Task-Aware BERT-based Sentiment Analysis from Multiple Essences of the Text

Jia-Hao Hsu[1], Chung-Hsien Wu[1] and Tsung-Hsien Yang[2]
[1]Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, TAIWAN
[2]Telecommunication Laboratories Chunghwa Telecom Co., Ltd., Taoyuan, Taiwan
E-mail: jiahaoxuu@gmail.com, chunghsienwu@gmail.com and yasamyang@cht.com.tw

Abstract— **Text sentiment analysis has always been an important topic in the research of human-computer interactions and is generally applied to help businesses monitor product satisfaction and understand customer needs. The research in this study intends to consider the sentiment of the text with a focus on capturing multiple essences of the text, such as words, events and sentence, in a specific task. First, an approach to automatically extract the task-specific emotional key terms in the corpus of a specific task is proposed. Task-specific key events are manually/automatically designed for the specific application task. The BERT-based model is employed to integrate the outputs from sentence, key terms and events of the input text for task-aware sentiment analysis. The pre-trained sentence-based BERT model is fine-tuned using the Ren-CECps, a large-size Chinese weblog emotion corpus. Then we transfer the encoder weights to a new model and initialize a new linear layer, and finally fine-tune this model to fit the specific task. For evaluation, the Telecom Domain Customer Service Corpus (TD-CSC), a telecommunications service dataset, was used. The experimental results show that the proposed BERT-based model using multiple essences improved the correct rate by 10% compared to that without using the multiple essence features.**

I. INTRODUCTION

In recent years, advanced technology has been continuously applied in all walks of life. Companies are committed to achieving a balance between product quality and use quality, to obtain higher customer satisfaction and increase revenue. In the past, the methods of obtaining customer satisfaction were market satisfaction surveys. But there were many interference factors in the process of such methods, and the authenticity of the customer's answers could not be accurately known, and it usually took a lot of time to obtain information. Therefore, this study proposes an approach to real-time monitoring of customer status such as emotional response in social media and customer conversations, to obtain customer emotional information to facilitate the smoothness and completeness of subsequent services.

Sentiment analysis [1] and emotion recognition [2], which can automatically mine unstructured data (social media, emails, customer service tickets, and more) for opinion and emotion, have been playing a crucial role in both commercial and research applications [3]. Recently, machine learning and deep learning algorithms have achieved a great success in sentiment analysis. As the research on deep learning becomes more and more mature, considerable progress has been made in the extraction of text features, such as word embedding, which can get quite good results with the use of deep learning methods. Common word embedding can be roughly divided into two categories: fixed-characterized word vectors and dynamically-characterized word vectors. The difference lies in whether the vectors of the same word are the same. For fixed representation of word vector, no matter which sentence the word appears, it will be encoded into the same word vector, such as Word2Vec [4], FastText [5] and GloVe [6]. Given the training text, the deep neural network can learn the relationship between words, and output a fixed vector for each word. In contrast, for dynamically represented word vectors, different word vector encodings of the same word can be obtained owing to the difference between the contexts. This type of methods can solve the problem of polysemous words, such as ELMo (Embeddings from Language Models) [7] and BERT (Bidirectional encoder representation transformer) [8]; the output of all word vectors will be affected by the current context in the network, and the degree of influence of distance will be considered. However, all the above-mentioned word vector encoding methods target the distributional semantics; that is, they only consider the frequency of the word in the text and the context of the word as a basis, and do not reflect sentiment characteristics. Gradually, some researchers have proposed approaches to the encoding of emotional word vectors (Emotional Word Embedding) [9, 10]. If the target can take emotion into consideration when encoding the word vector, the influence of the emotional word in the sentence can be improved. The sentiment amount of the sentence is used as the model encoding target, and the hidden layer vector in the model encoding process is taken as the emotional word vector of the text, which is used as the recognition input.

In the research of sentiment analysis, we often face the problem that the same word has different sentiments in different tasks/domains. For example: "There is only one frame for the signal". In the domain of telecommunications, the "one frame" in the sentence represents poor signal and contains negative sentiment. But in other domains, the word may not have positive sentiment. Therefore, application corpora in different domains cannot be compatible with each other, or the use of multi-domain corpus to build a sentiment analysis system in a specific domain can bring limited improvement.

