

Sentiment analysis of Chinese game reviews based on SKEP

Haishuo Wang

Mathematics School, Shandong University, Jinan, China

Corresponding e-mail: 202000091094@mail.sdu.edu.cn

Abstract. As China's game industry enters the era of the stock market, improving the quality of game products in all aspects is gradually becoming an important part of the development of China's game industry. There exists a large amount of game review data on the Internet, and sentiment analysis of them can provide important help for game makers to make decisions. However, the huge scale of game reviews and the extensive use of metaphors and sarcasm make it difficult to conduct sentiment analysis. In this paper, we experiment with a publicly available Chinese game review dataset using the ERNIE pre-training model equipped with the SKEP method, which improves the model's masking strategies and pre-training objectives to specialize the model's sentiment analysis ability. In addition, we compared the performance of ERNIE1.0, RoBERTa, MacBERT, BERT-wwm, and Nezha. The precision, recall, and F1 score of SKEP are higher than those of the models for comparison studies, which confirms the effectiveness of SKEP in this area.

Keywords: SKEP, games, sentiment analysis.

1. Introduction

China's game industry has developed rapidly in the last decade or so, and the number of game players has continued to increase. However, according to the report of Game Publishing Committee of CADPA (GPC), the Chinese game industry shrank for the first time in 2022 under the influence of various factors such as the epidemic, anti-addiction system, etc., and officially entered the era of stock market [1]. In such an environment, grasping players' emotional feedback on games through reviews to make real-time updates and adjustments to products is gradually becoming an important way for game makers to improve the retention rate and reputation of their products. Conducting large-scale manual analysis is undoubtedly inefficient, due to the massive and scattered features of game review data, so how to efficiently process and analyze game reviews is an urgent problem in the Chinese game industry at present. Driven by the rapid development of natural language processing (NLP), sentiment analysis has also gained tremendous momentum. The application of sentiment analysis techniques in the game reviews has great potential to reduce the cost and increase the efficiency of review data analysis.

Sentiment analysis is a cutting-edge research discipline at the intersection of multiple fields, which includes social sciences, psychological sciences, deep learning, and data mining, and is an important component of natural language processing [2]. In recent years, researchers have proposed a large number of methods and models for different tasks such as sentence-level sentiment classification, aspect-level sentiment classification, opinion target extraction, and multimodal sentiment analysis. In 2018, Google released BERT [3], which innovatively proposed masked language model (MLM) and next sentence prediction (NSP), and used it as the main tasks for model training. BERT has achieved excellent results on many tasks, making pre-trained models gradually become a hot topic in natural language processing, and more widely used in business, entertainment, information prediction, and personal emotion detection.

In this paper, we apply SKEP to sentiment analysis of game reviews. In the field of sentiment analysis, a pre-trained model is often trained based on a specific task. To address this situation, the Baidu research team and CAS proposed the SKEP method [4]. SKEP performs the fusion of different types of textual sentiment knowledge and can conduct many different sentiment analysis tasks through a single model. The SKEP method first performs unsupervised sentiment knowledge mining and divides the sentiment words in the text into two parts for masking, one is sentiment words and the other is aspect-sentiment pairs. The pre-training objectives of SKEP have three parts, Sentiment

Word prediction, Word Polarity prediction, and Aspect-sentiment Pair prediction. SKEP outperforms its baseline RoBERTa [5] and shows excellent results on three basic sentiment analysis tasks: sentence-level sentiment classification, aspect-level sentiment classification, and opinion target extraction on multiple datasets. We use a dataset based on reviews of about 300 games from the Chinese mobile game forum TapTap. SKEP obtains an F1 score of 0.86, which on average exceeds other models used for comparison by about 0.05, demonstrating that SKEP has superior performance on this problem and a strong potential for sentiment analysis for game reviews. Sentiment analysis of game reviews has a promising future, yet relatively little research has been conducted on this niche in the field of Chinese natural language processing using pre-trained models. In this paper, we confirm SKEP's excellent performance on this task through comparison studies to help the development of the game industry.

The paper is about the sentiment analysis of Chinese game reviews based on SKEP. In section 2, we review related works, including game review analysis, Chinese sentiment analysis, and an introduction to the model used for the comparison studies. In section 3, we introduce the pre-trained model ERNIE [6] and SKEP method, which are two main components of the model used in this paper. In section 4, we introduce the dataset, the experiment setting, and the results. Finally, the experiments are summarized and some aspects for improvement are proposed.

