A study on real-time internet comment system through sentiment analysis and deep learning application

[1]Hae-Jong Joo, [2*]Ho-Bin Song

Abstract

This paper proposes a big data sentiment analysis method and deep learning implementation method to provide a webtoon comment analysis web page for convenient comment confirmation and feedback of webtoon writers for the development of the cartoon industry in the video animation field. In order to solve the difficulty of automatic analysis due to the nature of Internet comments and provide various sentiment analysis information, LSTM(Long Short-Term Memory) algorithm, ranking algorithm, and word2vec algorithm are applied in parallel, and actual popular works are used to verify the validity. If the analysis method of this paper is used, it is easy to expand to other domestic and overseas platforms, and it is expected that it can be used in various video animation content fields, not limited to the webtoon field

Keywords: *Sentiment Analysis, Deep Learning, AI, Webtoon, Comment Analysis, Video Animation, LSTM Algorithm*

I. Introduction

Recently, the production and consumption of various online video contents are exploding due to the spread of personal smart devices including smartphones and the establishment of fast communication networks. The spread of such infrastructure is a reality that is accelerating. The performance of personal smart devices is taking place as a replacement device for PCs and laptops, and the communication speed is rapidly evolving to 5G beyond LTE, and WiFi can be easily used anywhere in daily life. This phenomenon is common all over the world, and in terms of Korea, it is showing a faster development speed[1].

Due to the development of these infrastructure conditions, the online market is gradually increasing its share in many areas such as movies, books, music, and art, beyond offline. Among them, the cartoon industry in the field of video animation shows a more prominent phenomenon. According to a statistical survey by the Korea Creative Content Agency (KOCCA), the development of the offline comics industry is on the decline, while the online comics industry is showing a rapid growth of more than 30% in all fields (number of companies, employees, sales, etc.). Recently, the production of movies and dramas based on Webtoon (Web+Carton) is also actively being made. If this market is included, the growth rate of the online animation market is increasing more rapidly[2].

The growth rate of the online animation market also changes in the surrounding environment, but accurate communication is limited due to easy accessibility and easy production and consumption of video animation content. The most important communication channel is the comments of consumers about the works, and in the case of popular works, there are 3 to 4,000 cases per day, and it is practically unreasonable to check them all.

Therefore, in this paper, in order to make it easier to communicate opinions between writers and readers about webtoons, which are the field of manga, among online video animation contents, a vast amount of real-time comment data is automatically transmitted through big data sentiment analysis. We will discuss how to provide web services through analytic deep learning models.

[1]*Professor, University of Kangnam, Dept. of KNU Cham-Injae College (hjjoo@kangnam.ac.kr)*
[2*] *Corresponding Author Professor, University of Mokwon, Dept. of Electrical & Electric Engineering (songhb@mokwon.ac.kr)*

Received: Mar. 27, 2023, Revised: Apr. 16, 2023, Accepted: Apr. 17, 2023

II. Related Works

2-1. Webtoon Market and Internet Comments

The webtoon market, a representative online animation service, is growing rapidly, and this phenomenon has been steadily appearing in statistics for the last five years. As shown in Figure 1 and Figure 2, the number of webtoon writers and works is steadily increasing. According to the data of Korea Creative Content Agency for 2019, the cumulative number of authors exceeds 7,000, and the cumulative number of works also exceeds 10,000. The average number of new authors increased by 1,153 over five years, and the number of works increased by 1,579 on average[2].

