Aspect-Based Sentiment Analysis on Mobile Game Reviews Using Deep Learning

Jiaxin Song
Graduate School of Computer and Information Sciences
Hosei University
jiaxin.song.5x@stu.hosei.ac.jp

Abstract—This paper proposes an aspect-based sentiment analysis method on mobile game reviews using deep learning, which can make better use of massive mobile game reviews data to judge users' emotional tendencies for different attributes of the game at a fine-grained level. Specifically, there are three models in our sentiment analysis method. The baseline model includes Bi-LSTM, FCN, and CRF for sentiment collocation extraction, matching, and classification. The iterative model updates the neural network structure and effectively improves the model's recall rate in the experiments. The joint model is based on the information passing mechanism and further improves the comprehensive performance of the model. We crawled more than 100,000 game review items from two well-known Chinese game review websites Bilibili and Taptap and manually annotated 3,000 items to construct the experiment dataset. Several experiments have been carried out to evaluate our methods. The experimental results show that our methods have achieved good results.

Keywords—Aspect-Based Sentiment Analysis, LSTM, FCN, CRF, Deep Learning.

I. INTRODUCTION

The mobile games have created considerable economic benefits in recent years. Game developers all around the world have published a large number of mobile games and users can download them easily in the app stores. Users can also submit their reviews about the games in these platforms conveniently. These massive mobile game reviews are important for game developers.

In order to effectively mine the value of the massive review data, researchers have study review analysis in different fields, such as movie reviews [1], [2], micro-blogs [3]–[5], and electronic commerce [6], [7]. Specifically in the field of game reviews, existing work [8]–[10] only uses some lexical-level features such as word frequency, combining with traditional learning-based methods to judge the sentiment of users, which cannot fully mine the emotional tendency.

This paper presents an aspect-based sentiment analysis method which can make use of massive review data to efficiently judge the sentiment tendency of users. To demonstrate the effectiveness of our approach, we crawl a large number of reviews from Bilibili [11] and Taptap [12] and build a comprehensive mobile game review dataset. Then we conduct extensive experiments for evaluation. Experimental results show that our approach achieves good performance.

The contributions of this paper can be summarized as follows:

- We present a deep learning-based sentiment analysis approach in the field of mobile game reviews, which is fresh and valuable. We can find more interesting conclusions via the latest dataset that we build and the meaningful classification rules that we design, helping game developers better utilize game review data to guide their future development.
- We design an aspect-based sentiment analysis method on mobile game reviews and propose three models for attribute-sentiment collocation extraction, matching, and classification. Compared with the baseline model, the iterative model and the joint model can effectively improve the accuracy.

This is the extended abstract of an extended version of the master's thesis that the author previously submitted to Wuhan University [13] for double degrees. The Wuhan version of the thesis proposed the system including sentiment analysis models (Section IV), semantic annotation rules and evaluation (Subsection VI-A), and presented the results of experiments on annotated dataset evaluation and model validity evaluation (Subsection VI-B). This Hosei version of the thesis additionally provides the testing of the processing of obscure words (Subsection VI-C) and the design of a new evaluation method based on statistical test (Subsection VI-D), which further verifies the advantages of the proposed method in the mobile game reviews sentiment analysis field.

II. RELATED WORK

A. *Aspect-Based Sentiment Analysis*

Aspect-Based Sentiment Analysis (ABSA) [14] can be divided into two sub-tasks: the first is the task of aspect description extraction and classification, and the second is sentiment analysis on the aspect level. ABSA is more fine-

grained than text-level and sentence-level sentiment analysis methods, and can obtain more accurate and detailed information from user reviews.

The goal of aspect description extraction is to extract and identify the evaluation objects from the reviews through a series of operations. For the product, the attribute description is the performance characteristics of the product in many aspects, such as appearance, function, price, and weight. In game review data, players often evaluate the various characteristics and attributes of a game from multiple perspectives and aspects of the game, such as game design, plots, pictures, sound effects, consumer experience, and technical level. Each attribute evaluation has good or bad, positive or negative, and other emotional tendencies.