2. Related Works

2.1 Sentiment analysis of game reviews

The study of game reviews has a long history. As early as 2009, Zagal et al. [7] revealed that game reviews contain a large amount of valuable content, such as game design suggestions, speculations about the intentions and goals of game creators, and advice to players on how to approach and best enjoy a particular game. Rajapakshe [8] suggested that proper analysis of the feedback collected from players would enable developers to identify key game features requested by players as well as bugs and imbalances in the game. Lin et al. [9] conducted an extensive analysis of game reviews on the Steam platform and compared the difference between positive and negative reviews, making a pioneering study for the application of binary sentiment classification in game review analysis.

Driven by the development of NLP, many sophisticated algorithms and models have been applied to sentiment analysis of game reviews to make it more intelligent and automated. Ruseti et al. [10] used traditional support vector machine, multinomial Naive-Bayes, and deep neural networks (DNN) for triple classification of pre-processed game review texts. Secui et al. [11] analyzed 9750 reviews using their own constructed model. Eight affective components were identified using a Principal Component Analysis (PCA) and a Discriminant Function Analysis based on the emerging components classified game reviews into three categories with a 55% accuracy. Vieira and Brandão [12] used a convolutional neural network (CNN) to analyze game reviews. ARIK [13] used Valence Aware Dictionary for sEntiment Reasoning (VADER), a dictionary and rule-based sentiment analysis tool specifically designed for predicting sentiments in social media texts to analyze reviews in the MMORPG sub-section on Reddit.

2.2 Chinese sentiment analysis

Driven by pre-training models, Chinese natural language processing has developed rapidly, and the field of sentiment analysis is no exception. Many pre-trained models have been applied to Chinese sentiment analysis datasets with excellent results. BERT is widely used in Chinese sentiment analysis. Xiang et al. [14] used a combination of BERT model and Hawkes process function to conduct sentiment analysis on dataset NLPCC2014. Li et al. [15] released a model based on BERT to analyze Chinese stock reviews. Other models have good performance on specific domains, for example, Luo et al. [16] applied the fine-tuned BiLSTM [17] model on a binary classified Chinese economics hotel review dataset. There are also some specific applications of SKEP used in this paper. Gai et al. [18] used the SKEP pre-trained model provided by Baidu intelligent cloud sentiment analysis platform to

identify 460,000 Weibo tweets from Xiamen, China in 2020 for studying inequalities between urban residents' sentiments and land values.

2.3 Comparison models

The model used in this paper is constructed by applying the SKEP method to the ERNIE model. ERNIE model was published by Zhang et al. [7] in 2019, which mainly improves the masking strategy of BERT. In addition to SKEP, the following models are used in this paper for comparison experiments. RoBERTa, an improved model of BERT proposed by Liu et al. [6] in 2019, BERT-wwm model, proposed by Cui et al. [18] by adding whole masking (wwm) to BERT, MacBERT proposed by Cui et al. [19] through improving the MLM task of BERT, Huawei's Nezha model [20], and ERNIE1.0 without SKEP. SKEP used in this paper specializes the ability of sentiment analysis in the design idea and training process, revealing its strong advantages in this domain compared to other pre-trained models mentioned above.

3. Methods

In this part, we illustrate what role ERNIE and SKEP play in the model of this paper, and give a brief description of them. This paper uses the SKEP pre-training model from PaddleNLP.

The model in this paper uses the transformer encoder structure of ERNIE and improves ERNIE according to the SKEP method. Specifically, the masking strategies and the pre-training objectives are changed. The model used in this paper first masks the tokens according to the masking strategies of SKEP. Then word embedding is performed and the word vectors are fed into ERNIE's network, i.e., multi-layer T-Encoder and K-Encoder structures. Then the parameters are tuned according to the pre-training objectives of SKEP. In conclusion, ERNIE forms the basic structure of the model in this paper, but its sentiment analysis ability mainly comes from SKEP.