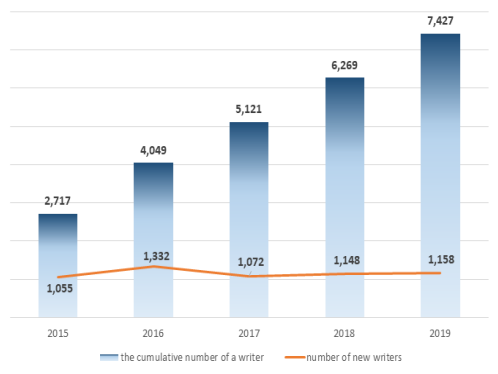


Figure 1. Author states by year

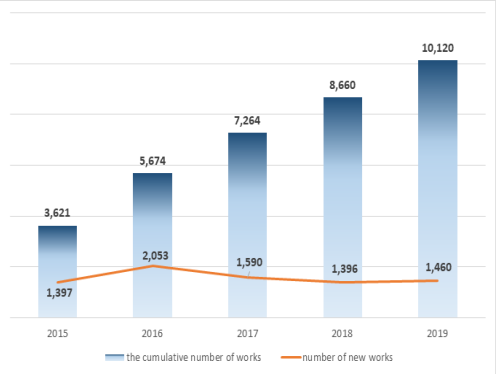
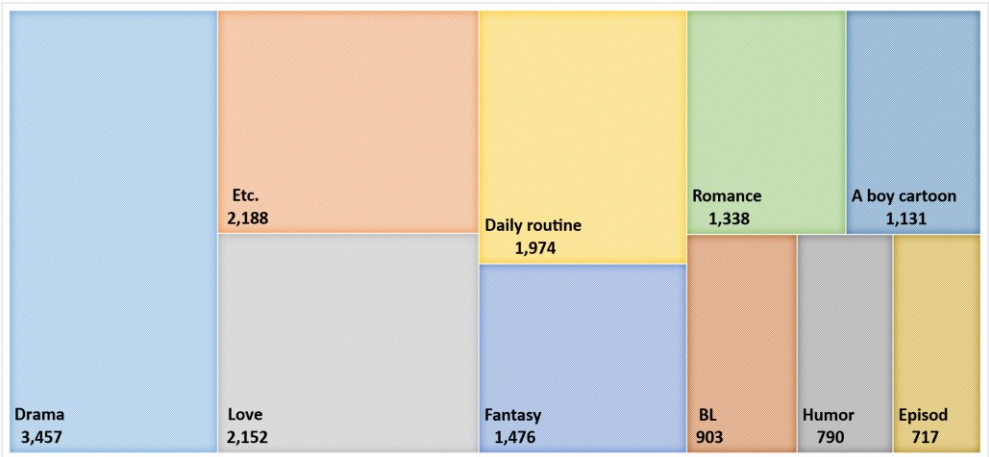


Figure 2. Works states by year



Drama	Etc.	Love	Daily routine	Fantasy	Romance	A boy cartoon	BL	Humor	Episod
21.4%	13.6%	13.3%	12.2%	9.2%	8.3%	7.0%	5.6%	4.9%	4.4%

Figure 3. Rank of works by genre

Figure 3 shows the total cumulative work of webtoons by genre. The most popular field is drama, which occupies more than 20%. Next, affection, daily life, and fantasy appear in order. Thanks to this popularity, large service companies are increasingly implementing webtoon services, and the share of large service platform operators is also increasing. As of 2020, as shown in Figure 4, "Naver Webtoon" accounts for more than 75%[2][3].