At present, sentiment word extraction and classification methods are always using sentiment dictionary. This kind of methods depends on the emotional dictionary, and its quality plays a decisive role in the judgment of emotional polarity. Compared to sentiment dictionary, methods based on neural networks are more efficient and more intelligent because of the strong learning ability of neural networks. Therefore, the methods based on neural networks have become the mainstream methods on ABSA tasks recently.

B. *Sentiment Analysis on Reviews*

In terms of movie reviews, Kumar *et al.* [1] used word frequency and lexicon as features, combining them with traditional machine learning methods. Yenter *et al.* [2] employed a Convolutional Neural Network (CNN) and Long Short-term Memory (LSTM) to handle movie reviews. In terms of electronic commerce, Yang *et al.* [6] modeled topic-sentiment correlation of customer reviews as graphical representation. Zhang *et al.* [7] proposed a trust representation model to conduct sentiment similarity analysis.

In the field of game reviews, the problem is more complicated. Most of the users of games are young people, who like short and personalized expressions. Therefore, game reviews often contain a large number of obscure words and difficult to grammatically define. Traditional emotional analysis methods are inefficient in dealing with such reviews. In this field, Zhu *et al.* [8] used lexical analysis for review analysis. Stråat *et al.* [9] proposed a manual aspect-based sentiment analysis approach. Existing work [8]–[10] only uses some lexical-level features such as word frequency, combining them with traditional learning-based methods to judge the sentiment of users, which cannot fully mine the emotional tendency. In contrast, our approach is based on the recent advances in deep learning, which can fully utilize the massive review data and better judge the emotional tendency of users.

III. PRELIMINARIES

A. *Word2vec*

Word2vec [15] is a class of neural network models that can produce for each unique word a corresponding vector in a continuous space in which the linguistic contexts of words can be observed. The distributed representation produced by Word2vec is better than the traditional one-hot representation. It can reflect the context of the words, making words with similar context closer in the embedding space. In this paper, Word2vec is used for data pre-training.

B. *Long Short-term Memory*

Long Short-term Memory (LSTM) is a variant of a Recurrent Neural Network (RNN) used in the field of deep learning, while the latter is typically designed for time series or sequence data.

Bidirectional Long Short-term Memory (Bi-LSTM) is an improved variant of traditional LSTM. In practice, it includes two LSTM layers (Fig. 1). One layer uses the direction of sequence as input, while the other uses the reverse. Bi-LSTM combines the outputs of two LSTM layers as the final output.

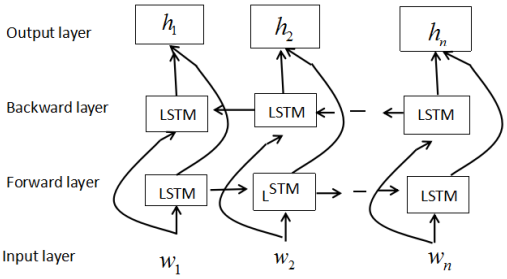

Fig. 1. The structure of LSTM

The basic idea of Bi-LSTM is that for each input sequence, two independent hidden layer states are used to record the early and late information of the current time, and then the information obtained by the two hidden layer states is combined as a result output. The calculation formula is as follows:

$$h_{t1} = \overrightarrow{\mathrm{LSTM}}(x_t)$$
$$h_{t2} = \overleftarrow{\mathrm{LSTM}}(x_t)$$
$$h_t = W_1 h_{t1} + W_2 h_{t2}$$

In the formula, the vectors refer to forward and backward layers of LSTM, the output h_t combines the processing results of forward and backward networks. This feature makes it handles historical information while taking future information into account, greatly enhancing the ability of LSTM to capture contextual relationships.

IV. PROPOSED METHOD

Previous research [1], [3]–[10] is mostly based on sentiment dictionaries and traditional machine learning algorithms. These methods can only extract the aspects defined in the sentiment dictionaries, which has great limitations, especially when dealing with data such as “nijigen” (a kind of art style originated from Japanese animation and is popular among teenagers) mobile game reviews, which contain a lot of fresh words and less common terms.

