

The Use of Sentiment Analysis and Latent Dirichlet Allocation Topic-Modeling (LDA) on Web Novel Content Quality Factor

Ji-Yoon Kim¹ and Chae-Kwan Lim²

¹Contents-AI Research Institute, Romantique, South Korea

²Department of Distribution Management, Tongmyong University, South Korea

cklim@tu.ac.kr

Abstract. In the content market of Korea, writers and producers of web novel contents have put more a lot of efforts into planning and creating well-selling web novel contents. They want to hire well-trained web novel content planners or user consumption analysis experts to utilize them for planning and creating web novel contents. Due to the fast changing market trend in the web novel content market and its practical and financial limitations, the existing traditional method of content creation has met some obstacles. This study conducted a web novel content quality analysis using LDA (Latent Dirichlet Allocation Topic-Modeling) and sentiment Analysis. It uses user-review data and classified data into the positive and the negative using sentiment analysis. In addition, the topics of positive data and negative ones were obtained by using the NLP (Natural Language Processing) model.

Keywords: Web novel content quality factor, Korean sentiment analysis, LDA topic modeling, deep learning

1. Introduction

The web novel content market in Korea is 600 billion won which is twice the conventional paper-back based publishing market in 2020, and 60-time growth has been made in the last seven years. Web novel contents companies are trying to secure new consumers while maintaining existing consumers for the growth of companies. In particular, they are putting in specialized manpower in creating and planning web novel contents. In addition, they are performing the work of design on new web novel contents and the related supports by analyzing existing web novel contents.

However, the amount of web novel content created with the growth of the web novel content market has also increased rapidly and there is a problem to analyze the feedback data of consumers which accumulate in large quantities. To solve these problems, web novel content companies are continuously expanding planning budget and putting new planners into creating web novel contents. However, there are limitations in manpower-based analysis, which has led to difficulties to maintain and improve the quality of new web novel contents.

In line with this, this study aimed at determining the efficiency and cost-friendly approaches using accumulated user data from the past to the present, unlike conventional creation methods of web novel contents.

Using these results, content creators can easily acquire information essential to improve the quality of web novel contents, and it is expected that information which should be considered during the creation of web novel contents can be easily obtained.

2. Literature Review

2.1. Theoretical Background

(1) LSTM (Long Short-Term Memory)

The existing Recurrent Neural Network (RNN) has a problem that the further you go through the network, the more necessary information disappears. Long Short-Term Memory (LSTM) is a model presented to solve the problem and can solve long-term dependencies.

(2) CNN (Convolutional Neural Network)

CNN (Convolutional Neural Network) uses Convolution Layer and Pooling Layer to extract the feature of input data as a proposed deep learning model to improve the recognition rate of images. Since the proposed CNN model was suggested, studies have been conducted using the characteristics of CNN models with lighter forms than Attention to process natural languages. Especially, it has characteristics that show fast learning and high performance in terms of sentiment classification.

(3) Attention

ATTENTION is proposed not only to compensate for the disadvantages of the Recurrent Neural Network (RNN) model similar to LSTM (Long Short-Term Memory) but also to pay attention to data with specific meanings. When analyzing a

particular sentence, ATTENTION gives high importance to the most important word in the sentence, and other words have low importance.

(1) LDA (Latent Dirichlet Allocation)

The Latent Dirichlet Allocation Topic-Modeling (LDA) is the most commonly used algorithm for topic modeling, and uses the topic distribution and word distribution of the inputted documents as variables. It is an algorithm that finds the optimal number of topics by using Dirichlet, which has the characteristic that the sum of all related elements is 1. The algorithm of LDA consists of the following structures.

