Journal of Korea Game Society 2021 Dec; 21(6): 75-86
http://dx.doi.org/10.7583/JKGS.2021.21.6.75

A Study on Brand Image Analysis of Gaming Business Corporation using KoBERT and Twitter Data

Hyunji Kim
Dankook University
hyunji93@gmail.com

ABSTRACT

Brand image refers to how customers, stakeholders and the market see and recognize the brand. A positive brand image leads to continuous purchases, but a negative brand image is directly linked to consumers′ buying behavior, such as stopping purchases, so from the corporate perspective, it needs to be quickly and accurately identified. Currently, methods of investigating brand images include surveys and SNS surveys, which have limited number of samples and are time-consuming and costly. Therefore, in this study, we are going to conduct an emotional analysis of text data on social media by utilizing the machine learning based KoBERT model, and then suggest how to use it for game corporate brand image analysis and verify its performance. The result has proved some degree of usability showing the same ranking within five brands when compared with the BRI Korea's brand reputation ranking.

요 약

브랜드 이미지는 고객, 이해관계자, 시장 전체가 해당 브랜드를 어떻게 보고 인지하는지를 뜻한다. 긍정적 브랜드 이미지는 계속적인 구매를 유발하지만, 부정적인 브랜드 이미지는 구매를 중단하게 만드는 등 소비자의 구매행동에 직결되기 때문에, 기업 입장에서는 빠르고 정확히 파악할 필요가 있다. 현재 브랜드 이미지를 조사하는 방법으로는 설문조사, SNS조사 등이 있는데, 표본의 수가 한정되고 시간과 비용이 많이 소요된다는 이슈가 있다. 따라서 본 연구에서는 딥러닝 기반의 KoBERT 모델을 활용하여 소셜미디어 상의 텍스트 데이터에 대한 감성분석을 실시한 후, 이를 브랜드 이미지 분석에 활용하는 방법을 제시하고, 이에 대한 성능을 검증하였다. 결과적으로, 다섯 개의 브랜드 이미지 순위를 매긴 결과가 한국기업평판연구소의 순위와 일치함으로써 본 연구의 사용성을 입증하였다.

Keywords : Brand Image(브랜드이미지), Machine Learning(기계학습), Social Media(소셜미디어), Transfer Learning(전이학습), Sentiment Analysis(감성분석), Natural Language Processing(자연어처리)

Received: Sep. 09. 2021 Revised: Nov. 16. 2021
Accepted: Dec. 06. 2021
Corresponding Author: Hyunji Kim(Dankook University)
E-mail. hyunji93@gmail.com

ISSN: 1598-4540 / eISSN: 2287-8211

1. Introduction

1.1 Purpose of the study

Korea's game industry is a key field in the software industry and digital content industry. According to data from the Korea Creative Content Agency, the total sales of the Korean game field grew 9% year-on-year in 2019, recording a total sale of 15.5 trillion won. The game field accounts for 12% of the total content industry (as of 2018) in terms of industry size, and is a strategic industry sector that accounts for 67% of total content exports in terms of export contribution.

In particular, as 33 game companies were listed on the stock market due to the rapid growth of the game industry (as of June 2020), the importance of the brand image of companies began to emerge. Among the listed companies, NCsoft and Netmarble have become super-large companies with a market capitalization of more than 10 trillion won, and a number of companies with a market capitalization of more than 1 trillion won (5 companies as of August 2020) have begun to emerge. The brand reputation of game companies can be seen as having a positive (+) correlation with the company's stock price [1], and for this reason, efforts to enhance the positive brand image using corporate social activities (CSR) are recognized as important.

On the other hand, the problem of the impact of the related corporate image on social negative perceptions caused by the controversy over the game's overindulgence and shutdown system cannot be ignored. In recent years, not only domestic regulatory issues but also international issues have emerged due to the WHO's push to grant disease codes, and game companies' brand image management issues are recognized as an essential element of corporate management strategies.