In this paper, deep learning technology is used to build an artificial neural network with learning ability. Without the

need to manually build rules, aspects can be extracted without collecting and constructing a sentiment dictionary. The model is smarter, more automated, and more efficient. This paper proposes a method including three models for extracting and classifying the collocation of aspect and opinion description, i.e., the baseline model, the iterative model, and the joint model. The baseline model includes Bi-LSTM and FCN neural networks and classification methods such as CRF. The iterative model improves the structure of the baseline model's extraction task and effectively improves the recall rate of the model. The joint model improves the overall structure of the model and further improves the comprehensive performance of the model.

A. *Baseline Model*

The baseline model is divided into four parts: the pre-training task, the extraction task, the matching task, and the classification and grading task (Fig. 2). Since the pre-training tasks of different models are almost the same, this part will mainly introduce the extraction task, the matching task, and the classification task.

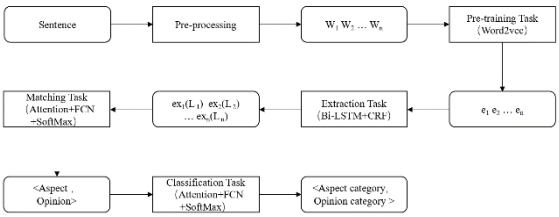

Fig. 2. The Structure of the Baseline Model

1) Extraction Task

The extraction task is used to extract the attribute description words (i.e., aspect) and the sentiment description words (i.e., opinion) from the sentence after word segmentation. The process can be regarded as a task of classifying words in the sentence. In the extraction task (Fig. 3), the first part uses the Bi-LSTM as the input of the pre-trained word vector sequence $\{e_1, e_2, \ldots, e_n\}$, and the output is a new distributed representation vector sequence $\{ex_1, ex_2, \ldots, ex_n\}$, the sequence vector contains not only the information of each word in the original input sequence, but also the information of the words around the word. That is, through the Bi-LSTM network, the model considers its context information when processing each word. In the second part, a fully connected neural network (FCN) is used to map the representation of words from the semantic space to the classification space by nonlinear transformation. The conditional random field (CRF) algorithm converts the representation of a word in the classification space into a label of a word by an approximate global optimization method, and outputs a label sequence $\{L_1, L_2, \ldots, L_n\}$.

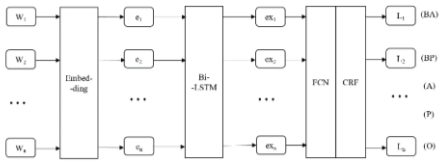

Fig. 3. The Structure of the Extraction Task

In Fig. 3, "W" indicates the independent word in the sentence after the pre-processing word segmentation; "e" indicates the word vector processed by the word embedding; "OF" and "OB" refer to the forward neural network (LSTM-F) and the output to the neural network (LSTM-B); "ex" indicates the word vector carrying the context information feature after processing with the Bi-LSTM neural network; "L" indicates the five types of BIO tag information processed by the FCN and CRF layers. Finally, the BIO tag information and the tagged word vector are used as the output of the extraction task.

2) Matching Task

The matching task is used to calculate the matching relationship between the attribute description and the sentiment description. In the matching task (Fig. 4), each attribute description in the sentence is matched with each sentiment description, that is, from the vector sequence $\{ex_1, ex_2, \ldots, ex_n\}$ generated by the extraction layer. The attribute description and the sentiment description that need to be matched are selected according to the BIO label. When the matching relationship between the attribute description and the sentiment description is predicted to be "0", it is considered that the two do not match, and the next operation is not performed. When the matching relationship between the attribute description and the sentiment description is predicted to be "1", the attribute description and the sentiment description evaluation pair are output to the classification task to complete the attribute classification and the sentiment classification.

