL to improve performance, especially for long text sequences. The Transformer (- XL) network is re-parameterized to eliminate ambiguity in permutation-based language modeling.

Through detailed experimental setup and results, the paper shows that the XLNet method is generally superior to BERT and RoBERTa in various tasks. The specific performance results are shown in the relevant experimental tables. In most cases, XLNet achieved higher scores on EM and F1 indicators than BERT and RoBERta.

3.3.2. BERT large. The paper titled "Unsupervised Data Augmentation for Consistency Training" delves into the critical challenges associated with consistency training in natural language processing (NLP) [11]. It introduces an innovative method that leverages unsupervised data enhancement to address these challenges, potentially reshaping the prevailing practices in training models for diverse NLP tasks. Consistency training plays a pivotal role in augmenting the robustness and generalization capability of NLP models, to maintain consistent predictions amid various disturbances. However, traditional training methods encounter obstacles such as limited labeled data and difficulties in capturing diverse data distributions. The proposed unsupervised data enhancement strategy enriches training data, enabling models to learn from more comprehensive datasets.

DOI: 10.54254/2755-2721/55/20241584

The research demonstrates the effectiveness of unsupervised data enhancement through extensive experiments across various NLP tasks. Results indicate significant improvements in model performance, particularly in accuracy, recall, and F1 score. Enhanced generalization capabilities are observed as models adapt better to different data distributions, contributing to an overall improvement in performance.

A major advantage of the proposed method lies in its capacity to reduce dependence on extensive labeled data, a common bottleneck in NLP training. Unsupervised data expansion allows learning from a broader, more diverse dataset, enhancing training efficiency and fostering the development of robust, adaptive NLP models. However, limitations exist, such as the reliance on the model's ability to generate meaningful data representations and challenges associated with the computational complexity when processing large datasets.

In conclusion, the study emphasizes the potential of unsupervised data enhancement for consistency training in NLP, offering a promising avenue to enhance model performance and robustness across various tasks. While the method has demonstrated substantial improvements, ongoing research and advancements in data enhancement technology are crucial to overcome existing limitations and maximize its effectiveness.

4. Prospects

Summarizing the above experimental methods and data, this paper has the following prospects for natural language processing applied to sentiment analysis in the future:

4.1. Context awareness and context understanding

In conclusion, the imperative for advancing sentiment analysis lies in addressing challenges associated with lengthy or intricately contextualized texts. The focal point of future developments should center on augmenting the model's contextual comprehension, achievable through strategic approaches:

1. Large-scale Pre-training Models: A continuous commitment to refining and advancing large-scale pre-training models is pivotal for effectively capturing contextual nuances in text, thereby elevating sentiment analysis performance. Particular emphasis should be given to minimizing the scale necessary for pre-training.

2. Multi-modal Information Fusion: Integration of diverse data modalities, encompassing text, image, audio, and video, facilitates the extraction of more nuanced emotional information. Recognizing the myriad ways emotions manifest on social media (text, emoticons, images, sounds), and leveraging multi-modal data promises a deeper understanding of sentiment.

3. Cross-sentence and Cross-document Understanding: The development of models adept at scrutinizing emotions across sentences and documents is imperative. This approach strives to comprehensively capture the overarching context and continuity of emotional expressions, thereby contributing to enhanced sentiment analysis in intricate textual landscapes.

4.2. Fine-grained sentiment analysis

Current emotional analysis usually divides emotions into positive, negative, and neutral. The future development direction will pay more attention to the fine-grained analysis of emotions, including emotional types and intensity. This will help to capture the diversity of emotional expression more accurately. The improvement directions include:

Emotional categories: develop models that can identify more emotional categories (such as joy, sadness, anger, surprise, etc.) to understand emotions in more detail.

Emotional intensity: develop a model that can identify the intensity and degree of emotion, to measure the intensity of emotional expression more accurately.

Emotional evolution: study how emotions evolve with time to better understand the dynamics of emotions.

Mixed emotions: The model can deal with the situation that there are multiple emotions in the text, such as emotions that contain both joy and worry.