Therefore, game companies need a methodology to frequently grasp the degree of consumer response to their brand image (or brand reputation) and establish appropriate response strategies when positive or negative issues related to industries and individual companies arise.

Traditional methods used for brand image surveys include surveys, focus group interviews, SNS surveys, and observations. However, these methods has several limitations including limited data, high sample costs, data analysts' subjectivity and difficulty in multiple and relative evaluations with competitors.

As a way to solve the above problems, this study aims to estimate the brand image or brand attitude of users based on the results by collecting opinions related to the brand uploaded to SNS in real time and analyzing the brand through text mining using machine learning techniques so that companies' self-prediction and trend analysis of brand images can be performed.

1.2 Basic Concepts of Brand Image

Regarding the brand image, Dobni and Zinkhan stated that when customers think of a brand, they include symbolism such as instantaneous images that pop up in their mind when they think of a particular brand, feelings and familiarity with the brand's goods or services[2]. Keller said it is an integrated impression of consumers on the brand[3]. Zhang stated that consumers are a key driver of brand assets that influence consumer behavior as a general perception and feeling of a specific brand[4].

This brand image consists of functional and emotional elements, and especially, the emotional elements originate from the information processing process of attributes that constitute the functional measure of an individual's brand experience and image, so the brand image is the result of comparing and contrasting the attributes of various companies.

In other words, it can be seen that a good brand image increases customer satisfaction with products and services provided by a company, which positively affects the company's business performance, contributing to enhancing corporate stock prices and corporate value.

On the other hand, since brand image affects consumer attitude and consumer loyalty to brand[5], it can be predicted that negative brand image will negatively affect the company. In particular, due to the rapid development of wired and wireless Internet, these negative viruses occur online, and their spread, speed, and scope are incomparably faster and wider than before.

In particular, SNS-based communities are the main starting point for these positive or negative brand images to be directly uttered, shared, and spread, and are often reproduced and processed through this.

Therefore, it is important for companies to be able to detect and cope with these changes quickly.

2. Methodologies

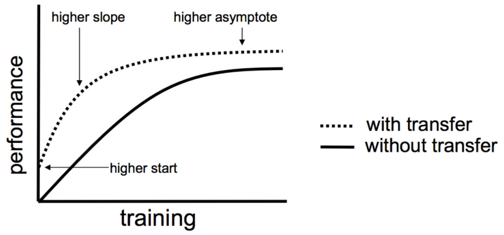
2.1 Machine Learning Approach

2.1.1 Transfer Learning

In the case of traditional machine learning models, a lot of time and resources are required since the number of parameters to be learned is very large and there are many matters to be considered when implementing the model, such as the type of activation function and the number of hidden layers. As an alternative to this, recently, transfer learning has been used in various fields.

Transfer learning is a machine learning method that reuses a model developed for work as a starting point for the model in the second task, and is a widely used approach in areas that require vast computing power and time, such as computer vision and natural language processing[6].

There are three advantages that can be earned when using this transfer learning[7]. Fisrt, The starting line is higher. The starting performance of the model is higher than that of not using transfer learning. Second, The slope is higher: When transfer learning is used, it shows a steeper rate of performance improvement. Lastly, asymptotes are higher: The fusion technique of the trained model is better than other methods.

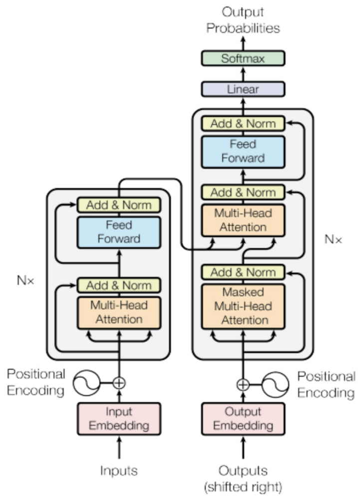


[Fig. 1] Strengths of Transfer Learning

2.1.2 Transformer

The transformer model has the characteristic of extracting features of each word using a self-attention mechanism to determine how important all other words in the sentence are in relation to the aforementioned words[8].

