Applied Artificial Intelligence
An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

A study on the evaluation of tokenizer performance in natural language processing

Sanghyun Choo & Wonjoon Kim

To cite this article: Sanghyun Choo & Wonjoon Kim (2023) A study on the evaluation of tokenizer performance in natural language processing, Applied Artificial Intelligence, 37:1, 2175112, DOI: 10.1080/08839514.2023.2175112

To link to this article: https://doi.org/10.1080/08839514.2023.2175112

Published online: 09 Feb 2023.

APPLIED ARTIFICIAL INTELLIGENCE
2023, VOL. 37, NO. 1, e2175112 (530 pages)
https://doi.org/10.1080/08839514.2023.2175112

A study on the evaluation of tokenizer performance in natural language processing

Sanghyun Choo [a] and Wonjoon Kim [b]

[a]Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC, USA; [b]Division of Future Convergence (HCI Science Major), Dongduk Women's University, Seoul, South Korea

ABSTRACT

The present study aims to compare and analyze the performance of two tokenizers, Mecab-Ko and SentencePiece, in the context of natural language processing for sentiment analysis. The study adopts a comparative approach, employing five algorithms - Naive Bayes (NB), k-Nearest Neighbor (kNN), Support Vector Machine (SVM), Artificial Neural Networks (ANN), and Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) - to evaluate the performance of each tokenizer. The performance was assessed based on four widely used metrics in the field, accuracy, precision, recall, and F1-score. The results indicated that SentencePiece performed better than Mecab-Ko. To ensure the validity of the results, paired t-tests were conducted on the evaluation outcomes. The study concludes that SentencePiece demonstrated superior classification performance, especially in the context of ANN and LSTM-RNN, when used to interpret customer sentiment based on Korean online reviews. Furthermore, SentencePiece can assign specific meanings to short words or jargon commonly used in product evaluations but not defined beforehand.

ARTICLE HISTORY

Received 17 June 2022
Revised 11 January 2023
Accepted 27 January 2023

Introduction

A user's favorable judgment of a product affects purchase behavior and brand loyalty to that product. Previously, product functional differences or low prices were the most critical factors in product purchase. Still, developers know that it is difficult to differentiate products in terms of performance, functional characteristics, and price as the technology level gap between companies narrows. For this reason, recent companies are focusing on improving appealing quality to maximize customer satisfaction in product and service development (Kim et al. 2018a; Park et al. 2019; Ryu, Son, and Kim 2020). Therefore, the focus is on product development to satisfy the affective needs of customers expressed in subjective and abstract ways, such as aesthetic experience and affective satisfaction. To quantitatively interpret the affective quality sought by customers–, affective analysis was introduced in product

CONTACT Wonjoon Kim, wjkim@dongduk.ac.kr Division of Future Convergence (HCI Science Major), Dongduk Women's University, Seoul, 02748, South Korea

development. The sentiment analysis methodologies have been applied to various products, such as automobiles, mobile phones, TVs, chairs, etc., because they can quantify customers' conceptual and implicit needs (Kim 2021; Son and Kim 2023).

Various studies related to affective analysis are proposed to understand customers' affective quality and apply it to product development (Henson et al. 2006; Park et al. 2019; Jain et al. 2022a). Especially, methodologies were developed to collect affective vocabulary to quantitatively measure the customer's emotion or impression of the product (Kim et al. 2018b; Lee et al. 2020). The most commonly used methods are literature research and interviews but collecting customer sentiment can cause problems. When performing affective evaluation for users, it is difficult to comprehensively grasp the results obtained from various age groups, genders, and ethnic groups, since evaluation experiments are generally performed on a small number of subjects (Moon et al. 2019). For this reason, it is difficult to generalize the model of affects derived based on the experimental results, and the possibility of bias due to the designer's intention cannot be excluded (Kim et al. 2018a).

To overcome this limitation, various studies on affective models' development through text mining techniques are being conducted (Jabreel et al. 2021; Kim et al. 2019b; Kitsios et al. 2021; Yang et al. 2020). Text mining refers to a technology that processes meaningful and valuable information in unstructured texts through natural language processing (NLP), and NLP is a computer understanding, analyzing, and interpreting the language used by humans in everyday life. It refers to a series of processes that make it possible (Collobert et al. 2011). Recently, various attempts have been made to understand customers' emotions through customer reviews on the web. Users can evaluate various experiences while using products and share opinions (Fang and Zhan 2015). e-commerce platforms such as Amazon utilize a system that allows only those who have purchased a product to give product reviews and ratings. Analyzing this can be used to improve the product by understanding users' real feelings about the product (Son and Kim 2023). In addition, as the population using various forms of social media on the web is increasing, and the age range is becoming more diverse (Auxier and Anderson 2021), web-based text mining has an advantage in deriving generalization of the results by dealing with a relatively large number of data compared to existing affective classification methods such as user surveys and expert interviews.