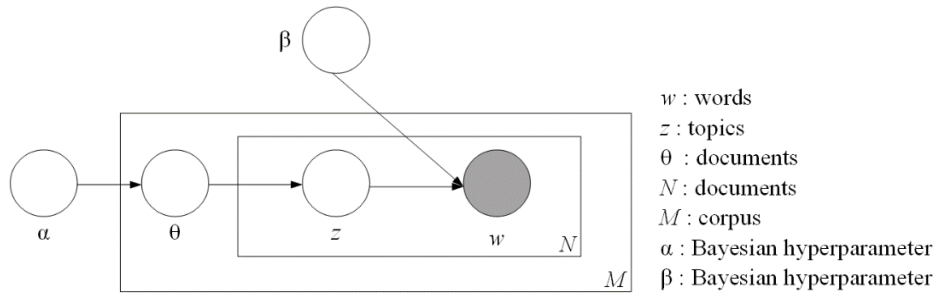


Fig. 1: Schematic of LDA model

2.2. Previous Studies

Research has been conducted to analyze the emotions of service users or ordinary people. These studies are representative of the studies that extract topics based on the sentences of users and analyze the emotions of users. The studies using traditional approaches are not using deep learning. Classic research is as follows. According to Hu, M. and Liu, B.'s research (Hu et al., 2004), as the online product market continues to grow, it is predicted that sellers would not easily analyze a large amount of accumulated user review data. Although they derived user data mining, positive and negative evaluation, and outlines of results, they argued that improvements in sentiment analysis methods in response to the use of nouns, adverbs and verbs should be made along with the improvements of pronoun recognition ability. In order to perform topic modeling of large data, (Hu et al., 2007) collected blog data and conducted a study to choose representative words as 'topic.' In the study of (Lee et al., 2009) it was proved that the correlation between nouns and products can be expressed through the high recall rate by combining only nouns among product review data. Lee, Y. J., Ji, J. H., Woo, G. and Cho, and H. G. proposed a system that allows users to distinguish topics simply by categorizing topics using clustering techniques based on similarity (Lee et al., 2009).

In addition, Song, J. S. and Lee, S. W. proposed the model of Automatic Construction of Positive/Negative Feature-Predicate Dictionary for improving the quality of opinion mining (Song et al., 2011). In the study of (Zhai et al., 2011), the study of methodology that abbreviates various words and expressions depicting the

same emotion into one feature was conducted, and they succeeded in producing the highest level of state-of-the-art by applying semi-supervised learning. In 2004, the proposal of LDA model led many researchers to analyze issues and improve products based on the model (Blei et al., 2003). C. Yang, L. Wu, K. Tan, C. Yu, Y. Zhou, Y. Tao, and Y. Song collected review data of electronic commerce products, analyzed them with LDA, and analyzed the internal variables of each product through linear regression to derive improvement points of products (Yang et al., 2021). The improvement points of offline hotel service were derived using LDA (Byeong-Cheol et al., 2020). Finally, the results of the study showed that the Yoon soo-nuk and Kim Min-chul combined big data with LDA to obtain microscopic conclusions (Yoon et al., 2020).

After AlphaGo and Lee Sedol's Deep Mind Challenge Match in 2016, various studies were conducted to apply the sentiment analysis for users by using deep learning. A. Dahou, S. Xiong, J. Zhou, M. H. Haddoud and P. Duan embedded the review data and then used the CNN model to analyze the user's sensitivity (Dahou et al., 2016).

In order to utilize the characteristics of review data with time series regression, studies were also conducted to analyze the sensitivity of reviews by combining embedded review data with LSTM and CNN (Alayba et al., 2018; Abu Farha et al., 2019; Omara et al., 2018). M. Heikal, M. Torki, and N. El-Makky proposed an approach to sentiment analysis models using deep and complex models using CNN, LSTM and CNN+LSTM structures (Heikal et al., 2018). M. Al Omari, M. Al-Hajj, A. Sabra, and N. Hammami conducted a study that proposed a new type of Hybrid CNN LSTM model for sentiment analysis only and could derive state-of-the-art performance (Al Omari et al., 2019). Unlike the above, studies were conducted on high-performance sentiment analysis model using ATTENTION utilized in RNN-based models (Vaswani et al., 2017). M. A. El-Affendi, K. Alajhi, and A. Hussain presented the Multilevel Parallel Attention Neural model for the sentiment analysis of Arabic sentences using ATTENTION (El-Affendi et al., 2021). W. Ali, Y. Yang, X. Qiu, Y. Ke, and Y. Wang conducted a study on sentiment analysis using Bidirectional-GRU model with simple structure (Ali et al., 2021).

3. Research Procedure

3.1. Morphological Analysis

To analyze natural language, morpheme analysis was performed using morpheme analyzer and KoNLPy (Korean Natural Language Processing in Python) which are commonly used in Korean research. The inputted morpheme data is classified into common nouns, adjectives, pre-nouns, and adverbs.

3.2. Data Set

To carry out this study, we collected data of generally used expressions and review data of consumers' using web novel contents. Data of generally-used expressions was collected to utilize the expressions Koreans generally use in their reviews. The review data of the content consumers' were collected to analyze the special data that the content consumers can express. Data collection was performed using Python-based Beautiful Soup. The following [Table 1] represents the data set used in the study.