As can be seen in [Fig. 2], the transformer consists of an encoder and a decoder. The role of the encoder is to encode an input (i.e., a sentence) of a state including several tensors. The state is then transferred to the decoder to generate an output. In machine translation, the encoder converts a source sentence such as "Hello World" into a state (e.g., a vector) that captures semantic information. The decoder then uses this state to generate a translated target sentence (e.g., "Hello world"). Encoders and decoders have some submodules, but both mainly use multi-head attention and feed-forward networks. Unlike existing encoder-decoder models, transformers introduced the concept of self-attention without using convolutional neural networks and recurrent neural networks. BERT is a model that uses only encoders among transformers' encoder-decoders.



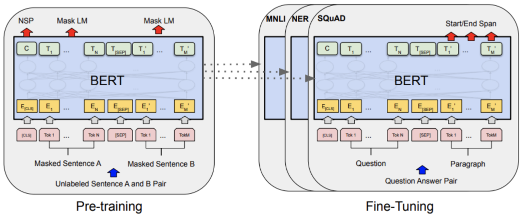
[Fig. 2] Structure of Transformer

2.1.3 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based language model built by Google. BERT showed excellent performance in a wide range of natural language processing areas, including question-answer (SQLUAD v1.1) and natural language inference (MNLI). Transformer has N encoding blocks consisting of multi-head attention and residual connection, and unlike reccurent neural networks commonly used in natural language processing, it has the advantage of being fast because it does not reccur. Instead of a one-way model that reads text inputs sequentially (left to right or right to left), the transformer encoder is considered a bidirectional model that reads the entire word sequence at once. This feature allows the model to learn the context of the word based on all its surroundings (left and right of the word)[9].

In the case of BERT, two tasks are used to pre-train the language model, the first is to randomly mask the words in the sentence and then predict the masked words, and the second is to determine whether the sentence B is appropriate as the next sentence in A. To solve the above tasks, BERT additionally uses token embedding, position embedding for token location information, and segment embedding to distinguish between A and B sentences, and finally collects the above three input embeddings to make them one embedding value [10]. In addition, the sum is used as an input by applying Layer Normalization and Dropout between neurons of the same layer.

BERT showed performance improvement in many natural language processing problems by conducting pre-training using the above two tasks and then fine-tuning the pre-trained model to process new tasks.



[Fig. 3] Pre-training and Fine-tuning of BERT

2.1.4 KoBERT

KoBERT[11] is an abbreviation of Korean BERT and is a BERT-based Korean language model. KoBERT is a model developed by SKT's T-Brain to overcome the limitations of existing BERT's Korean language performance, and learned with Korean sentences collected from Wikipedia and news. The KoBERT model used in this study is a model pre-trained with more than 5 million Korean Wikipedia sentence data and more than 20 million Korean news sentence data to learn Korean characteristics. The dictionary size of this BERT model is 8,002, and in order to tokenize Korean text, a Sentence Piece Tokenizer is separately learned and provided based on Korean Wikipedia and news text.

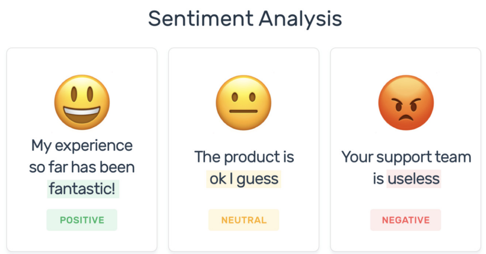
2.2 Text Mining

Text mining (also referred to as text analysis) is an artificial intelligence technology that uses natural language processing to convert unstructured text from documents and databases into normalized structured data suitable for analysis or to drive machine learning algorithms. Text mining can be used for various analysis purposes using text data, and in this study, text data on social media is collected and used to convert into structured data.