NLP generally goes through tokenization, cleaning, stemming, and lemmatization, among which tokenization is the process of dividing a given corpus into units called tokens. The morpheme analysis tokenizer was mainly used in the existing Korean-based online review for sentiment analysis (Lim and Kim 2014). However, online product reviews include many new words, such as abbreviations, slang/jargon, and emoticons (Vidal, Ares, and Jaeger 2018), and these buzzwords spread quickly and are used for a short period (Hashimoto et al. 2021). In general, out-of-vocabulary (OOV) issues, which are not defined in advance, frequently occur in sentiment analysis, translation, and document restoration and occur because input language cannot be processed because there is no dictionary or database. Therefore, various attempts are being made to solve this problem in the field of NLP (Arora and Kansal 2019; Kaewpitakkun, Shirai, and Mohd 2014; Pota et al. 2019).

Recently, Google proposed SentencePiece, a sub-word text tokenizer. It was possible to generate a vocabulary model including OOV by learning a sub-word model based on sentences entered in the text without depending on a specific language. A previous study on sentiment analysis on English, Japanese, Chinese, and French customer reviews improved accuracy using SentencePiece. Therefore, it is necessary to examine the applicability of SentencePiece to improve the accuracy of affective classification based on product reviews using Korean.

In a situation where competition among companies due to technological development is intensifying, it is emerging as one of the core values of corporate management to identify the various types of experiences that occur when customers use products or services and systematically manage affective quality.

The research goal of this study is to compare and evaluate the performance of tokenizers, which is one of the critical technologies for advancing customers' emotional classification through NLP. In particular, this study aims to identify a tokenizer that can reflect the characteristics of online product reviews that contain non-verbal expressions, abbreviations, slang, and neologisms. This study confirmed whether the performance of SentencePiece used in other languages, such as English and French, is the same in Korean. For evaluation, Mecab-Ko, a supervised learning-based tokenizer with proven effectiveness as a Korean tokenizer, was selected, and smartphone online product reviews were used as evaluation data. The authors would like to compare and evaluate the performance of a learning-based tokenizer and propose an algorithm that can optimize the classification performance for improving customer sentiment classification through NLP.

Background

Sentiment Analysis and Online Customer Reviews

Sentiment analysis grasps different levels of human emotions toward a product or service through NLP methods (Liu 2012; Farha and Magdy 2021). Due to the proliferation of online reviews, social media, and multimedia sharing platforms, vast amounts of text data in digital form have been accumulated on the web. Sentiment analysis has grown to be the most active research field in NLP. User opinions are the output of human activities using products and services and are a significant factor influencing the decision-making of others. In addition, its importance has spread to the omnidirectional domain (Balbi, Misuraca, and Scepi 2018).

Online customer reviews are defined as reviews of products posted on the website (Rose, Hair, and Clark 2011). The e-WOM (electronic word of mouse) expressed through online review provides product experiences such as product quality, price, and purchase reviews. The online retail market represented by Amazon allows customers to express their values and experiences freely, and these expressions can influence other people's choices (Gruen, Osmonbekov, and Czaplewski 2006). Therefore, online store managers strive to manage e-WOM effectively and efficiently (Litvin, Goldsmith, and Pan 2008).

In the product design area, interest in ways to meet customer needs in terms of affective experience in customers is increasing. Based on e-WOM, many studies have been conducted to identify the affective experiences of customers from various perspectives (Duarte, E Silva, and Ferreira 2018; Yoo, Sanders, and Moon 2013). These studies investigate the effects of e-WOM on customer's affective experience, preference, and ultimately purchase intention and behavior (Decker and Trusov 2010). Recently, studies have been mainly conducted to classify affective experience in online reviews using learning algorithms such as artificial neural networks (ANN) and support vector machines (SVM) (Jain et al. 2021, 2022b).

SentencePiece

SentencePiece, proposed by Kudo and Richardson (2018), is an unsupervised text tokenizer and decoder for a neural network-based text generation model. SentencePiece is implemented with byte-pair encoding (BPE), a two-sub-word classification algorithm, and a unigram language model by extending the learning concept directly from the original sentence. The basic principle of BPE is to compress the strings by merging the strings that appear most in the text corpus and repeat the process of merging and adding high-frequency strings repeatedly until the size of the vocabulary set reaches the desired level.

SentencePiece consists of four components: Normalizer, Trainer, Encoder, and Decoder. Normalizer is a step to standardize semantically equivalent Unicode characters in a normalized form, and Trainer is a step to learn a sub-word segmentation model in the normalized corpus. The encoder is a process of normalizing input text and tokenizing in sub-word order using the sub-word model trained by Trainer. Finally, the Decoder is a step of converting sub-word order into normalized text. SentencePiece has the advantage that it does not depend on the form or characteristics of the language as it performs word separation according to the frequency of appearance without prior knowledge of each language, such as morphemes.