978-988-14768-9-0/21/$31.00 ©2021 APSIPA

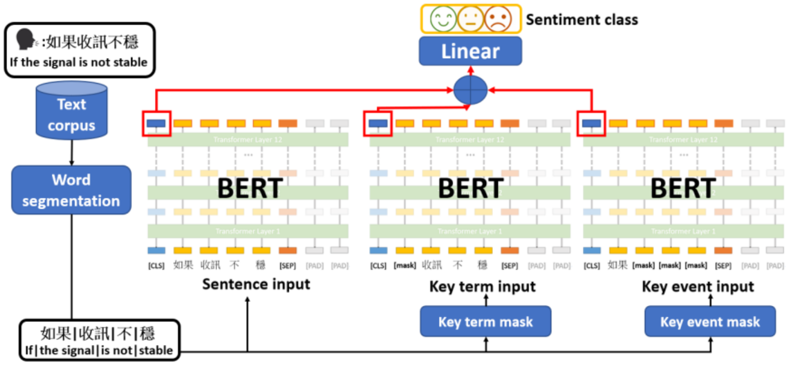


Fig. 1. Proposed system framework

However, the current study to solve the problem of word sentiment difference mostly focuses on detecting domain-invariant features from the input, or searching for keywords that can represent the same sentiment in different domains [11, 12]. Through adversarial training of the language model, the text input is characterized by some emotional words that can be found across domains. The disadvantage is that the amount of these representative words that can represent common emotions in different domains will be reduced. Therefore, the method proposed in this study focuses on integrating emotional key terms and key events based on the corpus of a specific task. The main contribution of this study is two-fold: First, combining multiple essences of the text, such as words, events and sentence to improve the performance of sentiment analysis in a specific task. The second is that the key terms and events BERT-based models with transfer learning can pay more attention to the important and task-related essences of the text.

II. Methods

The system architecture proposed in this study is shown in Fig. 1. We first perform word segmentation on the input sentence. In order to cover multiple essences of the input text for task-aware sentiment detection, the embeddings of the input sentence, emotional key terms, and key events are extracted, respectively, using the BERT-based models followed by transfer learning. First, the text input is fed to a BERT-based model for sentence embedding. The sentence-based BERT model is pre-trained using the Ren-CECps [13, 14], a large-size Chinese weblog emotion corpus. Then we transfer the encoder weights to a new model and initialize a new linear layer, and fine-tune this model to fit the target task. Key terms in the input sentence are extracted using an algorithm described in the following section for the target task. For key term embedding, the extracted key terms along with the other context of the sentence which are masked are fed to the BERT-based model to obtain key term embedding. The events composed of word patterns along with the masked content are fed to another BERT-based model for event embedding extraction. Finally, a fully connected layer is employed as the classifier by integrating three embeddings from three BERT models for sentiment classification. The method of masking the position of the non-critical content can retain the position information of the key content in the sentence, and provide more attention to these key contents.

A. Key Term and Key Event Extraction

The method for key term extraction is based on term frequency. First, we use the Chinese word segmentation system, Jieba, to segment the sentence, and use *tf-idf* to calculate the score of each *n*-gram term in the positive and negative corpus. The dataset is divided into positive and negative documents based on the sentiment labels. In equation 1, *s* means the document type and *t* means the term. It is calculated to identify the importance of the term *t* presenting the *s-th* sentiment [15, 16]. When the terms with length *n* are obtained, we calculate their spam-score, and finally select the top *N* terms into the key term set. The parameters *n* and *N* will be verified by subsequent experiments. The terms obtained, such as "want to cancel the contract," "network instability," "not solving my problem," etc., are the key terms which include multiple words.

$$spam\text{-}score_t^s = tf_{t,s} * idf_t - tf_{t,\neg s} * idf_t \,, \quad s \in \{positive \,, negative\} \tag{1}$$

On the other hand, the key events, which is composed of connected or disconnected phrase patterns are extracted automatically like key term extraction followed by manual selection. Manual selection allows the model to learn unseen, important information that may not have appeared in the corpus. The manually defined key event set is shown in Tab. 1.

Tab. 1 Key event set

Sentiment	**Context in the event set**
Positive	|be interested in / be satisfied with| + |some service|
Neutral	|if / assume|
Negative	|cancel / complain| + |some service|; Change to other phone carriers

B. Key Content Embedding

We use the positive and negative key terms and key events as the input, and replace each remaining content with a mask token **[MASK]** which is defined in the vocabulary set of BERT, retaining the position information. This method is like giving more attention to the key content, and it can retain the position information and sentence sequence information of the key content in the sentence.