2.2.1 Sentiment Analysis

Sentiment analysis (or opinion mining) is a natural language processing technique used to interpret and classify emotions in subjective data. Sentiment analysis is often performed on text data to detect emotions such as email, survey responses, and social media data. Sentiment analysis is the process of detecting positive or negative emotions in text. Sentiment analysis refers to systematically identifying, extracting, quantifying, and researching emotional states and subjective information using natural language processing, text analysis, computer linguistics, and biometric recognition. Sentiment analysis is widely applied to data analysis on customer opinions such as reviews and survey responses, online and social media, and medical data. Since users express their thoughts and emotions more openly than ever on the Internet, Sentiment analysis is becoming an essential tool for grasping and understanding those emotions.

By automatically analyzing customer feedback, such as postings on social media, brands can learn what satisfies or frustrates customers, allowing them to customize their products and services to suit their needs.

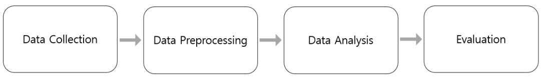


[Fig. 4] An Example of Sentiment Analysis

Figure 4[12] shows an example of sentiment analysis. In the figure above, the sentence "My experience so far has been fantastic!" was labeled as positive, the sentence "The product is okay I guess" was labeled neutral, and the sentence "Your support team is useless" was labeled as negative. As in the example above, sentiment analysis can be said to be determined by reading emotions from various text data such as customer feedback, reviews, and social media posting.

3. Experiment

3.1 Experimental setup



[Fig. 5] Experiment Flowchart

Figure 5 shows the schematic diagram of the overall flow for brand image analysis analysis. This study first acquired data by crawling searched tweets by setting the brand name as a search term on Twitter. After that, unnecessary information such as special characters was deleted, and Twitter text data was preprocessed by performing spelling and spacing tests. Data analysis consisted of pre-learning, transfer learning, and test stages using KoBERT. Data analysis consisted of pre-learning, transfer learning, and test stages using KoBERT. On KoBERT, a pre-trained model trained with Korean news data and Wikipedia data, transfer learning was conducted with NSMC (Naver Sentiment Movie Corpus) data, which is Naver movie review data, and sentiment analysis was performed with the trained model. Since the data was unlabeled data because Twitter crawling data was used as test data, performance evaluation was conducted by comparing it with the corporate reputation figures and brand rankings released by the Data Lab of the Korea Corporate Reputation Research Institute.

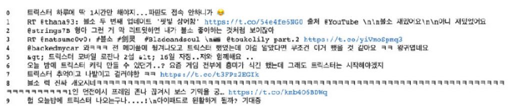
3.2 Data Collection

In this experiment, the brand image of five domestic game companies was analyzed. In order to compare and verify the analysis of the final positive evaluation ranking, the top five companies listed in the game-listed company brand reputation index data of the "Korea Corporate Reputation Research Institute Data Lab" in December 2020 were pre-set for analysis. Social media for data collection was limited to Twitter.

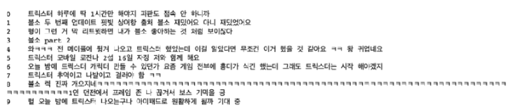
As the first step in the experiment, tweets related to five game companies (NCsoft, Netmarble, Com2us, Neptune, and Pearl Abyss) were collected from Twitter. Twitter API was used, and crawling was performed by setting the name of each game company as a search word. In the part of collecting data by executing crawling using the Twitter API, a streaming method was used to collect tweets updated in real time because there was a limit to the number of crawling permits for data already existing in Twitter policy. Data used in this experiment was collected from December 10 to December 15, 2020.

3.3 Data Preprocessing

In addition to the text data needed for the experiment, the collected data include dates, regions, and number of Likes. Since there are additional variables such as numbers, and raw data containing special characters such as exclamation marks, retweet IDs, emojis such as hearts, urls, and English, the process of deleting unnecessary information is necessary. After deleting unnecessary information, text preprocessing processes such as spacing and spelling tests were performed for easier text analysis.