DOI: 10.54254/2755-2721/55/20241584

4.3. Model interpretability

Deep learning models have made remarkable achievements in emotion analysis, but they are often regarded as "black boxes" and challenging to explain their decision-making process. The future development direction will focus on improving the interpretability of the emotion analysis model so that users and decision-makers can understand the working principle of the model. The improvement direction includes:

Visualization tools: develop visualization tools to show the model's decision-making process in emotional analysis, such as which words or sentences impact the results of emotional analysis.

Explanatory methods: explore explanatory methods, such as attention heat map and importance score, to explain the prediction results of the model.

5. Conclusion

Emotion analysis, a pivotal facet within the realm of emotion computing, is dedicated to the recognition and analysis of emotions in text. The comprehensive review covers a spectrum of affective analysis methods, spanning from traditional dictionary-based approaches to machine learning and deep learning methods. The dictionary-based approach relies on emotion dictionaries for classifying text emotions, but its efficacy is constrained by limitations in coverage and accuracy. Machine learning methods, leveraging classification algorithms and feature engineering, conduct emotion analysis but necessitate a substantial amount of tagged data and artificial feature extraction. On the other hand, deep learning methods, represented by RNNs, CNNs, and Transformers, excel in capturing intricate emotional features in text, albeit requiring significant data and computational resources. This paper meticulously illuminates the strengths and weaknesses inherent in each method, emphasizing the potential advantages of deep learning for advancing emotion analysis.

In this paper, a comprehensive literature review is carried out in the field of emotion analysis, and the development, application, and prospect of natural language processing technology in this field are discussed. Through the analysis of several typical affective analysis models based on deep learning, including BERT, ALBERT, XLNet, etc., this paper compares the efficiency differences of these models when dealing with corpora of different sizes, and discusses their advantages and limitations. Research shows that the model based on Transformer architecture can capture semantic information more accurately than traditional methods, and is more effective for sentence-level and document-level context modeling. Experimental results on binary classification, multi-category segmentation, and large-scale data sets show that these deep-learning models have achieved significant advantages. In conclusion, this paper proposes that deep learning technology has great potential in improving the performance of emotion analysis. Looking forward to the future, the development of natural language processing in the field of emotion analysis needs to focus on enhancing the context-understanding ability of models, conducting more fine-grained multi-category emotion analysis, and improving the interpretability of model results. The research of this paper hopes to provide a helpful overview and enlightenment for scholars and practitioners in this field.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł. Polosukhin I 2017 Attention is all you need *Neural Information Processing Systems; Curran Associates, Inc.*

[2] Devlin J, Chang M W, Lee K, Toutanova K 2018 Bert: Pre-training of deep bidirectional transformers for language understanding

[3] Zhenzhong L, Mingda C, Sebastian G, Kevin G, Piyush S, Radu S 2019 ALBERT: A lite BERT for self-supervised learning of language representations *arXiv,1909.11942*

DOI: 10.54254/2755-2721/55/20241584

[4] Colin R, Noam S, Adam R, Katherine L, Sharan N, Michael M, Yanqi Z, Wei L, Peter J L 2019 Exploring the limits of transfer learning with a unified text-to-text transformer *arXiv,1910.10683*
[5] Yinhan L, Myle O, Naman G, Jingfei D, Mandar J, Danqi C, Omer L, Mike L, Luke Z, Veselin S 2019 RoBERTa: a robustly optimized BERT Pretraining approach *arXiv,1907.11692*
[6] Armen A, Anchit G, Akshat S, Xilun C, Luke Z, Sonal G 2021 Muppet: massive multi-task representations with pre-finetuning *arXiv,2101.11038*
[7] Kevin C, Minh-Thang L, Quoc V, Christopher D M 2020 ELECTRA: Pre-training text encoders as discriminators rather than generators *arXiv,2003.10555*
[8] Franz A H 2022 An algorithm for routing vectors in sequences *arXiv,2211.11754*
[9] Amirhossein A, Sascha R, Yadollah Y 2023 LM-CPPF: paraphrasing-guided data augmentation for contrastive prompt-based few-shot fine-tuning *arXiv,2305.18169*
[10] Zhilin Y, Zihang D, Yiming Y, Jaime C, Ruslan S, Quoc V L 2019 XLNet: generalized autoregressive pretraining for language understanding *arXiv,1906.08237*
[11] Qizhe X, Zihang D, Eduard H, Minh-Thang L, Quoc V L 2019 Unsupervised data augmentation for consistency training *arXiv,1904.12848*