Table 1: Data set

The Kind of data	The No. of Data
Labeled General Sentences	100,000
Labeled Contents Review Data	100,000

(1) Features of data set

The sentiment analysis model using data of generally-used expressions has a characteristic that it is difficult to properly analyze web novel content review data. This is because many expressions in web novel content review data are difficult to judge other than people who understand and share a specific culture and there is a characteristic that new coinages online, popular buzzwords in society, and emojis are widely used. Also, the way of expression used according to time, age, and generation is very different in the characteristics of these data. Therefore, this system is designed to build a user's sentiment analysis model by utilizing data of generally-used expressions and web novel content review data.

3.3. Analysis Method

In this section, this study proposes a system that can automatically analyze the emotions of web novel content reviewers. The following Fig. 2 shows the overall structure of the system; this system consists of three areas.

(1) User sentiment analysis

Language used by humans is characterized by creating words with new meanings over 7 time. Therefore, it is essential to develop a model that can analyze the user's emotions, including both new words that change over time and the words used by web novel content reviewers. The user emotion analysis model was trained by combining web novel content review data with general review data.

(2) Topic acquisition

Positive topics and negative topics can be analyzed through the positive and negative data by using user emotion analysis. By using the positive-negative topics, the characteristics of web novel contents that are well sold and the characteristics of web novel contents that are low in sales can be obtained.

(3) Strategy establishment

Through the analysis of the difference between the positive and negative topics, it is possible to analyze web novel contents with low sales volume to establish some strategies for content creation and its improvement.

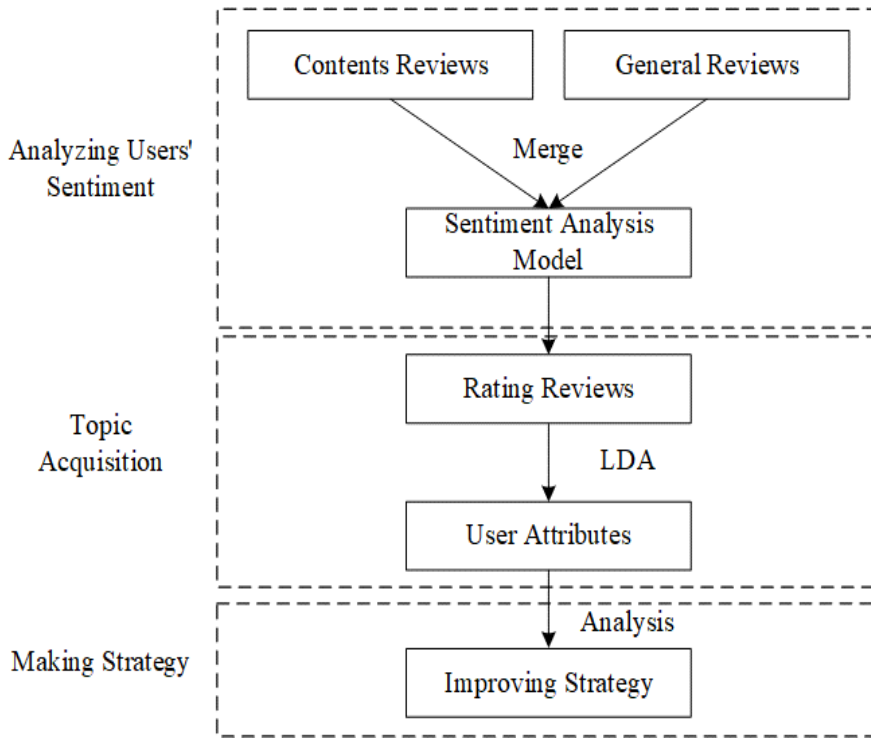


Fig. 2: Overall of methodology

3.4. User Sentiment Analysis Model

Recently, studies have been conducted to improve the accuracy of the user emotion evaluation model by applying CNN, ATTENTION, etc. In addition, most of the models have been studied based on time series encoder-decoder models. In this system, five models were adopted and used to select the model that can analyze contents review and general review the most smoothly. The following table is the type of model used in the user emotion evaluation model and the number of emotional classifications.