[Fig. 7] Sample Text data before Preprocessing

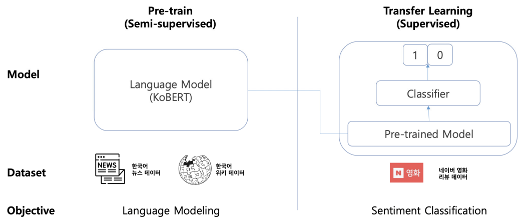


[Fig. 8] Sample Text data after Preprocessing

3.4 Data Analysis

3.4.1 Model Structure

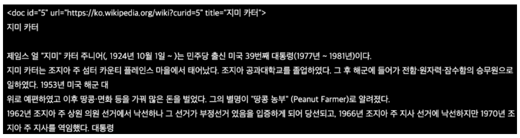
Natural language processing using BERT is largely divided into two stages: a language modeling process in which a giant encoder embeds input sentences and fine-tune them to perform several natural language processing problems. In this experiment, it was structured into a pre-trained model using Korean news data and Korean wiki data, and a model that performs Transfer Learning with Naver movie review data so that the task of classifying the sentiment of text can be performed.



[Fig. 6] Model Structure of this experiment

3.4.2 Model Training

In this experiment, a pre-learning model called 'KoBERT' was used. KoBERT is pre-trained with more than 5,000,000 Korean Wikipedia sentence data and more than 20,000,000 Korean news sentence data, and trained to learn the characteristics of "Korean" sentences.



<doc id="5" url="https://ko.wikipedia.org/wiki?curid=5" title="지미 카터">
지미 카터

제임스 얼 "지미" 카터 주니어(, 1924년 10월 1일 ~)는 민주당 출신 미국 39번째 대통령(1977년 ~ 1981년)이다.
지미 카터는 조지아 주 섬터 카운티 플레인스 마을에서 태어났다. 조지아 공과대학교를 졸업하였다. 그 후 해군에 들어가 전함·원자력·잠수함의 승무원으로 일하였다. 1953년 미국 해군 대
위로 예편하였고 이후 땅콩·면화 등을 가꿔 많은 돈을 벌었다. 그의 별명이 "땅콩 농부" (Peanut Farmer)로 알려졌다.
1962년 조지아 주 상원 의원 선거에서 낙선하나 그 선거가 부정선거 였음을 입증하게 되어 당선되고, 1966년 조지아 주 지사 선거에 낙선하지만 1970년 조지아 주 지사를 역임했다. 대통령

[Fig. 9] Example of Korean Wikipedia data

About 200,000 Naver movie review data were used for the transfer learning, which was built using movie ratings and one-line reviews uploaded on Naver. It was built by labeling 1-4 out of 10 movie stars as negative and 9-10 reviews as positive.

```
id      document        label
9976970 아 더빙.. 진짜 짜증나네요 목소리        0
3819312 흠...포스터보고 초딩영화줄....오버연기조차 가볍지 않구나        1
10265843        너무재밓었다그래서보는것을추천한다      0
9045019 교도소 이야기구먼 ..솔직히 재미는 없다..평점 조정       0
6483659 사이몬페그의 익살스런 연기가 돋보였던 영화!스파이더맨에서 늙어보이기만 했던 커스틴 던스트가 너무나도 이뻐보였다  1
5403919 막 걸음마 뗀 3세부터 초등학교 1학년생인 8살용영화.ㅋㅋㅋ...별반개도 아까움.     0
7797314 원작의 긴장감을 제대로 살려내지못했다.  0
9443947 별 반개도 아깝다 욕나온다 이응경 길용우 연기생활이몇년인지..정말 발로해도 그것보단 낫겟다 납치.감금만반복반복..이드라마는
7156791 액션이 없는데도 재미 있는 몇안되는 영화 1
```

[Fig. 10] Example of Naver movie corpus data

By training the classifier through transfer learning using Naver movie review data, it is possible to learn the characteristics that appear in daily conversations on the Internet, so transfer learning was conducted with the data. 150,000 data were used for training of the classifier and 50,000 data for testing, showing about 88 percent training accuracy. In this trained model, positive and negative classifications were conducted by adding Twitter data previously collected and preprocessed.