C. *Fine-tuning of Classification Models*

In this study, we use the BERT model [17] as the sentence embedding model. The BERT model is pre-trained with a large amount of text data. In this study, the Chinese Wikipedia corpus is used as the training text for unsupervised learning. As the BERT model is a general model, it is not possible to fully present the word relationships for a specific task, such as emotional relationships, logical relationships, and telecommunications service relationship required by this study. Therefore, we will use such a pretrained BERT model, and perform fine-tuning to adapt the model to fit a specific task. Like most downstream tasks that use pre-trained BERT, we use BERT model for sentence embedding. A classification layer is employed to classify the embedding vector, trained based on the weighted cross entropy as the loss function [18, 19], shown in equation 2, for model adjustments. Weighted adjustments w_{class} according to the number of categories in the classification is used to solve the problem of data imbalance. The classification target can be freely defined. Here we use the positive and negative sentiments in the telecommunications service as the classification target.

$$\mathrm{loss}(x, class) = \frac{1}{w_{class}}\left(-x[class] + \log\left(\sum \exp(x[j])\right)\right) \qquad (2)$$

The three BERT models in Fig. 1 are independent models. We fine-tune the sentence BERT model using an emotion-domain corpus, Ren-CECps, followed by the transfer learning to the telecommunications task. The key terms BERT model is fine-tuned based on the corpus of the telecommunications domain and only key terms are used as the input, while the other terms are masked. Besides, we use the key events in telecommunications corpus for model fine-tuning the key events BERT model. Finally, the three BERT-based models are combined by integrating sentence, key terms and key events BERT models for sentiment classification.

III. Experiments

For evaluation, as the corpus used was the Telecom Domain Customer Service Corpus (TD-CSC), most of the words used were related to telecommunications content, such as “signal,” “network,” and even service plan names. The objective of the sentiment analysis task is to identify the user’s statement on the satisfaction valence of Telecom Customer Service System. If the service was satisfied, the text was labeled as positive, dissatisfied text was labeled as negative, and for no significant evaluation, a neutral label was given. The input was the customer feedback questions collected from the Chinese customer service dialogue system. The output was the user's sentiment with Telecom Customer Service, and it was divided into positive, neutral and negative sentiments.

A. *Corpus*

The corpus used in this study is shown in Tab. 2. The table contains the Telecom Domain Customer Service Corpus (TD-CSC) and the TD-augmented Corpus. TD-CSC was the original corpus collected from the customer service system in Telecom Domain, and was annotated as the text corpus with sentiment labels. In view of the privacy issues related to customer personal information, TD-augmented corpus was included for this study. TD-augmented corpus was an augmented corpus of TD-CSC, which was augmented by discarding the user information and simulating the data in TD-CSC. It contains the characteristics of the original corpus, and excludes personal information from the original corpus. Compared with the original corpus, the distribution of each category of the augmented corpus was more balanced and more suitable for model training.

Tab. 2 The distribution of each category of the corpus.

Corpus	Positive	Neutral	Negative
TD-CSC	36	14268	1447
TD-augmented	343	3309	1248

B. *Experimental Results*

To evaluate the reliability of the method proposed in this research, we evaluated the original BERT-based model and the proposed systems with key terms and key events. Five-fold cross-validation method was adopted for evaluation.

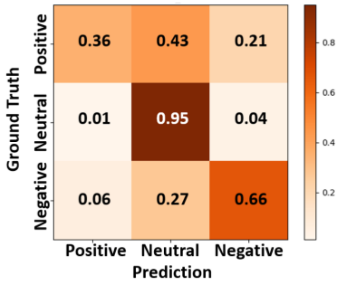


Fig. 2 Confusion matrix of the original BERT model

The confusion matrix obtained in the original BERT-based model is shown in Fig. 2. The test data, TD-augmented, was the input of the BERT model fine-tuned from the model trained on the emotion corpus, Ren-CECps, to evaluate the accuracy of the BERT model in the telecommunication domain. It can be seen from the figure that the neutral data performed well. But we can also find that the model predicted most of the data as neutral. The reason for this problem may be that the model has not grasped the key information to predict the positive and negative sentiments in the telecommunication domain.