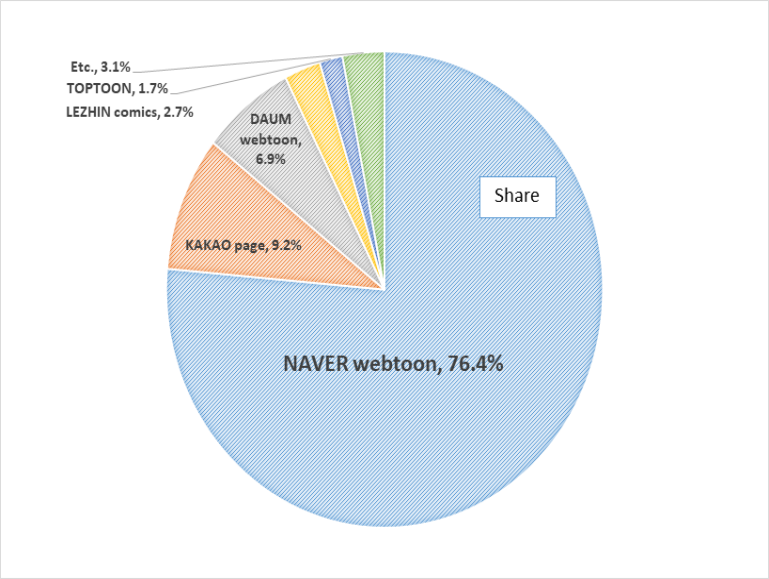


Figure 4. Service rank in webtoons

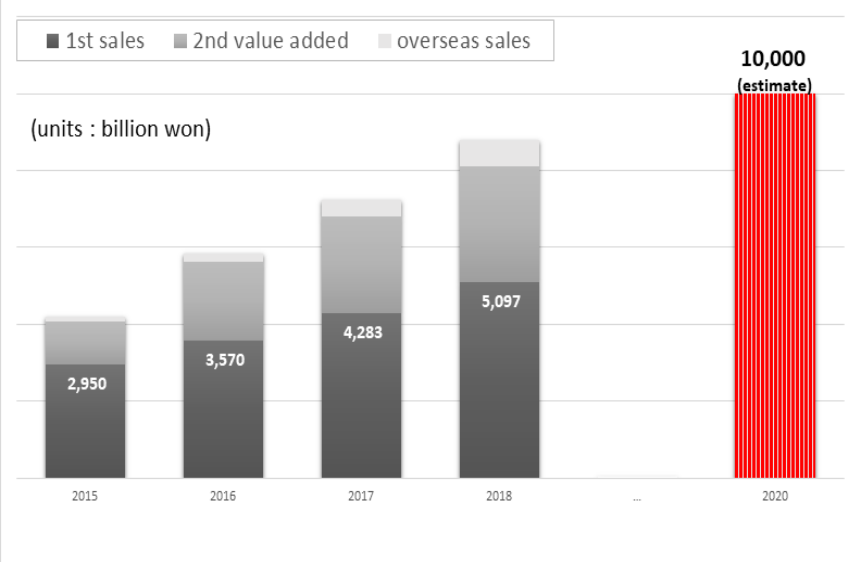


Figure 5. Webtoon market size

Looking at the sales status of the webtoon market(Figure 5), both primary sales, secondary value added, and overseas are increasing, and in 2020, total sales are expected to be over 10,000 billion won, plus secondary processing using webtoons(movies, games, drama, etc.), it will become a bigger market. In addition to being the author's pure creation, webtoon is a system that is frequently published every week, and communication with readers is one of the factors that can connect the popularity and sales of the work. Based on Naver, which occupies the largest share, an average of 3,000 to 4,000 comments per day are posted on popular works, and it is a reality that it is impossible for authors to analyze them individually. A questionnaire was conducted for writers and writers aspiring students, and through the analysis of the results, it can be seen that the identification of comments plays an important role in the work.

2-2. Deep learning methodology through big data analysis

Deep learning methodology is being used to solve various problems that have not been solved in various applications such as computer vision and natural language processing, and has excellent performance in solving most problems. The outstanding performance of deep learning has attracted many people's attention, and has established itself as an important concept indispensable for recent artificial intelligence research. However, the deep learning methodology has been pointed out in some aspects compared to its excellent performance, which causes difficulties in practical application of the deep learning model. Figure 6 compares the characteristics of current deep learning-based artificial intelligence technology with the characteristics of next-generation artificial intelligence technology that can be explained and predicted in the future. The current deep learning methodology has improved predictive performance better than the existing machine learning methodology, but it needs improvement in supervised learning-centered learning methodology, black box model with low explanatory power, and simple prediction at the present time[4][5].

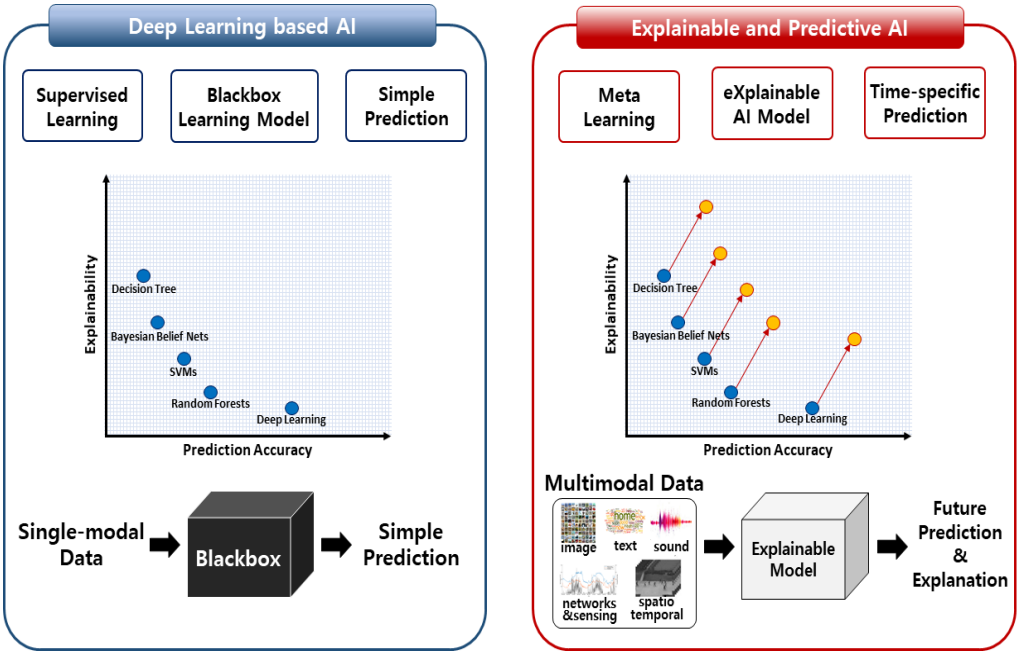


Figure 6. Comparison of the characteristics of current artificial intelligence technology (left) and the characteristics of next-generation artificial intelligence technology that can be explained and predicted in the future (right)

Except for some specially structured models or learning methodologies, general deep learning models perform learning in a supervised learning method. In supervised learning, learning data consisting entirely of data and labels must be constructed, but many data in the real world do not have labels attached, and it takes a lot of cost and time to build a quality dataset. Moreover, even if the training data set is difficult to construct, it is difficult to guarantee the performance of the model when a class imbalance problem of training data occurs, such as a very small amount of data of a specific class. In addition, the supervised learning-based deep learning model has a disadvantage that it cannot cope with concepts other than labels included in the training data at all[6]. For example, if a model is trained using training data to perform classification on 100 concepts in a specific domain, one of the existing 100 concepts is used for input data of new concepts not included in the actual 100 concepts. It has to be classified as a concept. Since the real world is composed of numerous concepts that are not included in the learning data, the characteristics of such supervised learning-based deep learning models are a great limitation to the application of the model.