Table 2. Sentiment analysis model list for comparison

Type of Model	Sentiment Classes
LSTM + ATTENTION	2
CNN	2
CNN, ATTENTION and LSTM	2
CNN, CNN	2
CNN + LSTM	2

(1) LSTM + ATTENTION

This model complements LSTM with weak long-term dependencies and combines ATTENTION and LSTM to extract the characteristics of user sentences well. The following figure shows LSTM + ATTENTION model, which is analyzed by inputting embedded data and using LSTM and ATTENTION.

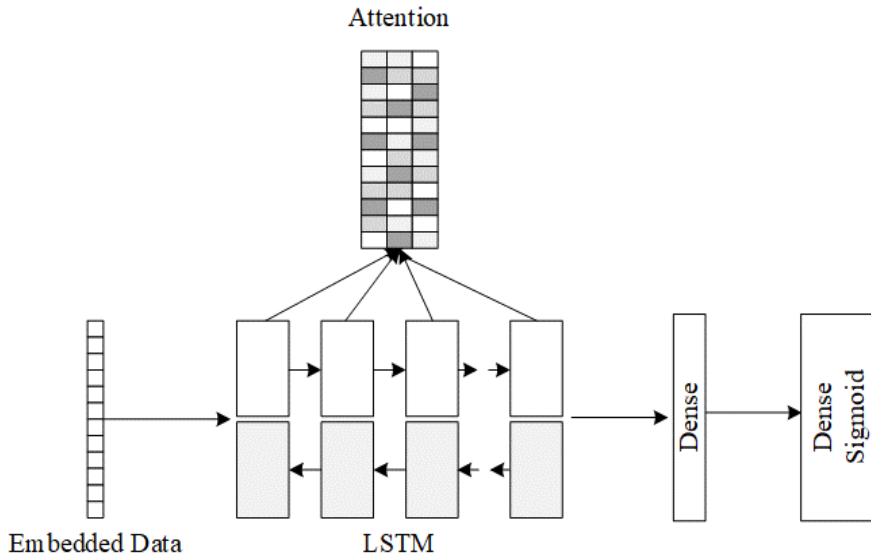


Fig. 3: LSTM+ATTENTION model structure for sentiment analysis

(2) CNN

It is a model that is widely used in sentiment analysis model. It analyzes emotional information using CNN. The feature of embedded data is drawn and the sensibility evaluation of user is performed based on the drawn result value by utilizing the extraction of feature which is the advantage of CNN model.

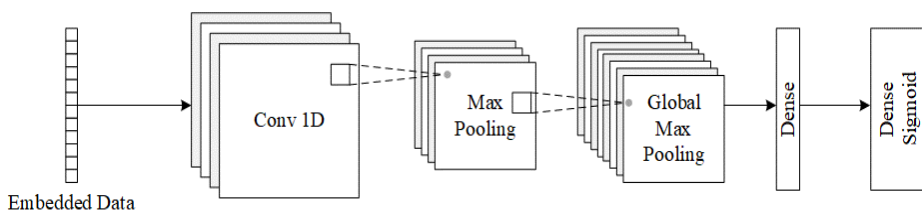


Fig. 4: CNN model for sentiment analysis

(3) CNN, ATTENTION and LSTM

CNN and ATTENTION-LSTM are made into parallel structure, and the features of embedded data are extracted, and the extracted result values (feature map) are combined in the Concatenate Layer. The combined data contains semantic data and performs the user's sensibility evaluation based on this data.