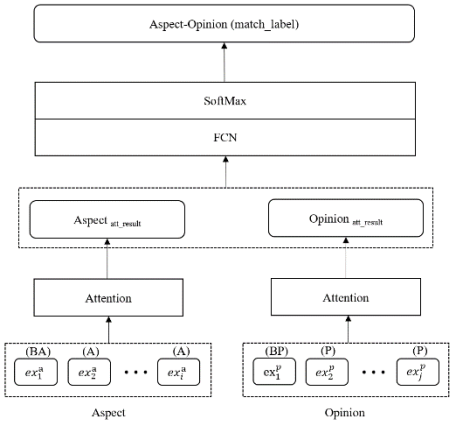

Fig. 4. The Structure of the Matching Task

3) Classification Task

The classification task (Fig. 5) is used to classify the attribute description and classify the sentiment description, which can be abstracted into the task of multi-label classification of the phrase. In the actual language environment, attribute description and sentiment description are semantically mutually auxiliary expressions. Therefore, in the attribute classification and sentiment classification tasks, it is necessary to simultaneously use the matched pair ⟨attribute description, sentiment description⟩ as input, and complete the attribute classification and sentiment classification tasks independently through two networks

with different structural parameters, and output the evaluation.

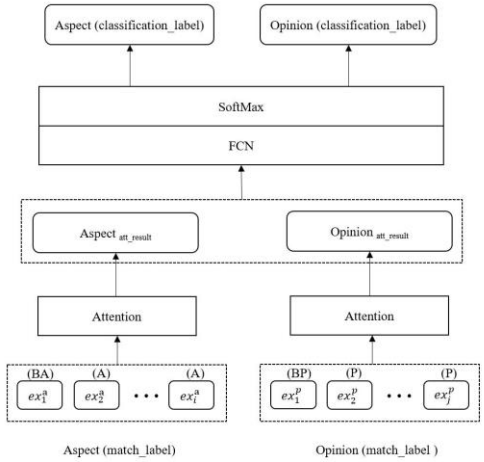

Fig. 5. The Structure of the Classification Task

B. *Iterative Model*

The iterative model is divided into four parts: the pre-training task, the extraction task, the matching task, and the classification and grading task (Fig. 6). Since the pre-training tasks of different models are the same, and the extraction task and matching task structure of the iterative model are basically the same as the baseline model, this section will mainly introduce the extraction task of the iterative model.

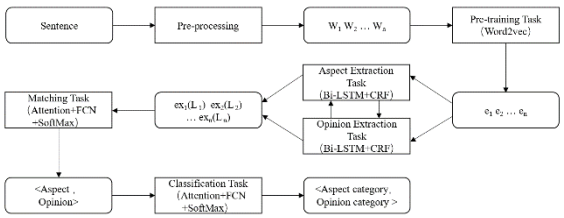

Fig. 6. The Structure of the Iterative Model

In the extraction task of the iterative model, two independent Bi-LSTM-FCN-CRF neural networks are established, one for predicting the attribute description (aspect), which is the classification of A, BA, and O, and the other for predicting sentiments description (opinion), which is a classification of P, BP, and O. Since attribute descriptions and sentimental descriptions usually appear in pairs, that is, users are expressing their sentimental views on certain attributes of an item. Therefore, the extraction process of the two can be considered. The extraction of the attribute description can improve the extraction of the sentiment description. The extraction of the sentiment description can also improve the extraction performance of the attribute description. In this paper, the model design will be carried out according to the above settings. The output of the attribute description prediction network is used as the input of the sentiment description prediction network. At the same time, the output of the sentiment description prediction network is used as the input of the attribute description prediction network, and the attribute description and the sentiment description label are fully utilized. This process combines Bi-LSTM's contextual information integration capabilities to increase recall rates.

C. *Joint Model*

Aspect-based sentiment analysis tasks are usually done in the pipeline mode, first performing attribute description extraction, and then performing sentiment analysis on the extracted attribute descriptions. Although easier to implement, this approach does not take advantage of the joint information from the two subtasks and does not use all of the information sources available for training. For example, in the above neural network model, the information of the extraction task, the matching task, and the classification task may be related, which may help the training or prediction of other subtasks.