Next, the deep learning model has the disadvantage of being a black box. The deep learning model is a kind of complex nonlinear function that is trained to output a specific value for input data, but does not

provide a basis for why the value was output[4][7]. A few years ago, Microsoft unveiled its artificial intelligence chatbot, which has been criticized a lot for its model being trained to speak racist. In other words, even for a deep learning model that shows excellent performance in many fields, the risk factors as in the previous example always exist because it is possible to determine whether it has worked properly only by checking the result.

The deep learning model is known as a universal approximation function, and based on this characteristic, it is applied to the industrial field including the existing approximation, classification, and prediction problems, showing excellent performance[4][8][9]. However, these problems are generally researched centering on the prediction of the current point of view, and by using only the existing deep learning methodology, the spatio-temporal image is unfolded in frames over time, away from tasks such as simple image classification. It is not easy to solve problems such as predicting future situations based on learning data composed of data.

III. Big data sentiment analysis for Internet comments

The process of 'Big Data Sentiment Analysis for Internet Comments' proposed in this paper is shown in Figure 7. As shown in Figure 7, it consists of the processes of data collection, data preprocessing, and data analysis, and detailed explanations are provided in each section.

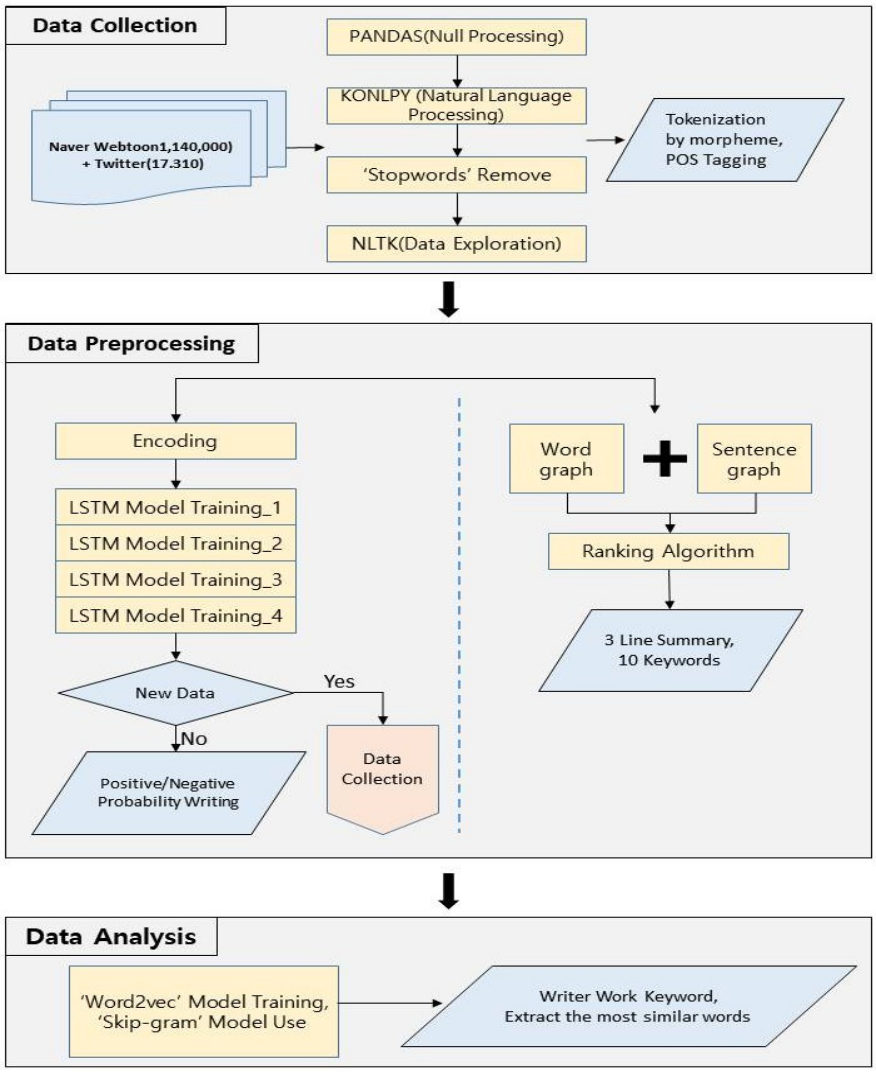


Figure 7. The process of sentiment analysis

3-1. Comment Collection

In order to analyze the comments of the webtoons that this paper intends to proceed with, the data shown in Table 1 were collected and used for analysis. Based on Naver works, comments from popular works and comments from unpopular works were collected on pure genre.