3.1 ERNIE

ERNIE is a pre-trained language model proposed by Baidu. Considering that the random mask strategy of BERT may cause the model not to take full advantage of the lexical and grammatical structure in the training data, ERNIE has improved the masking strategy. The model used in this paper is built based on ERNIE. ERNIE uses two different masking strategies, a phrase-based masking strategy and an entity-based masking strategy, where phrases or entities consisting of multiple words will be masked together, which can take full advantage of lexical, syntactic, and knowledge information. In addition, unlike the transformer model used by BERT, the network structure of ERNIE consists of two main encoders, one is the underlying textual encoder (T-Encoder) responsible to capture basic lexical and syntactic information from the input tokens, and the other is the upper knowledgeable encoder (K-Encoder) responsible to integrate extra token-oriented knowledge information into textual information from the underlying layer. The number of T-Encoders is N and the number of K-Encoders is M . In the experiment of the original paper, both N and M are set to 6. Similar to BERT, ERNIE also uses MLM and NSP as pre-training tasks. Moreover, ERNIE proposes a new pre-training task called denoising entity auto-encoder (dEA). This paper uses the 1.0 version of ERNIE.

3.2 SKEP

SKEP first performs unsupervised sentiment knowledge mining using Pointwise Mutual Information (PMI) [22]. PMI relies on a small number of pre-given sentiment words and their polarities, and then more sentiment words and their polarities are mined using these sentiment words. For aspect-sentiment pairs, the sentiment word and its nearest noun will be considered as an aspect-sentiment pair. The maximum distance between the sentiment word and noun will not be more than 3 tokens.

SKEP applies the following three masking strategies, Aspect-sentiment Pair Masking, Sentiment Word Masking, and Common Token Masking, which mainly rely on the sentiment information. Firstly, at most 2 aspect-sentiment pairs are randomly selected to mask. All tokens of a pair are replaced by [MASK]. This step is called Aspect-sentiment Pair Masking. Then, the unmasked sentiment words are randomly selected to mask and replaced by [MASK]. The numbers of masked tokens in this step are limited to less than 10%. This step is called Sentiment Word Masking. Finally, if the numbers of masked tokens are less than 10%, Common Token Masking would be conducted.

For these masking strategies, SKEP uses three training objectives to tell the transformer encoder to recover the replaced sentiment information. The three objectives, Sentiment Word (SW) prediction L_{sw} , Word Polarity (WP) prediction L_{wp} and Aspect-sentiment Pair (AP) prediction L_{ap} are jointly optimized.

The SW prediction objective L_{sw} is as follows:

$$\hat{y}_i = \text{softmax}(\tilde{x}_i \mathbf{W} + \mathbf{b}) \quad (1)$$

$$L_{sw} = - \sum_{i=1}^n m_i y_i \log \hat{y}_i \quad (2)$$

Here, $\tilde{x}_i$ is the output from the encoder. y_i is the one-hot representation of original token x_i . $\mathbf{W}$ and $\mathbf{b}$ are parameters from the output layer. $m_i = 1$ if the i -th position of the sequence is a masked sentiment word, otherwise $m_i = 0$.

L_{wp} is similar to L_{sw} . The difference is that L_{wp} calculated the loss of the polarities instead of tokens, and polarities only have two kinds, positive and negative.

The AP prediction objective L_{ap} is as follows:

$$\hat{y}_a = \text{sigmoid}(\tilde{x}_1 \mathbf{W}_{ap} + \mathbf{b}_{ap}) \quad (3)$$

$$L_{ap} = - \sum_{a=1}^A y_a \log \hat{y}_a \quad (4)$$

Here, $\tilde{x}_1$ is the output vector in the position of [CLS]. A is the number of masked aspect-sentiment pairs in a corrupted sequence. y_a is the sparse representation of a target aspect-sentiment pair, and $\hat{y}_a$ is the evaluation of y_a .

The overall pre-training objective is:

$$L = L_{sw} + L_{wp} + L_{ap} \quad (5)$$

The operation process of SKEP is shown in Figure 1.