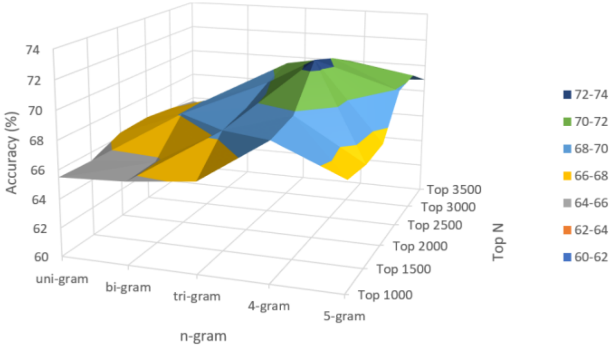
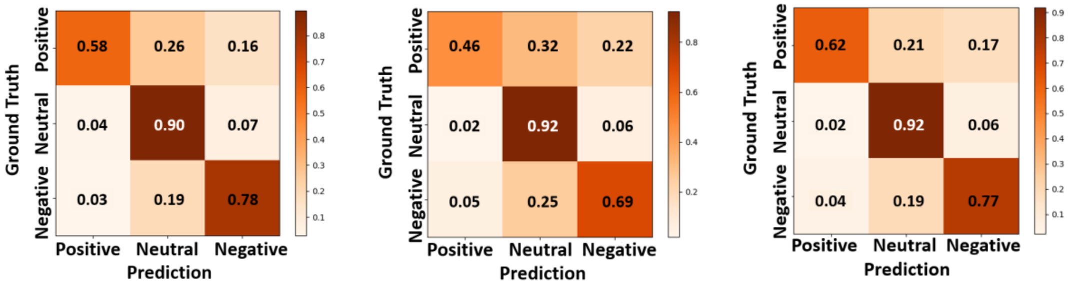


Fig. 3 Accuracy of different parameters of key term extraction

First, we evaluated the performance of key term extraction. The method proposed in [20] was used to calculate the total



(a) (b) (c)

Fig. 4 Confusion matrices of the original BERT-based model combined with (a) the key terms BERT-based model, (b) the key events BERT-based model, and (c) the key terms and key events BERT-based model.

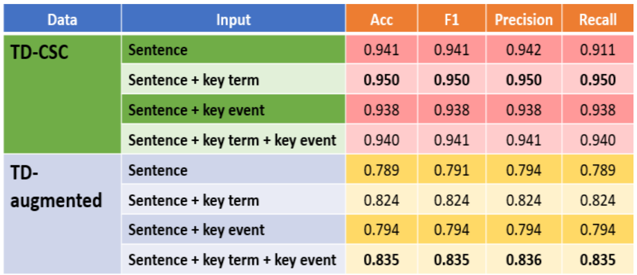
score of the sentence. The correct rate was used to find the best parameters for key term extraction, as shown in Fig. 3. The adjustable parameters are the number of *n*-grams and the number of top *N* terms selected into the key term set. The final key term set contained term lengths from uni-gram to 4-gram, and each length considered the top 2000 high-score terms.

Fig. 4(a) and Fig. 4(b) are the results of the sentence BERT model combining key terms and key events, respectively. Fig. 4(c) is the result of the sentence BERT model combining both key terms and key events. The system can effectively identify the correct results for the input of key terms and key events. The results show that the performance for classifying both positive and negative sentiments could be improved. It is worth mentioning that because there are more negative data than the positive data in the corpus, the improvement of the negative data is higher than that of the positive data.

Tab. 3 The results of data with and without key terms/events

	Data with key terms (75.7%)	**Without key term (24.3%)**	**Whole data**
Baseline (BERT)	79.51%	77.11%	78.89%
Adding key terms	84.63%	76.27%	82.44%
	Data with key events (5.6%)	**Without key event (94.4%)**	**Whole data**
Baseline (BERT)	80.55%	78.79%	78.89%
Adding key events	92.16%	78.68%	79.44%

We also evaluated the performance of the two systems with and without key terms and key events, and the results are shown in Tab. 3. The key terms appeared in about 75.7% of the data in 5 folds of the test data, while the key events appeared in about 5.6% of the data. Compared with key terms, key events occurred less frequently, and the overall performance that can be improved by key events was very limited. However, if key events were included in the data, the sentiment could be mostly correctly identified, and the accuracy was as high as 92.16%.



Data	Input	Acc	F1	Precision	Recall
TD-CSC	Sentence	0.941	0.941	0.942	0.911
	Sentence + key term	**0.950**	**0.950**	**0.950**	**0.950**
	Sentence + key event	0.938	0.938	0.938	0.938
	Sentence + key term + key event	0.940	0.941	0.941	0.940
TD-augmented	Sentence	0.789	0.791	0.794	0.789
	Sentence + key term	0.824	0.824	0.824	0.824
	Sentence + key event	0.794	0.794	0.794	0.794
	Sentence + key term + key event	**0.835**	**0.835**	**0.836**	**0.835**

Fig. 5 Total results of all methods and two corpora

Fig. 5 shows the complete experimental results, including the results of TD-CSC and TD-augmented corpora. Combining key terms can improve the accuracy of TD-augmented corpus. However, for TD-CSC corpus, because there were too many neutral data, not many positive and negative key terms and events can be extracted. The improvement of combining two features was not significant.