Inspired by the Message Passing Architecture of the Interactive Multi-task Learning Network (IMN) proposed by He et al. [16], we propose a multi-task joint neural network model (Fig. 7). This model allows informational interactions between multiple layers of tasks through an information transfer mechanism, and information is passed to different tasks through a shared set of potential vector iterations.

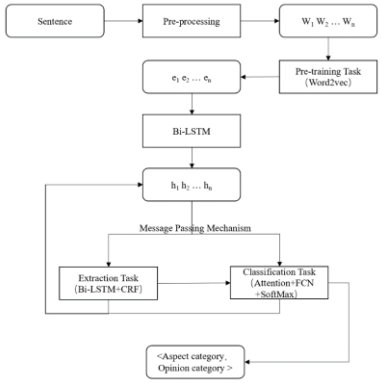

Fig. 7. The Structure of the Joint Model

The shared hidden vector sequence $\{h_1, h_2, \ldots, h_n\}$ is used as input to specific tasks for different tasks, combining information from different tasks. The value of the output vector is used as the prediction result after 3 iterations of information passing. Since the structure of the joint model has been combined with the input information of the extraction task and the classification task, in the actual experiment, it is observed that the information matching the task in the baseline model will interfere with the extraction task and the classification task, resulting in negative impact on the model efficiency. This paper omits the matching layer task when applying to the federated model. The extraction task and the classification task can be mutually influenced and combined more efficiently.

V. IMPLEMENTATION

The overall workflow is displayed in Fig. 8. The method is mainly composed of four parts, data processing, model training, model prediction, and result evaluation. This paper uses Python v3.7 and TensorFlow v1.13 to implement the

model. SPSS Statistics v23 is used for statistical analysis.

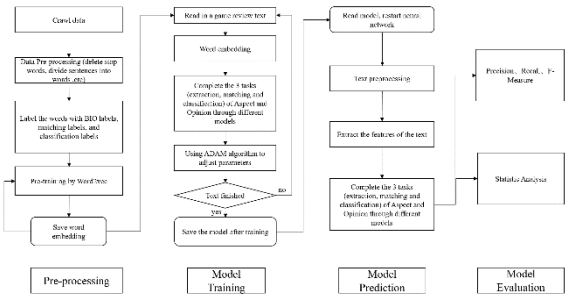

Fig. 8. The Workflow of the Project

VI. EVALUATION

A. Dataset

In this paper, we crawled a total of 100,000 reviews from 6 mobile games on Bilibili and Taptap, and selected 3,000 reviews for manual annotation. The aspect description in each review is marked as five types, Game Play (GP), Hybrid Aesthetics (HA), Culture Icon (CI), Consumption Experience (CE), and Technical Implementation (TI). Opinion description is marked as three scores: 1 means negative, 2 means neutral, and 3 means positive. 80% of the experimental data set are used as the training set, and 20% are used as the testing set.

B. Model Accuracy

As shown in Fig. 9, the classification result accuracy of each model is more than 75%. The iterative model improves the extraction task of the baseline model, so that the recall rate is increased to about 85%. The joint model further improves the overall structure, its F1-measure is about 89%, which means the comprehensive performance of the model is further improved. The experimental results are consistent with expectations, which means that the model accuracy meets the requirements of practical application.

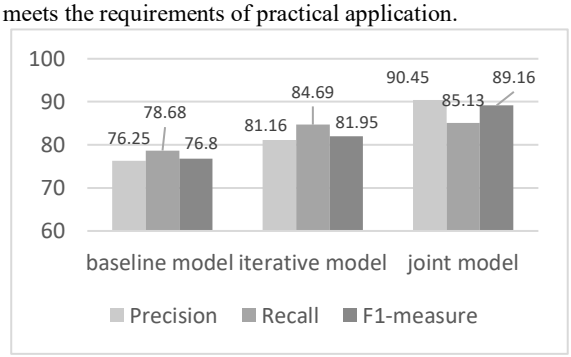

Fig. 9. The Experimental Results of Three Models.