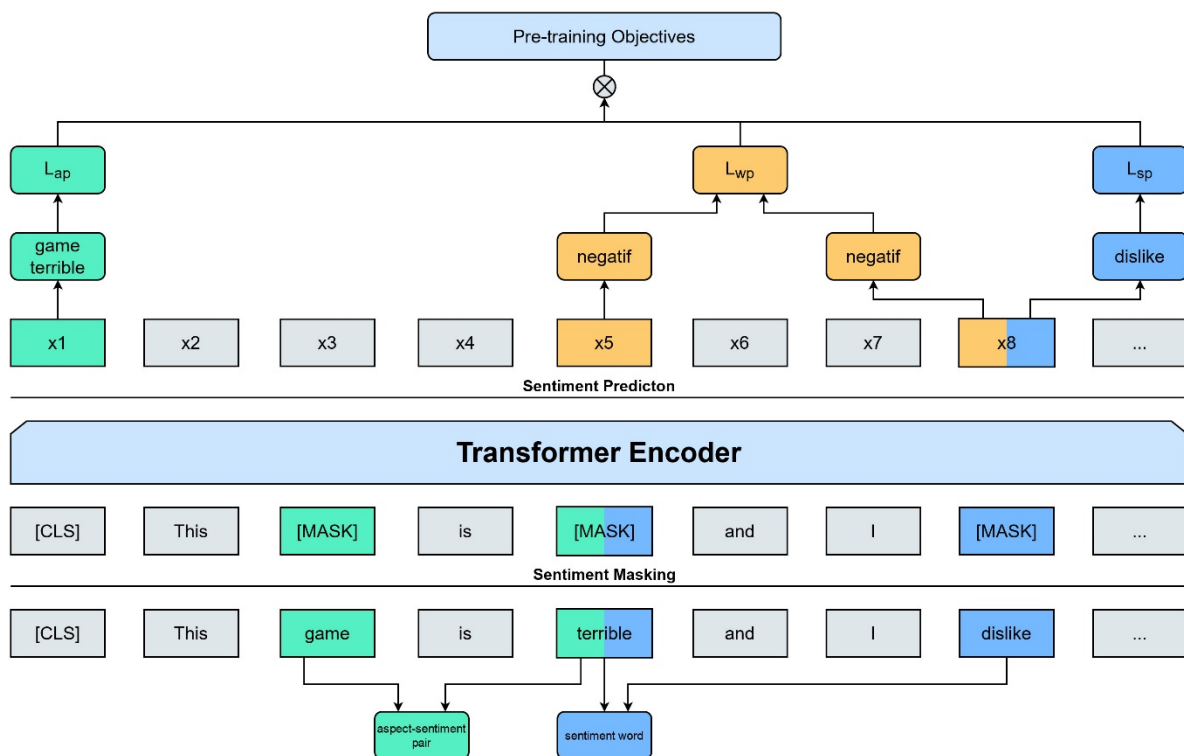


Figure 1. First, the sequence is masked according to the sentiment knowledge, and then word embedding is performed. After that, the word vector will be fed into the transformer encoder. The final output result is formed into probability distribution by the softmax layer and used to calculate the objective functions.

4. Experiment

4.1 Dataset

The dataset used in this paper is a public dataset published in AI Studio, based on game reviews from the Chinese mobile game forum TapTap. The dataset collects 4888 reviews from about 300 games. Reviews on TapTap are divided into five levels from one star to five stars, and this dataset regards reviews below three stars as negative and marked as 0, and others as positive and marked as 1.

To improve the fine-tuning efficiency, this paper removes the excessively long texts in the dataset that are longer than 512, and divides the original dataset into training set, validation set, and testing set in the ratio of 8:1:1. The processed dataset has a total of 4279 data, including 2260 negative reviews and 2019 positive reviews. The text length distribution is shown in Table 1. It can be seen that the text length distribution of this dataset varies widely, which is consistent with the realistic game reviews.

Table 1. The text length distribution.

	Length
mean	180.6130
std	123.7908
min	31.0000
25%	74.0000
50%	154.0000
75%	254.0000
max	512.0000

Examples selected in the training dataset are shown in Table 2. As can be seen from the examples, the sentiments of game reviews are relatively hidden. Some reviews do not have obvious sentiment words, but a clear sentiment tendency can be felt during reading. Not only do negative reviews contain a lot of metaphors and sarcasm, but also positive reviews may have suggestions that do not look so positive. Therefore, it is difficult to extract sentiment information from these game reviews.

Table 2. English translation of some examples in the training dataset.

Review	Sentiment polarity
If you really integrate all three games, it would be great, but this copying is too low-level.	negative
... this familiar painting style (Don't Starve)... this familiar name (MC)... one star is enough.	negative
I want to give a rating of four and a half stars, but unfortunately it's not possible. The small robot is very cute and the game is also very interesting, but when paying attention to how to make the small robot reach the finish line, it is also important to be careful not to let the small robot be hit by the mechanism out of the map haha. I gave half a star less because of the channels. Some channels are too dark, and the little robots just "disappear" when they enter the channel. I turn the map around and look for the little robots.	positive
Small as the sparrow is, it possesses all its internal organs. This game restores real chess, but it cannot be online. I hope to create a friend system and a friend battle system.	positive
No, can't you have the heart to score a point for such a conscientious game? It must be a five-star review. But using only one account to make a five-star rating cannot express my sincerity. I will give you a full five-star rating in five accounts, one by one.	negative

4.2 Experiment setting

In this paper, we use a 12-core Tesla A100 for model training and prediction. For hyperparameters, the batch size is set to 8, max sequence length is set to 512, learning rate is set to 2e-5, epochs is set to 4, warmup proportion is set to 0.1, and weight decay is 0.01.