IV. Conclusions

This study aims to establish a sentiment analysis system, trying to capture and react sentiment information in user interaction and dialogue customer service system in the task of telecommunications. The research in this study intends to consider the sentiment of the text with a focus on capturing multiple essences of the text, such as words, events and sentence, in a specific task. The three BERT-based models are combined by integrating sentence, key terms and key events BERT models for final sentiment classification.

In the experimental results, the proposed method can improve the performance for positive and negative sentiment analysis. Compared with the baseline, the accuracy for the TD-augmented corpus was improved by 4.7%. Because the key terms appear more frequently, combining key terms for model construction can get a greater efficiency improvement. Although the number of the data with key events is few for the overall evaluation, it can be perfectly identified if there is a key event in the data. In practical applications, a good definition of key events provides a significant effect for sentiment analysis.

In sentiment analysis, it is often necessary to deal with the corpus of a specific task, which has the problem of insufficient amount of data. How to extract the key identification basis from the limited or unbalanced corpus is a problem. In practical applications, the corpus of customer service may not contain text only. Incorporating other information, such as speech, emoji, symbol, etc., for sentiment analysis is also the future direction of this study.

Acknowledgment

The authors would like to thank the Ministry of Science and Technology, Taiwan for partially supporting this research under contract no. 108-2221-E-006 -108 -MY3.

REFERENCES

[1] Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowledge-Based Systems,* p. 107134, 2021.

[2] Jia-Hao Hsu, Ming-Hsiang Su, Chung-Hsien Wu, and Yi-Hsuan Chen, "Speech Emotion Recognition Considering Nonverbal Vocalization in Affective Conversations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 29, pp. 1675-1686, 2021.

[3] Kun-Yi Huang, Chung-Hsien Wu, Ming-Hsiang Su, and Yu-Ting Kuo, "Detecting unipolar and bipolar depressive disorders from elicited speech responses using latent affective structure model," *IEEE Transactions on Affective Computing,* vol. 11, no. 3, pp. 393-404, 2018.

[4] Yoav Goldberg and Omer Levy, "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method," *arXiv preprint arXiv:1402.3722,* 2014.

[5] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759,* 2016.

[6] Jeffrey Pennington, Richard Socher, and Christopher D Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532-1543.

[7] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365,* 2018.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805,* 2018.

[9] Ameeta Agrawal, Aijun An, and Manos Papagelis, "Learning emotion-enriched word representations," in *Proceedings of the 27th international conference on computational linguistics*, 2018, pp. 950-961.

[10] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann, "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm," *arXiv preprint arXiv:1708.00524,* 2017.

[11] Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao, "Adversarial and domain-aware bert for cross-domain sentiment analysis," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4019-4028.

[12] Chung-Hsien Wu, Ze-Jing Chuang, and Yu-Chung Lin, "Emotion recognition from text using semantic labels and separable mixture models," *ACM transactions on Asian language information processing (TALIP),* vol. 5, no. 2, pp. 165-183, 2006.

[13] Ji Li and Fuji Ren, "Creating a Chinese emotion lexicon based on corpus Ren-CECps," in *2011 IEEE International Conference on Cloud Computing and Intelligence Systems*, 2011: IEEE, pp. 80-84.

[14] Ning Liu, Fuji Ren, Xiao Sun, and Changqin Quan, "Microblogging hot events emotion analysis based on Ren-CECps," in *Proceedings of the 2013 IEEE/SICE International Symposium on System Integration*, 2013: IEEE, pp. 233-238.

[15] Francisco Jánez-Martino, Eduardo Fidalgo, Santiago González-Martínez, and Javier Velasco-Mata, "Classification of spam emails through hierarchical clustering and supervised learning," *arXiv preprint arXiv:2005.08773,* 2020.

[16] Aditya Gaydhani, Vikrant Doma, Shrikant Kendre, and Laxmi Bhagwat, "Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach," *arXiv preprint arXiv:1809.08651,* 2018.

[17] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu, "Pre-training with whole word masking for chinese bert," *arXiv preprint arXiv:1906.08101,* 2019.

[18] Yafen Dong, Xiaohong Shen, Zhe Jiang, and Haiyan Wang, "Recognition of imbalanced underwater acoustic datasets with exponentially weighted cross-entropy loss," *Applied Acoustics,* vol. 174, p. 107740, 2021.

[19] Zhilu Zhang and Mert R Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *arXiv preprint arXiv:1805.07836,* 2018.

[20] Dibyendu Seal, Uttam K Roy, and Rohini Basak, "Sentence-level emotion detection from text based on semantic rules," in *Information and Communication Technology for Sustainable Development*: Springer, 2020, pp. 423-430.