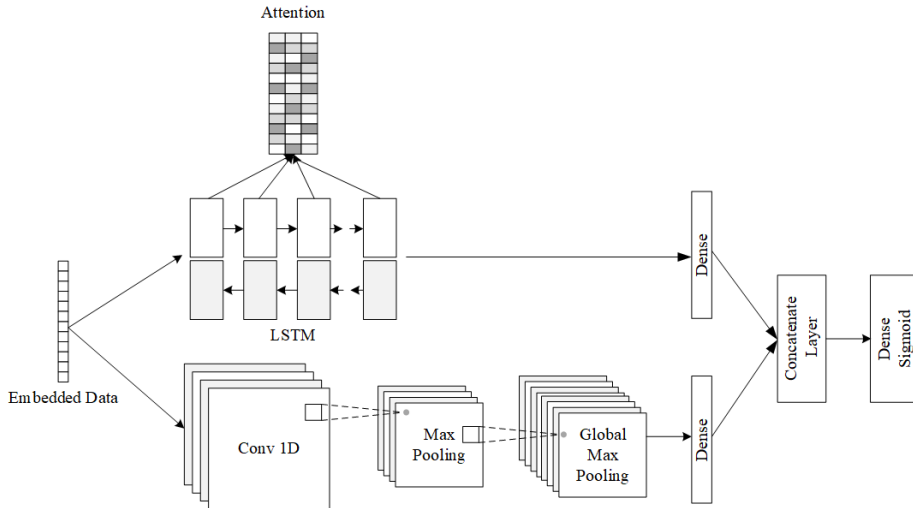


Fig. 5: CNN, attention and LSTM model structure for sentiment analysis

(4) CNN, CNN

It is a CNN parallel structure that can examine the meaning of embeddings more closely. In this structure, two CNNs contain different sizes of filters respectively and extract features of embeddings using different filters. The extracted data is combined in the Concatenate Layer, and the user's sensibility evaluation is performed using the data.

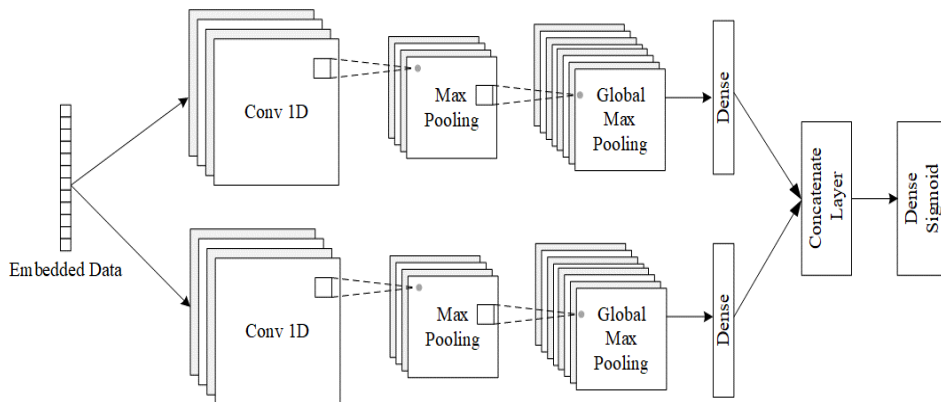


Fig. 6: CNN, CNN model structure for sentiment analysis

(5) CNN + LSTM

Embedded data have relatively low length compared to general time series data. Therefore, CNN used a model that can replace the role of ATTENTION. Embedded data extracted from CNN is inputted to LSTM and the sentiment analysis of user review data is performed based on the analysis result value of the sentence trained in LSTM.

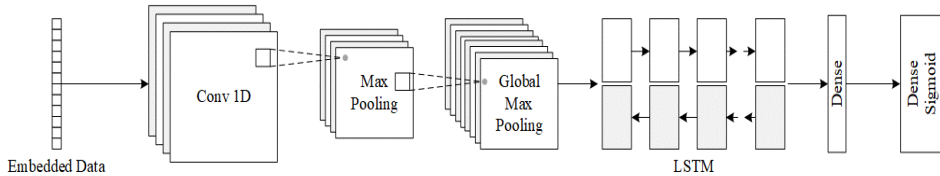


Fig. 7: CNN + LSTM model structure for sentiment analysis

(6) Model Assessment Method

To evaluate the performance of the model, accuracy, precision, recall rate, and F1 score, which are model performance indicators used traditionally or conventionally, were used. The model performance indicators are constructed as follows.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1 - score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

(7) Selecting a user emotion evaluation model

In accuracy and recall rate, CNN, CNN model showed the best performance, and in terms of recall rate and F1-Score, CNN, ATTENTION and LSTM models showed the best performance. However, given the case that content review data of new web novels occur and the characteristics of this system that require an update of the model at a rapid pace, this study uses CNN and CNN models with relatively light models. The following table shows the results of the performance experiment.

Table 3: List of experimental result of each model

Model	Accuracy	Precision	Recall	F1-Score
LSTM + ATTENTION	80.99	80.51	82.81	81.58
CNN	80.82	80.07	82.06	81.29
CNN, ATTENTION and LSTM	81.55	82.27	81.37	81.75
CNN, CNN	81.51	81.24	82.95	82.02
CNN + LSTM	80.99	80.51	82.81	81.58

3.5. Sentiment Analysis

In this study, the Latent Dirichlet Allocation (LDA) which is widely used for topic modeling was utilized to analyze the collected data. The LDA model is a model using probability, and extracts the latent themes in text data using principal component analysis. For the assessment of the topic extracted from LDA, the perplexity and consistency, which are widely used in the LDA model evaluation, were utilized. Since the current perplexity does not always give a clear answer, it has a certain level of perplexity and consistency, and the number of topics optimized is obtained. For LDA model evaluation, the number of topic models was changed from 1 to 20. The following figure shows the number of topics selected in the positive review. If the number of topics is small, it can be seen that the consistency is high. Generally, LDA topic modeling has the characteristic of selecting the number of topics within 10. However, considering the characteristics of web novel content data that need to analyze various topics, six topics were selected, which are the number of topics that can extract the appropriate number of perplexity and consistency.