4.3 Results

In this paper, the F1 score is used as an evaluation criterion which is formulated as formula 6, formula 7, and formula 8.

Confusion Matrix		Predicted Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

Figure 2. Confusion matrix.

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$\frac{1}{F1_{score}} = \frac{1}{2} * \left(\frac{1}{P} + \frac{1}{R} \right) \quad (8)$$

In formula 8, P represents Precision in formula 6 and R represents Recall in formula 7.

We use ERNIE1.0, RoBERTa, MacBERT, BERT-wwm, and Nezha models provided by PaddleNLP for comparison studies. The results are shown in Table 3.

Table 3. The experiment results, where SKEP represents the SKEP pre-trained model used in this paper.

Model	Precision	Recall	F1 score
ERNIE1.0	0.8331	0.8321	0.8326
RoBERTa	0.8117	0.8057	0.8077
MacBERT	0.8224	0.8187	0.8201
BERT-wwm	0.7979	0.8003	0.7985
Nezha	0.8024	0.8046	0.8030
SKEP	0.8659	0.8606	0.8625

It can be seen that SKEP is significantly better than other models, whose F1 score on average exceeds other models by about 0.05. Particularly, ERNIE1.0 is the original ERNIE model without the SKEP method, and its prediction result is inferior to that of the ERNIE model with SKEP, which is the model used in this paper, confirming the effectiveness of SKEP.

5. Conclusion

Game reviews are an important source of information for game makers to make decisions, and how to mine and analyze game reviews efficiently is a problem that needs to be solved in the game industry. In this paper, we use PaddleNLP's SKEP pre-trained model to perform sentiment analysis on a Chinese game review dataset and use several models for comparison experiments.

By introducing unsupervised sentiment knowledge mining methods, improving the masking strategy, and proposing new pre-training objectives, SKEP specializes the ability of sentiment analysis. SKEP achieves a better result with an F1 score of 0.8625, which on average exceeds other models used for comparison by about 0.05, demonstrating that SKEP has superior performance on this problem and a strong potential for sentiment analysis for game reviews. In addition to the sentence-level sentiment classification used in this paper, SKEP can also perform both aspect-level sentiment classification and opinion target extraction, which means that the application of SKEP on game reviews is very promising. In conclusion, SKEP will be beneficial to make the analysis of game reviews more intelligent and automated and help upgrade the game industry.

6. Discussion

Although we have taken a very large number of factors into account, the model can continue to be enhanced, and this is where we come in later.

On the one hand, there are fewer open datasets for sentiment analysis of Chinese game reviews, and thus the dataset used in this paper has some limitations. In the process of sentiment labelling, the dataset used in this paper treats reviews with less than three stars as negative reviews and labels them as 0, while others as positive reviews and labels them as 1. However, the corresponding relationship between stars and sentiment may be different for different players, and such a division may blur part of the sentiment information. A direct five-category sentiment analysis problem according to star ratings may achieve better results, but since the dataset does not provide raw data, this work can only be done on other data sources.

On the other hand, game reviews themselves have special characteristics. The heavy use of sarcasm, suggestions, and comparisons with other games in game reviews creates difficulties in

correctly determining the sentiment tendency of game reviews [23]. Moreover, different game products and game communities may have specific sentiment words, which means one word may have a significant sentiment tendency for one game, but not for others. In this case, SKEP's unsupervised sentiment knowledge mining method can effectively mine the sentiment tendency of such words. However, since all texts use the same set of sentiment words, this may also lead to incorrect classification. Such a phenomenon may also occur in other domains of sentiment analysis.