Table 1. Collected data information

	Naver Webtoon	Twitter
Collected information	date, episode information, star rating, comment, number of Likes, number of dislikes, number of comments	date, tweet, number of tweets
Collected data	-Random comment data: 130,000 cases -Comments on popular genre works: 850,000 cases -Pure Genre Unpopular Works Comments: 160,000 cases	-Tweets mentioning popular webtoons of pure genre: 16,466 cases -Pure genre unpopular webtoon mention tweet: 844 cases
File Format	CSV	CSV

Using the collected data, natural language processing and data search were performed in the following order prior to pre-processing.

① Using 'PANDAS', null processing was performed, and comments processed by Naver's bad comments cleanbot were removed.

② Natural language processing was conducted using 'KONLPY'. The sentences were tokenized in units of morphemes, part of speech tagging was performed, and stopwords were removed.

③'NLTK' was used to search the data, and the frequency graph of the token was verified

3-2. Data preprocessing

Figure 8 is an example showing some of the results of tokenization and category tagging. Each morpheme is tokenized, and you can see that it is tagged with category.

```
[['추국/Noun', '아/Josa', 'ㅠㅠㅠ/KoreanParticle'],
 ['백매/Noun', '약간/Noun', '무미랑/Noun', '같다/Adjective',
'.../Punctuation', '완전/Noun', '천재/Noun', '.../Punctuation'],
 ['도마뱀/Noun', 'ㅋㅋㅋ/KoreanParticle']]
```

Figure 8. The result of tokenization

Sentiment analysis was performed using the collected data using LSTM (Long Short-Term Memory) algorithm[10][11][12], one of the Recurrent Neural Networks(RNN). The LSTM is a structure with multiple layers and can efficiently overcome the disadvantages of RNN. LSTM model training was performed using one DENSE layer and Sigmoid function, and Epoch 4 training was performed. As a result, when new data(comment) is input, a pre-processing pipeline(tokenization, encoding, null processing) is performed, and a positive/negative

probability result is output. The threshold used in the analysis of this paper was applied as shown in Equation (1), and Figure 9 shows the results of some examples.

$$\begin{aligned} score > 0.7 &: \ positive \\ score > 0.5 &: \ medium \\ score \leq 0.5 &: \ negative \end{aligned} \qquad \text{(Equation 1)}$$

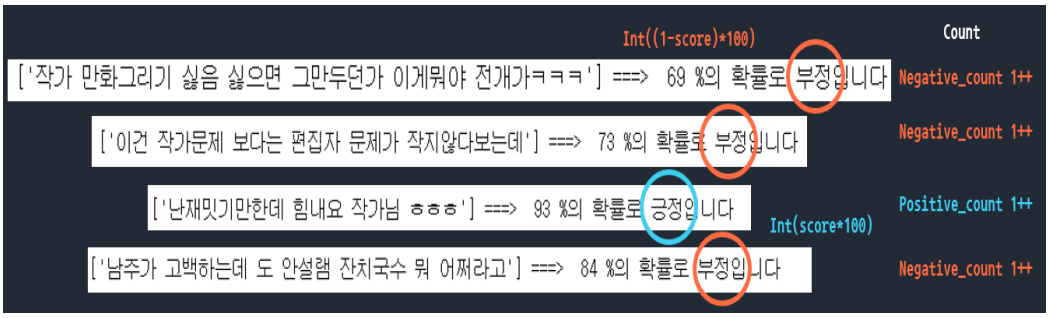


Figure 9. Example of sentiment analysis result

3-3. Big data sentiment analysis

Big data sentiment analysis performed two tasks for data analysis using preprocessed data. First, the comments were summarized using the textrank algorithm[5][13], and then the words related to the author and work keywords were extracted using word2vec[10][14]. In the comment summary, after using a word graph that calculates the similarity between words using a countvectorizer and a sentence graph that calculates the similarity between sentences using the TF-IDF model[6][12], a ranking algorithm that outputs sentences and words with high ranking using this graph is used. 3 lines of summary and 10 keywords were printed. Word2vec's skip-gram model was used to extract words most similar to the author's and work keywords in order to understand the relationship between words for the tokenized words. Figure 10 shows the result of such a word2vec model. As a result of visualizing the author as a keyword in tensorboard, it shows the top 15 words with the highest similarity.

Figure 9. Visualization result of word2vec model for author keyword (top 15)

IV. Result of analysis

In order to verify the validity of the big data sentiment analysis conducted in this paper, the analysis was conducted using “August Blizzard”, one of the most popular works in Naver, and the results are shown. Figure 11 shows the pipe chart as the result of the sentiment analysis for one episode. Positive, negative, and medium represent 1,601 times, 2,463 times, and 698 times, respectively, and when expressed as a ratio, negative accounts for 51%.