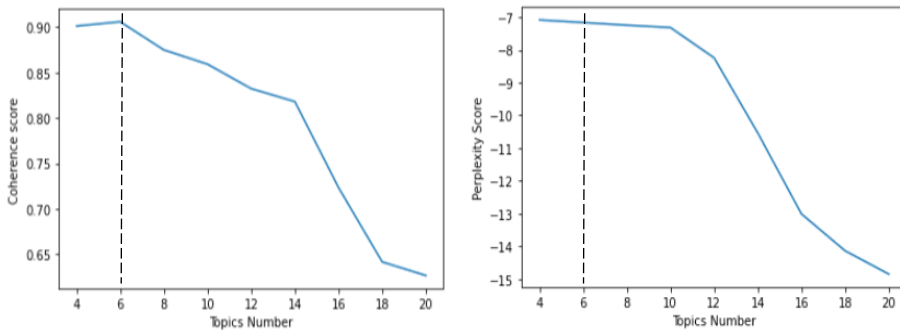


Fig. 8: Optimal number of positive topics

The following figure shows the number of topics selected from negative review. Negative review topic shows a characteristic that is continuously and rapidly decreasing.

However, in order to analyze various topics, the perplexity and consistency are satisfied, and the number of topics that can analyze the topic is selected.

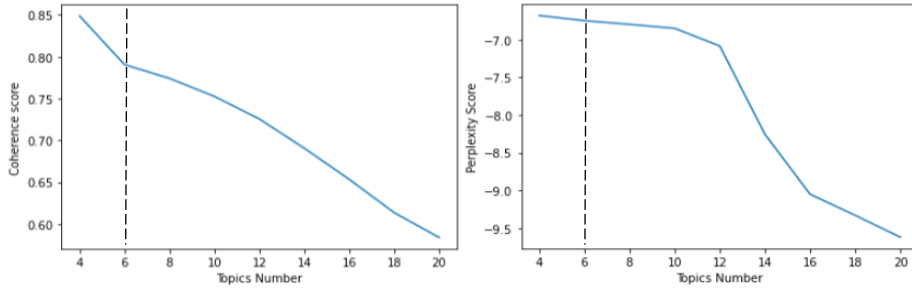


Fig. 9: Optimal number of negative topics

4. Test Result

4.1. Positive Topic Analysis

Based on LDA analysis, we obtained six positive topics. However, since there are many words in each topic, the key words were selected so that each topic could be intuitively analyzed. The selected key words are classified into related purchase, emotional delivery, content development, differentiation, and level of philosophical completeness. The related purchase consists of 'serially published, purchase, read-out, ebook, fun, again' which indicates that readers purchased serially published web novel contents as e-books. Emotional delivery consists of words that users feel good through web novel contents and consists of 'feeling, thought, expectation, woman, middle part, and finale.'

The content development consists of satisfaction felt by the user through the development of solid web novel contents, and consists of 'heart, person, mindset, man, charm, touching'. Differentiation is composed of satisfaction of users who feel different from existing web novel contents and includes 'as expected, different, slightly, novel, and highly recommended'. Philosophy consists of words that make users who read web novel contents think variously through immersion and includes 'content, a little, appearance, thought, attention, once'. Completeness expresses the user's opinion on the integrated part of web novel contents, and includes 'novel, part, major, main character, romance, plot'.