References

- [1] Game Publishing Committee of CADPA. (2023). China Gaming Industry Report(2022). China Audio-video and Digital Publishing Association. <http://www.cadpa.org.cn/3271/202302/41574.html> W. Strunk Jr., E.B. White, The Elements of Style, third ed., Macmillan, New York, 1979.
- [2] Jana, R. K., & Maity, S. (2022). An Accuracy based Comparative Study on Different Techniques and challenges for Sentiment Analysis. *Pervasive Computing and Social Networking: Proceedings of ICPCSN 2022*, 601-619.
- [3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- [4] Tian, H., Gao, C., Xiao, X., Liu, H., He, B., Wu, H., ... & Wu, F. (2020). SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. *arXiv preprint arXiv:2005.05635*.
- [5] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- [6] Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. (2019). ERNIE: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*.
- [7] Zagal, J. P., Ladd, A., & Johnson, T. (2009, April). Characterizing and understanding game reviews. In *Proceedings of the 4th international Conference on Foundations of Digital Games* (pp. 215-222).
- [8] Rajapakshe, U. (2019, May). Development Centric Player Feedback Analysis for Video Games: A Review. In *2019 International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS)* (pp. 190-194). IEEE.
- [9] Lin, D., Bezemer, C. P., Zou, Y., & Hassan, A. E. (2019). An empirical study of game reviews on the Steam platform. *Empirical Software Engineering*, 24, 170-207.
- [10] Ruseti, S., Sirbu, M. D., Calin, M. A., Dascalu, M., Trausan-Matu, S., & Militaru, G. (2020). Comprehensive exploration of game reviews extraction and opinion mining using nlp techniques. In *Fourth International Congress on Information and Communication Technology: ICICT 2019, London, Volume 1* (pp. 323-331). Springer Singapore.
- [11] Secui, A., Sirbu, M. D., Dascalu, M., Crossley, S., Ruseti, S., & Trausan-Matu, S. (2016). Expressing sentiments in game reviews. In *Artificial Intelligence: Methodology, Systems, and Applications: 17th International Conference, AIMSA 2016, Varna, Bulgaria, September 7-10, 2016, Proceedings 17* (pp. 352-355). Springer International Publishing.
- [12] Vieira, A., & Brandão, W. (2019, September). Evaluating Acceptance of Video Games using Convolutional Neural Networks for Sentiment Analysis of User Reviews. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media* (pp. 273-274).
- [13] Kaan, A. R. I. K. (2022). Social Media Content Review of Popular MMORPG Games: Reddit Comment Scraping and Sentiment Analysis. *Journal of Emerging Computer Technologies*, 2(1), 13-21.
- [14] Xiang, N., qian Jia, Q., & dong Wang, Y. (2021, July). Sentiment analysis of Chinese Weibo combining BERT model and Hawkes process. In *Proceedings of the 2021 5th International Conference on Deep Learning Technologies* (pp. 59-65).
- [15] Li, M., Chen, L., Zhao, J., & Li, Q. (2021). Sentiment analysis of Chinese stock reviews based on BERT model. *Applied Intelligence*, 51, 5016-5024.
- [16] Luo, J., Huang, S., & Wang, R. (2021). A fine-grained sentiment analysis of online guest reviews of economy hotels in China. *Journal of Hospitality Marketing & Management*, 30(1), 71-95.

- [17] Luo, L., Yang, Z., Yang, P., Zhang, Y., Wang, L., Lin, H., & Wang, J. (2018). An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics*, 34(8), 1381-1388.
- [18] Gai, Z., Fan, C., Shen, S., Ge, Y., Shi, Z., Li, S., ... & Cao, Y. (2022). Using Social Media Data to Explore Urban Land Value and Sentiment Inequality: A Case Study of Xiamen, China. *Wireless Communications & Mobile Computing*, 2022.
- [19] Cui, Y., Che, W., Liu, T., Qin, B., & Yang, Z. (2021). Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 3504-3514.
- [20] Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., & Hu, G. (2020). Revisiting pre-trained models for Chinese natural language processing. *arXiv preprint arXiv:2004.13922*.
- [21] Wei, J., Ren, X., Li, X., Huang, W., Liao, Y., Wang, Y., ... & Liu, Q. (2019). Nezha: Neural contextualized representation for chinese language understanding. *arXiv preprint arXiv:1909.00204*.
- [22] Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *arXiv preprint cs/0212032*.
- [23] Viggiato, M., Lin, D., Hindle, A., & Bezemer, C. P. (2021). What Causes Wrong Sentiment Classifications of Game Reviews?. *IEEE Transactions on Games*, 14(3), 350-363.