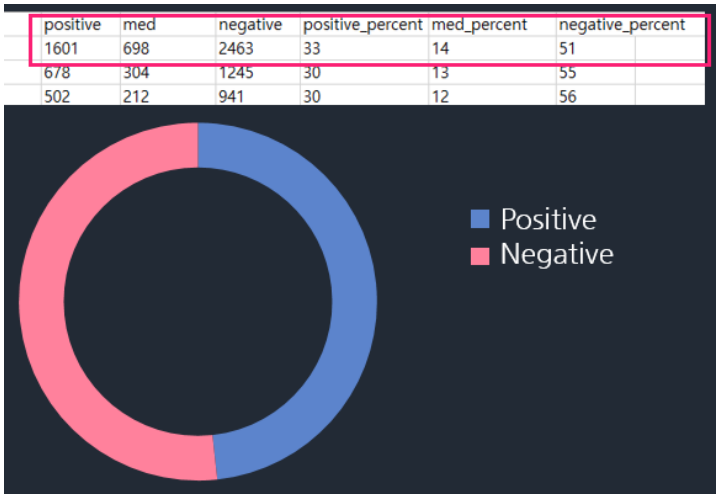


Figure 11. Emotional analysis result(“August blizzard”)

Figure 12 is a chart showing some of the results of keywords related to the author's keyword and the top 10 frequencies for the first to 42 episodes of the work, and the text visualization results for the words that appear.

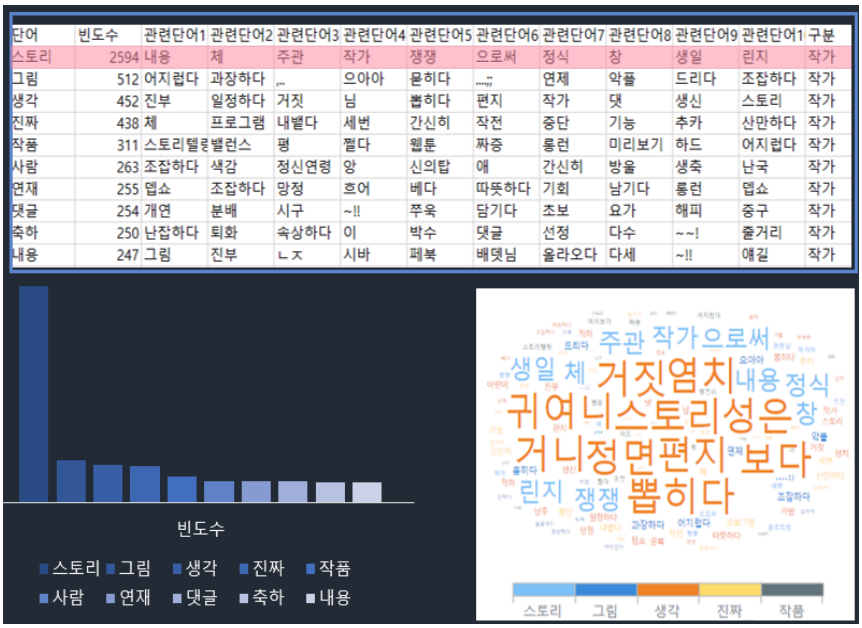


Figure 11. Visualization result for artist's keyword (for all works)

Figure 13 ~ Figure 16 show the final result of this paper, the dashboard, and it was implemented so that it can be transformed into a web or app form as needed. Figure 13 is a screen that provides a list of works provided by the platform in order to proceed with the analysis, and it is configured so that it is possible to search and select by episode or genre as needed. This screen can be replaced with a platform screen as needed.

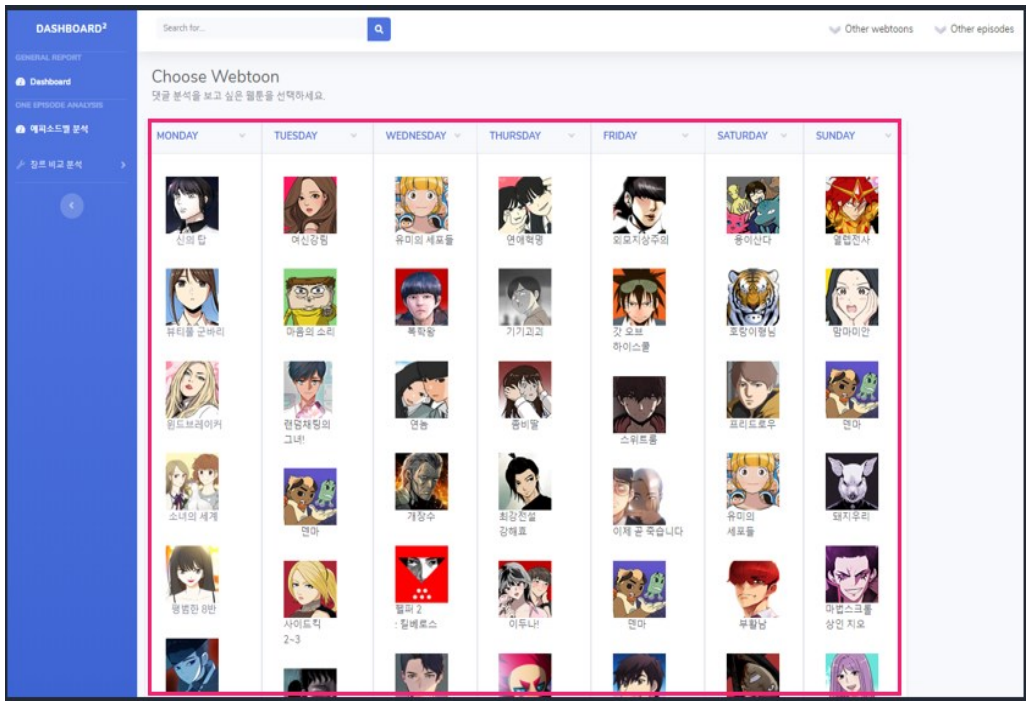


Figure 13. Result dashboard screen (work selection screen)