C. Test Case

In order to test the processing of obscure words, we take a review as sample α to explain the prediction process of the neural network model. As shown in Fig. 10, each number in the vector sequence $\{e_1, e_2, \ldots, e_n\}$ is the index of each word in the vocabulary list. Among them, "0" means that the word is not in the vocabulary list, which is a word that does not appear in the training set or appears very infrequently. Those obscure words are uniformly set as "UNK". In this test case, we can observe that the model can identify an unknown word through context information, then extract the word or phrase as an attribute description, and pair it with the corresponding sentiment description, and finally correctly predict the classification of this collocation.

Form	Sample α
	Pre-processing
Original Senctence	{角色设计，立绘配音，游戏性挺不错；交易系统基建系统还是不好玩。希望之后公测能完善一些鸡肋的小细节。}
$\{W_1, W_2 \ldots W_{16}\}$	{'角色', '设计', '立绘', '配音', '游戏性', '挺不错', '交易系统', '基建', '系统', '不好玩', '希望', '之后', '公测', '完善', '鸡肋', '小', '细节'}
	Pre-training
$\{e_1, e_2 \ldots e_{16}\}$	{55, 360, 194, 1061, 861, 1668, 0, 1708, 149, 2291,35, 105, 298, 1433, 2087, 117, 846} #0=not in vocabulary
	Extraction Task
$\{L_1, L_2 \ldots L_{16}\}$	{1, 2, 1, 1, 1, 3, 1, 2, 2, 3, 0, 0, 0, 0, 0, 0, 0}
Aspect (Attribute Description)	{'角色', '设计'}, {'立绘'}, {'配音'}, {'游戏性'}, {'UNK', '基建', '系统'}
Opinion (Sentiment Description)	{'挺不错'}, {'不好玩'}
	Matching Task
<Aspect,Opinion>	input {'UNK', '基建', '系统', '不好玩'} ; output 1
	input {'角色', '设计', '不好玩'} ; output 0
	Classification Task
<Aspect,Opinion>	input {'UNK', '基建', '系统', '不好玩'}
	output {0.3699155 0.2569093 0.147147 0.19643514 0.02959312} #Probability
	output 1, GP
<Aspect category, Opinion category>	output {GP, 1}

Fig. 10. The Process of Sample α

D. Statistical Results

The dataset for statistical test contains 500 reviews randomly selected from the original dataset. 10% of invalid reviews were deleted. The user score in the original data is retained as the real score, while the average sentiment score of each aspect description predicted by model is defined as the calculated score. This paper makes a statistical analysis of them in order to find interesting conclusions. In this paper, Wilcoxon Signed Rank Test and Kruskal-Wallis Test in nonparametric test are carried out by using SPSS for statistical test.

The Wilcoxon Signed Rank Test result shows that the median of difference between the real score and the calculated score equals 0 (Fig. 11), which means the distribution of the calculated score around the real score is relatively uniform, in line with expectations.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between real_score and cal_score equals 0.	Related-Samples Wilcoxon Signed Rank Test	.267	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Fig. 11. The Result of the Wilcoxon Signed Rank Test

The Kruskal-Wallis Test result shows that the real score and the calculated score have the different distribution (Fig. 12). The "Abscissa" real_score in the figure is the real score, and the "Ordinate" cal_score shows the distribution of the calculated score. Because the review website provides a 5-point scale, users can only submit five kinds of scores and the distribution of the real score is discrete, while the calculated score contains more information. We can also observe from the figure that most of the calculated score is

distributed around the real score, showing the same trend as the real score, consistent with our expectation that the calculated score includes more detailed information on the basis of the real score.

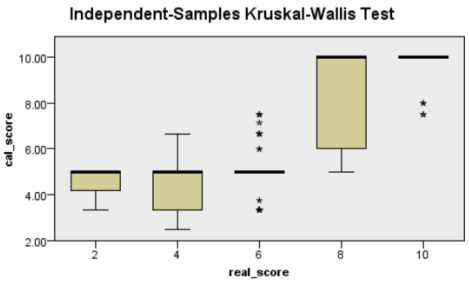

Fig. 12. The Result of the Kruskal-Wallis Test

VII. DISCUSSION

Through the experiments, we can find that the proposed method has many advantages. It can do the implementation of extraction and categorization at a fine-grained level and with a high efficiency. Compared with the baseline model, the iterative model improves the recall rate effectively, and the joint model further improves the comprehensive performance.