3.5 Evaluation

3.5.1 Result

The table below shows the ratio of positive and negative labels finally derived by brand.

[Table 1] Ratio of Pos/Neg Label by Brands

Brand \ Label	Positive	Negative
NCSoft	88.30%	11.69%
Netmarble	66.89%	33.10%
Com2us	59.24%	40.75%
Neptune	56.66%	43.33%
PearlAbyss	52.57%	47.42%

According to text data from social media collected, among the brand images of the five game companies to be analyzed, NCsoft had the highest positive percentage, and Pearl Abyss had the highest negative percentage and the lowest positive percentage. Based on the absolute gap in data, NCsoft's most positive image shows a large gap, while Com2us, Neptune, and Pearl Abyss do not show a large absolute gap between companies.

3.5.2 Comparative Verification

To confirm that the positive ratio by each brand shows similar trends with the real values, the result from this experiment was compared with the Communication Index of listed game companies from the Data Lab of the Korea Corporate Reputation Research Institute, which is related to positive and negative indices. For comparative verification between the two methodologies with different analysis criteria, the two values were normalized to derive relative value differences, and the rankings and values were compared through bar graphs.

[Table 2] Comparison of Results

Brand \ Source	BRI Korea DataLab	Result of the experiment
NCSoft	1.00	1.00
Netmarble	0.78	0.40
Com2us	0.51	0.19
Neptune	0.03	0.11
PearlAbyss	0.00	0.00

[Table 2] is a table that normalizes and compares each value. With normalization, min-max normalization was performed to normalize all values between 0 and 1.

As for the degree of positiveness of the brand indicated through normalization, as confirmed in Table 1, it can be reconfirmed that the results of the bottom three companies are close to each other, with NCsoft showing its distinctiveness. These results show some differences from the values derived from the Data Lab of the Reputation Research Institute, but there is no change in the ranking of brands ranked first to fifth.

4. Conclusion

Brand Image Analysis using machine learning techniques and social media data is actively performed in countries where the widely spoken language is English. However, the experiment is not actively performed in Korea and not many researches are done due to the linguistic limitations - the complexity of Korean language. Almost all of the researches including Brand Image Analysis still choose survey as their main research methodology in Korea. In this study, it was shown that it is possible to analyze the brand's image through data on social media, natural language processing techniques, and sentiment analysis techniques. Social media data can be collected at relatively little time and low cost compared to surveys, and users express more honest and transparent opinions because it is not a controlled environment. However, as it is a space where opinions can be freely expressed, more effort is needed in preprocessing to use such text data. In this experiment, text preprocessing was performed and used through unnecessary text removal, spelling, and spacing tests. In addition, in order to overcome the limitations of Korean language, which is difficult to train characteristics with small data, the pre-learned KoBERT model was used to learn only Korean characteristics, thereby reducing time and cost and improving performance. As a result, if there is collected data and a trained model, the test can be performed within a very short time. Also word clouds for each positive-negative label can be created to see keywords that affected the brand's positive-negative evaluation, and using them as feedback on service operation or management. In addition, when compared to the ranking of the "Data Lab of the Korea Corporate Reputation Research Institute," there was a slight difference in slope, but there was no change in the ranking, so it was possible to confirm some degree of usability.

In this experiment, emotions are classified in a binary way only as positive/negative, but future studies will attempt to perform transfer learning using the "single conversation dataset containing Korean emotion information" provided by the Korea Electronics Research Institute's Intelligence Information Flagship R&D, instead of NSMC (Naver Sentiment Movie Corpus). The dataset has seven emotions of joy, sadness, surprise, anger, fear, disgust, and neutrality as labels, and is a dataset built by collecting social media posts and online comments, so similar or better performance from Naver movie review data used in this experiment and more diversely observe opinions or feedback from consumers or users can be expected.