Figure 14. Results dashboard (top visualization of author and work mention comments)

Figure 14 is a screen that visualizes the author and top comments related to the work for the selected work, and is implemented so that it can be compared at a glance by composing it on one screen. The upper part of the screen is the composition of the artist, and the lower part is the composition of the work. In addition, the left is the reaction to each paid reader, and the right is the result of comments from people who have canceled each subscription.

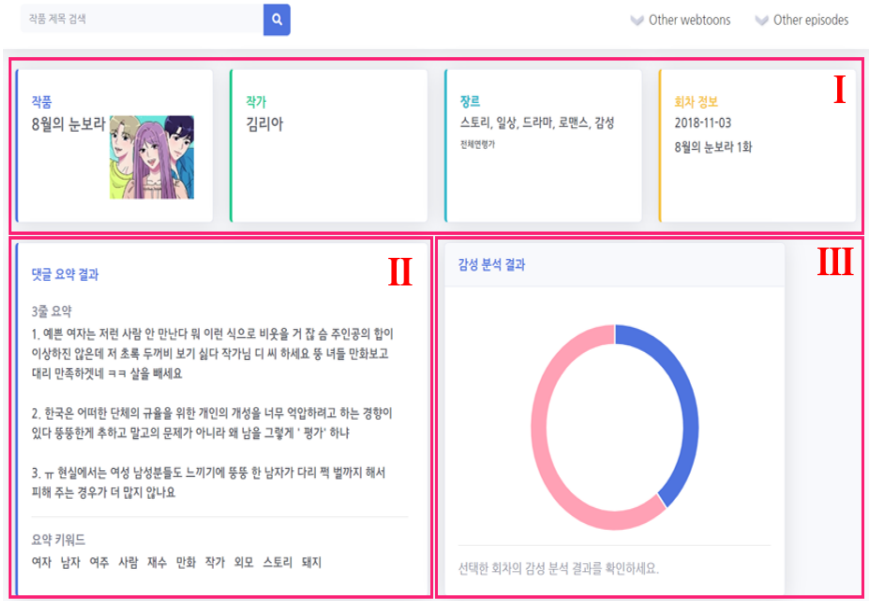


Figure 15. Result dashboard (work outline, comment summary, sentiment analysis)

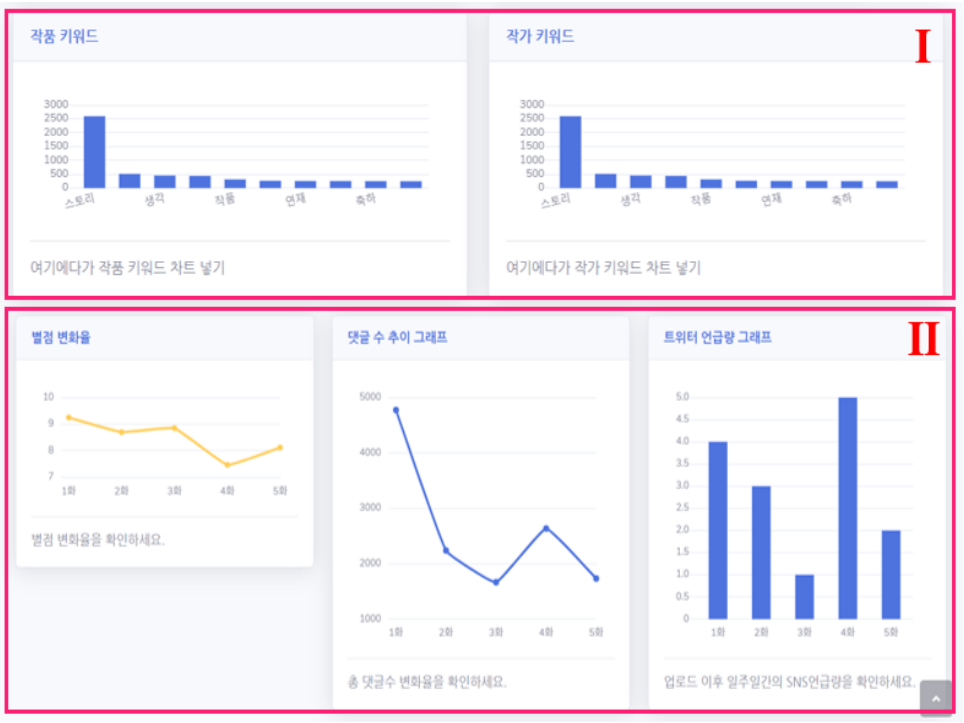


Figure 16. Result dashboard (Artist/Article Keywords, Star Rating/Comment Trend/Twitter)

Figure 15 is the second screen of the analysis result, showing the work outline, comment summary, and sentiment analysis. The outline of the work in 'Zone I' represents the general content of the analyzed work, and is displayed using data provided by the platform.