Table 4: The result of positive topic modeling

Topic	Name	Score	Topic	Name	Score
related purchase	serially published	0.067	differentiation	as expected	0.063
	purchase	0.051		different	0.034
	read-out	0.049		slightly	0.032
	e-book	0.047		a little	0.028
	fun	0.036		novel	0.027
	again	0.033		highly recommended	0.026
emotional delivery	feeling	0.092	philosophy	content	0.067
	thinking	0.054		a little	0.049
	expectation	0.049		appearance	0.042
	woman	0.043		thinking	0.028
	middle part	0.042		attention	0.027
	finale	0.023		once	0.023
content development	heart	0.046	completeness	novel	0.066
	person	0.045		part	0.037
	mindset	0.035		major	0.035
	man	0.030		main character	0.033
	charm	0.029		romance	0.032
	touching	0.026		plot	0.025

4.2. Negative Topic Analysis

Using LDA analysis, six negative topics were obtained. In addition, the key words were selected so that the users who gave negative evaluations of web novel contents can intuitively understand topics. The selected theme words are classified into expectation, content development, related purchase, immersion, characteristics, and completeness. To begin with, expectation consists of words negatively evaluated by users who purchased web novel contents with expectation, and includes 'parts, expectation, degree, first, understanding, and e-book'. Content development consists of the words left by a user who felt negative about the content development of web novel contents. It is composed of 'middle part, main character, story, charm, romance, setting.' The related purchase was based on the words from the users who purchased web contents based on other users' review but made negative evaluation unlike that of the reviewers. It consists of 'see, buy, review, book review, again, and purchase.' The immersion is an evaluation left by users who are disappointed in the immersion of web novel contents, and consists of 'work, rating, concentration, individual, this time, a bit.'

Characteristics consists of 'love, people, major, woman, man, plot', which are the words left by users who are disappointed with the characters in the web novel contents. Completeness is composed of data from users who do not feel fun about web novel

contents with a plain story, and consists of 'fun, a little, story, disappointment, this, and plain'.

Table 5: The result of negative topic modeling

Topic	Name	Score	Topic	Name	Score
expectation	part	0.037	immersion	work	0.077
	expectation	0.034		rating	0.056
	degree	0.030		concentration	0.046
	first	0.028		individual	0.043
	understanding	0.027		this time	0.032
	e-book	0.022		a bit	0.028
content development	middle part	0.048	characteristics	love	0.062
	main character	0.043		person	0.036
	story	0.037		major	0.031
	charm	0.035		woman	0.028
	romance	0.029		man	0.026
	setting	0.025		plot	0.020
related purchase	see	0.062	completeness	fun	0.071
	buy	0.055		a little	0.067
	review	0.039		story	0.034
	book review	0.032		disappointment	0.034
	again	0.025		this	0.030
	purchase	0.025		plain	0.024

5. Conclusion

The purpose of this study is to find the factors of quality improvement of web novel contents through sentiment analysis of web novel contents users. Therefore, this study proposed a system that can be used for planning of high-quality web novel contents based on accumulated user review data using LDA modeling method so as to overcome the limitations of existing human resources-based web novel content analysis. First, the user's sentiment was classified. In addition, according to the results of the classification, topic modelings were performed to derive representative topics that can express positive web novel contents and negative web novel contents. As a result of such classification, high-quality web novel contents that received positive evaluation were able to derive representative key words such as related purchase, emotional delivery, content development, differentiation, philosophy, and completeness. As for the negative web novel contents, key words such as expectation, content development, related purchase, immersion, character, and completeness were derived. The main quality factors of web novel contents were secured by using the above key words.

This study has made positive contributions to deriving academic and practical implications related to the web novel users emerging as new demand who utilize various devices in recent years by exploring quality factors that affect their satisfaction with use and intention to use continuously.

It is necessary to study the new Attention model that can elaborately extract the characteristic data of web novel contents review data in the future and to study the improvement of model completeness through securing data of various expression methods -emojis, new words, special character combinations, etc.- that occur continuously.

Acknowledgments

Following are results of a study on the “Leaders in Industry-university Cooperation +” Project, supported by the Ministry of Education and National Research Foundation of Korea.