The methods also have some disadvantages. For example, the annotation dataset is obtained manually and the dataset is relatively small. There are some new deep learning techniques to handle the text classification task. In the future, we can use more novel deep learning algorithms to improve our method.

VIII. CONCLUSION

In this paper, we presented an aspect-based sentiment analysis approach for mobile game reviews, which can fully utilize the massive data and better judge the emotional tendency of users. To evaluate the effectiveness of our approach, we built a comprehensive mobile game review dataset and conducted extensive experiments for evaluation. The experiments showed that our approach achieved good performance.

Currently, the annotation data is only a small scale, and we need to annotate more data in the future. We can also use more deep learning algorithms to improve the performance, such as graph convolutional networks. Another future direction is to adapt the models to another domain.

REFERENCES

[1] H. M. K. Kumar, B. S. Harish, and H. K. Darshan, “Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method,” *IJIMAI*, vol. 5, no. 5, pp. 109–114, 2019.

[2] A. Yenter and A. Verma, “Deep CNN-LSTM with combined kernels from multiple branches for IMDb review sentiment analysis,” in *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, 2017, pp. 540–546.

[3] R. M., V. R. Hulipalled, K. R. Venugopal, and L. M. Patnaik, “Consumer insight mining: Aspect based Twitter opinion mining of mobile phone reviews,” *Appl. Soft Comput.*, vol. 68, pp. 765–773, Jul. 2018.

[4] J. Jmal and R. Faiz, “Customer Review Summarization Approach Using Twitter and SentiWordNet,” in *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, 2013, pp. 33:1--33:8.

[5] F. H. Khan, S. Bashir, and U. Qamar, “TOM: Twitter opinion mining framework using hybrid classification scheme,” *Decis. Support Syst.*, vol. 57, pp. 245–257, Jan. 2014.

[6] Q. Yang, Y. Rao, H. Xie, J. Wang, F. L. Wang, and W. H. Chan, “Segment-level joint topic-sentiment model for online review analysis,” *IEEE Intell. Syst.*, vol. 34, no. 1, pp. 43–50, 2019.

[7] S. Zhang and H. Zhong, “Mining Users Trust From E-Commerce Reviews Based on Sentiment Similarity Analysis,” *IEEE Access*, vol. 7, pp. 13523–13535, 2019.

[8] M. Zhu and X. Fang, “A Lexical Analysis of Nouns and Adjectives from Online Game Reviews,” in *Human-Computer Interaction: Interaction Technologies*, 2015, pp. 670–680.

[9] B. Stråat, H. Verhagen, and H. Warpefelt, “Probing User Opinions in an Indirect Way: An Aspect Based Sentiment Analysis of Game Reviews,” in *Proceedings of the 21st International Academic Mindtrek Conference*, 2017, pp. 1–7.

[10] B. Stråat and H. Verhagen, “Using User Created Game Reviews for Sentiment Analysis: A Method for Researching User Attitudes,” in *Proceedings of the 1st Workshop on Games-Human Interaction co-located with CHItaly 2017, the 12th Edition of the biannual Conference of the Italian*, 2017.

[11] Bilibili, “Video Sharing Website and Nijigen Cultural Community.” 2019.

[12] Taptap, “High-quality Mobile Game Recommendation and Sharing Community.” 2019.

[13] J. Song, “Sentiment Analysis of Mobile Game Review based on Deep Learning.” Master's Thesis, Wuhan University, 2019. In Chinese.

[14] H. H. Do, P. W. C. Prasad, A. Maag, and A. Alsadoon, “Deep Learning for Aspect-Based Sentiment Analysis : A Comparative Review,” *Expert Syst. Appl.*, vol. 118, pp. 272–299, 2019.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” *CoRR*, vol. abs/1301.3, 2013.

[16] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, “An Interactive Multi-Task Learning Network for End-to-End Aspect-Based Sentiment Analysis,” pp. 504–515, 2019.