The summary of comments in ‘Section II’ shows a three-line summary and key keywords as a result of word2vec. 'Zone III' is the result of sentiment analysis using the LSTM algorithm.

Figure 16 is the third screen as a result of division, showing keywords and amount of change. 'Section I' shows the top 10 words related to each keyword for a work and an artist. 'Section II' represents the rate of change of star ratings, the amount of change in the number of comments, and the number of articles mentioned on Twitter for a specific period.

V. Conclusion

In this paper, a big data-based sentiment analysis method and implementation method are proposed to provide webtoon comment analysis web pages for convenient comment confirmation and feedback of webtoon writers. Big data analysis method was used to analyze a large amount of comments. In order to solve the difficulty of automatic analysis due to the nature of comments, LSTM algorithm, Ranking algorithm, and Word2vec algorithm are applied in parallel to provide various information, and to verify validity, actual popular works were implemented. It was possible to judge the possibility of using a big data analysis. If this analysis method is used, it is easy to expand to other domestic and overseas platforms, and it is expected that it can be used in various content fields, not limited to the webtoon field.

VI. References

[1] 1. Philipp A. Rauschnabel, Reto Felix, Chris Hinsch, Augmented reality marketing: How mobile ARapps can improve brands through inspiration, Journal of Retailing and Consumer Services, Volume 49, pp.43-53, 2019.
[2] Graphic Nevel Industry White Paper, Korea Creative Content Agency, 2017.
[3] Young-Kyu Kima and Min Ho Ryu, Towards Entrepreneurial Organization: From the case of Organizational Process Innovation in Naver, Procedia Computer Science 122, Information Technology and Quantitative Management (ITQM 2017), pp.663–670. 2017.
[4] Boemer F, Lao Y, Cammarota R, Wierzynski C (2019) nGraph-HE: a graph compiler for deep learning on Homomorphically encrypted data. ACM International Conference on Computing Frontiers 2019:1–27
[5] Nallapati, Ramesh, Zhou, Bowen, dos Santos, Cicero, Gulcehre, Caglar, Xiang, Bing. “Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond.” Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, Berlin, Germany, pp.280-290, Aug 2016.
[6] Barret Zoph and Quoc V. Le. Neural Architecture Search with Reinforcement Learning. arXiv e-prints, art. arXiv:1611.01578, November 2016.
[7] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, Polosukhin, Illia. “Attention Is All You Need” 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. Dec 2017.
[8] Kunkel, J., Loepp, B., & Ziegler, J.. “A 3D item space visualization for presenting and manipulating user preferences in collaborative filtering.” In Proceedings of the 22nd International Conference on Intelligent User Interfaces (pp. 3-15). ACM. March, 2017.
[9] Yann Lecun, Leon Bottou, Yoshua Bengio, Patrick Haffner, "GradientBased Learning Applied to Document Recognition," Proceedings of the IEEE 86.11, 1998.
[10] Mihalcea, Rada. “Graph-based ranking algorithms for sentence extraction, applied to text summarization.” In Proceedings of the 42nd Annual Meeting of the Association for Computational Lingusitics(ACL 2004) (companion volume), Barcelona, Spain. 2004.
[11] Yang You, “MIC-SVM: Designing a Highly Efficient Support Vector Machine for Advanced Modern Multi-core and Many-Core Architectures,”2014 IEEE 28th International, 2014.
[12] Dony, Robert D., and Simon Haykin. "Neural network approaches to image compression,"

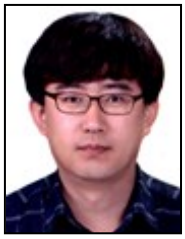
Proceedings of the IEEE 83.2, 1995.
[13] Quoc Le, Tomas Milokov, "Distributed Representations of Sentences and Documents," Proceedings of the 31st International Conference on Machine Learning, 2014.
[14] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, "Efficient Estimation of word Representations in Vector Space," arXiv:1301.3781, 2013.

Authors

Hae-Jong Joo

2008 : Ph.D of Computer Education, Cumberland University
2010 : Ph.D of Computer Engineering & Science, Myongji University

Research Interests : Data Science, Intelligence SW, Data Mining, Metaverse Platform, Video Big-Data QC



Ho-Bin Song

2006 : Ph.D Of Electrical Engineering, Myongji University

Research Interests : Big-data, AI, Power Electronics, Electric Vehicle