References

- Abu Farha I. & Magdy, W. (2019). Mazajak: An online arabic sentiment analyser,” in Proc. 4th Arabic Natural Lang. Process. Workshop, Florence, Italy. 192-198.
- Alayba, A. M., Palade, V., England, M., & Iqbal, R. (2018). A combined CNN and LSTM model for arabic sentiment analysis, in machine learning and knowledge extraction. A. Holzinger, P. Kieseberg, A. M. Tjoa, E. Weippl, Eds. Cham, Switzerland: Springer. 179-191.
- Ali, W., Yang, Y., Qiu, X., Ke, Y., & Wang, Y. (2021). Aspect-level sentiment analysis based on bidirectional-GRU in SIoT. May 6, IEEEAccess 9, 69938-69950.
- Al Omari, M., Al-Hajj, M., Sabra, A., & Hammami, N. (2019). Hybrid CNNs-LSTM deep analyzer for arabic opinion mining. in Proc. 6th Int. Conf. Social Netw. Anal., Manage. Secur. (SNAMS), Granada, Spain, Oct. 364-368.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*. 3, 993-1022.
- Byeong-Cheol, L., & Doo-San, K. (2020). Deriving the determinants of hotel service quality using OTA reviews. *Journal of Hotel & Resort*. 19(4), 41-58.
- Chae-Kwan, Lim and Kyeng-Bong. “The effect of service quality of legal services on customer satisfaction. *International Journal of Smart Business and Technology*. 5(2), 7-14. <http://dx.doi.org/10.21742/IJSBT.2017.5.2.02>.

Dahou, A., Xiong, S., Zhou, J., Haddoud, M. H., & Duan, P. (2016). Word embeddings and convolutional neural network for arabic sentiment classification. in Proc. 26th Int. Conf. Comput. Linguistics Tech. Papers, Osaka, Japan, Dec., 2418-2427.

De Venuto, D. & Bellotti, F. "A CNN based personalized product recommendation model. *International Journal of Smart Business and Technology*. Global Vision Press. 8(1), 19-30.

El-Affendi, M. A., Alrajhi, K., & Hussain, A. (2021). A novel deep learning-based multilevel parallel attention neural (MPAN) model for multidomain arabic sentiment analysis. *IEEEAccess*. 9, 7508-7518.

Heikal, M., Torki, M., & El-Makky, N. (2018). Sentiment analysis of arabic tweets using deep learning. *Procedia Comput. Sci.*, 142, 114-122, DOI:10.1016/j.procs.2018.10.466.

Hu, M. & Liu, B. (2004). Mining and summarizing customer reviews. *KDD*, 168-177.

Hu, M. H., Sun, A., & Lim, E. P. (2007). Comments-oriented blog summarization by sentence extraction. *CIKM*. 901-904.

Jin, Z. & L. C. -K. (2020). A study on the influencing factors of customer satisfaction and continuous use intention in mobile payment service. *International Journal of Smart Business and Technology*, 8(2), 25-30.

Lee, W. C., Lee, H. A., & Lee, K. J. (2009). Automatic product feature extraction for efficient analysis of product reviews using term statistics. *Journal of Korean Institute of Information Scientists and Engineers*. 16B.6, 497-502.

Lee, Y. J., Ji, J. H., Woo, G., & Cho, H. G. (2009). Analysis and visualization for comment messages of internet posts. *Journal of The Korea Contents Association*. 9(7), 45-56.

Lim, C. -K. (2021). The effect of the quality of entrepreneurship education and students' participation on entrepreneurial competence and entrepreneurial intention. *Asia-pacific Journal of Convergent Research Interchange*. FuCoS, ISSN: 2508-9080 (Print); 2671-5325 (Online), 7(7), 37-49. <http://dx.doi.org/10.47116/apjcri.2021.07.04>

Omara, E., Mosa, M., & Ismail, N. (2018). Deep convolutional network for arabic sentiment analysis. in Proc. Int. Japan-Africa Conf. Electron., Commun. Computations (JAC-ECC), Alexandria, Egypt, Dec. 155-159.

Song, J. S. & Lee, S. W. (2011). Automatic construction of positive/negative feature predicate dictionary for polarity classification of product reviews. *Korean Institute of Information Scientists and Engineers*. 38(3), 157-168.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., & Gomez, A. N. (2017). Attention is all you need. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Yang, C., Wu, L., Tan, K., Yu, C., Zhou, Y., Tao, Y., & Song, Y. (2021). Online user review analysis for product evaluation and improvement. *Journal of Theoretical and Applied Electronic Commerce Research*. 16(5), 1598-1611.

Yoon, S. & Kim, M. (2020). Topic modeling on fine dust issues using LDA analysis. *Journal of Energy Engineering*. 29(2), 23-29.

Zhai, Z., Liu, B., Xu, H., & Jia, P. (2011). Clustering product features for opinion mining. ACM. 347-354.

Zhao, C. & Liu, J. (2015). A novel dynamic virtual machine consolidation algorithm based on correlation coefficient analysis and bin packing. *International Journal of Smart Business and Technology*. 3(1), 1-14.
<http://dx.doi.org/10.21742/IJSBT.2015.3.1.01>.