In addition, an advanced Korean Natural

Language Preprocessing model named KoGPT(Korean Generative Pre-trained Transformer) has recently been unveiled by KakaoBrain. KakaoBrain KoGPT is trained on rayn dataset, which is a dataset known to contain profanity, lewd, political changed, and other harsh language. This can cause trouble when the model is generating texts, however, it is expected to show better performance when the model classifies social media text data which contain countless informal terms.

REFERENCES

[1] Changseop Lee et al.. “A Study on the Regulation of the Disease Code of Game Addiction and the Classification of Guilty Likes by Game Companies.”. International JOURNAL OF CONTENTS, 19.10: 364-371, 2019.

[2] DOBNI, Dawn; ZINKHAN, George M. In search of brand image: A foundation analysis. ACR North American Advances, 1990.

[3] KELLER, Kevin Lane. Branding perspectives on social marketing. ACR North American Advances, 1998.

[4] ZHANG, Yi, et al. The impact of brand image on consumer behavior: A literature review. Open journal of business and management, 3.01: 58, 2015.

[5] RICHARD, Michael D.; ALLAWAY, Arthur W. Service quality attributes and choice behaviour. Journal of Services Marketing, 1993.

[6] Dipanjan (DJ) Sarkar “A Comprehensive Hands-on Guide to Transfer Learning with Real-World Applications in Deep Learning”, towardsdatascience, 2020.

[7] Lisa Torrey and Jude Shavlik, Transfer Learning. University of Wisconsin, 2009.

[8] VASWANI, Ashish, et al. Attention is all you need. In: Advances in neural information processing systems. p. 5998-6008. 2017.

[9] DEVLIN, Jacob, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[10] Joonyoung Choi, Heeseok Lim. “Establishment of an e-commerce data-based emotional analysis model using a natural language processing model.” Journal of Korea Convergence Society, 11.11: 33-39, 2020.

[11] SKT Brain "KoBERT," Github, 2018.

[12] Khristion Lambert “Harnessing the Power of Natural Language Processing to Create Better Patient Satisfaction”, Medium, 2020

[13] Takumi Kato, Kazuhiko Tsuda, “A Management Method of the Corporate Brand Image Based on Customers’ Perception”, Procedia Computer Science, Volume 126, Pages 1368-1377, 2018.

[14] Soyoung Kang, A Study on the Effects of Corporate Social Responsibility Assessment on Corporate Brand Image and Favorability : Focusing on the moderating effect of ordinary interest in CSR, JOURNAL OF THE KOREA CONTENTS ASSOCIATION, 21(7), 206-221, 2021.

[15] Jihoon Han, “The Relationship among Influencer Interpersonal Trust, Brand Image and Purchase Intention for SNS(Social NetworkService) Users”, JOURNAL OF THE KOREA CONTENTS ASSOCIATION, 20(1), 31-44, 2020.

[16] Junyoung Chung, Çaglar Gülçehre, Kyunghyun Cho, and Yoshua Bengio. “Empirical evaluation of gated recurrent neural networks on sequence modeling”. CoRR, abs/1412.3555, 2014.

[17] Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. “Sequence to sequence learning with neural networks”. In Advances in Neural Information Processing Systems, pages 3104 – 3112, 2014.

[18] T. Dietterich. Hierarchical reinforcement learning with the MAXQ value function decomposition. Journal of Artificial Intelligence Research, 13:227 – 303, 2000.

[19] L. C. Wnter, "The Effect of Brand Advertising on Company Image: Implications for Corporate Advertising," Journal of Advertising Research, Vol.26, No.2, pp.54-59,

1986.
[20] Hyuncheol Choi, Statistics in Social Sciences, Nanam, 2016.

김 현 지 (Kim, Hyunji)

약 력 : 2021 단국대학교 데이터사이언스학 석사

관심분야 : 데이터 분석, 기계학습, 미디어