Because of these advantages, previous studies using SentencePiece in sentiment analysis of web reviews have been reported (Polignano et al. 2019). Bérard et al. (2019) used the concept of SentencePiece to translate restaurant reviews into English-French. They proposed task-specific measurement indicators based on sentiment analysis or translation accuracy for each domain. Su, Yu, and Luo (2020) proposed XLNetCN, a new algorithm that complements pre-training-based models such as BERT and XLNet, mainly used in emotion analysis. They verified the superiority of the proposed model through restaurant and notebook review data. Bataa and Wu (2019) conducted a Japanese sentiment analysis based on transfer learning using SentencePiece. However, few studies have conducted sentiment analysis based on SentencePiece on product reviews written in Korean. Since the excellence of SentencePiece was investigated when analyzing emotions in various languages, it is necessary to conduct a study on Korean-based product reviews.

Method

Target Sample & Data Source

In this study, DANAWA (www.danawa.com), the largest price comparison site in Korea, crawled about 160,000 product reviews corresponding to the smartphone category. Among them, 153,257 product reviews were finally extracted by refining the contents of duplicates, product reviews in a form other than text such as emoticons, products with no or only stars, and those with less than two words.

In addition, in this study, we omitted the general pre-processing of NLP because we tried to find the possibility that various types of abbreviations, slang words, and inscriptions used on the web can more accurately classify customer emotions in reviews used.

To perform labeling to indicate positives and negatives using the product review data of the DANAWA site, the rating information evaluated for the smartphone purchased by the customer was used in the product review data. Ratings are rated and can be rated by 1–5 customers. In this study, the product review data corresponding to the ratings of 5 stars and 4 stars were considered positive. The review data corresponding to the ratings of 1, 2, and 3 stars were considered negative, and labeling was performed.

Tokenization

In this study, two types of tokenization methods were compared to compare and evaluate the performance of Korean tokenization: Mecab-Ko, a Korean morpheme analysis engine, and SentencePiece. Looking at comparative studies related to morpheme analysis, there was a performance difference for each POS tagger proposed for each language (Alluhaibi et al. 2021). Mecab-Ko is one of the morpheme analyzers widely used to tokenize Korean sentences (Moon et al. 2022). In the existing studies on POS tagger for Korean, it has been reported that the performance of Mecab based on the Sejong corpus is superior to that of other analyzers. Therefore, in this research, Mecab-Ko was selected as a supervised learning-based morpheme analyzer for comparative study. SentencePiece is a new tokenization method for neural network machine translation of unsupervised text based on a data-centric approach (Kudo and Richardson 2018).

Unlike the map-based approach, SentencePiece does not require pre-tokenized data sets, so it is more suitable for text corpora containing mixed language or words that are not in the dictionary. One of the critical differences between the SentencePiece algorithm and the general morpheme analysis engine is specifying the number of tokens in advance. According to Taniguchi, Konomi, and Goda (2019), selecting an appropriate number of tokens is essential because the number of tokens affects the classification accuracy of the SentencePiece algorithm. In the case of SentencePiece, the optimal k was set to 5000, 10000, 15000, 20000, 25000, and 30,000. In addition, Mecab-Ko and SentencePiece must use UTF-8 encoded text as input values. All product review texts collected in DANAWA used as target sites in this study were UTF-8 encoded, separate encoding Used without conversion.

Classification Algorithms and Performance Index

In this study, five types of supervised learning-based classification algorithms: Naive Bayes (NB), k-nearest neighbors (kNN), Support Vector Machine (SVM), artificial neural networks (ANN), and long short-term memory recurrent neural networks (LSTM-RNN) were used to evaluate the performance of sentiment classification in Korean products reviews. Especially, the LSTM-RNN algorithm was adopted in this study for comparison with the existing machine-learning method. Since the performance of LSTM-RNN has been proven in studies related to sentiment analysis in the existing NLP, LSTM-RNN was selected as a representative instead of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) in this study (Lin et al. 2020).

NB is a probabilistic classification algorithm based on Bayes' Theorem. Given a class, this algorithm assumes that all properties are independent of each other and is generally used widely to reduce parameters. The NB consists of Bayesian rules, conditional independence assumptions, and classification rules for input data. Compared to other complex graphic models, NB has the advantage that the number of data required to estimate parameters required for classification is small. kNN is an algorithm that calculates the distance between a training sample and a test sample in a data set and classifies it based on adjacent elements. It has the advantage of straightforward interpretation and short calculation time. SVM is mainly used for classification problems as a supervised learning model for pattern recognition and data analysis (Kim et al. 2018a). This algorithm aims to find a hyperplane that can maximize the margins in the feature space where the data is mapped.

ANN is a statistical learning algorithm modeled by simulating human brain neural networks (Zou, Han, and So 2008). In general, to achieve reliable performance, this model requires many interconnected neurons (Kim et al. 2019a). According to the signaling method, the neural network model is divided into a feed forward NN and a recurrent NN and is applied and utilized in research related to deep learning.

RNN was developed to model sequence data. RNN forms a cyclic structure in which hidden nodes are connected to edges with a certain direction in the neural network structure and is a structure that can receive input and output regardless of the size of sequence data (Lipton, Berkowitz, and Elkan 2015). To solve this problem, Hochreiter and Schmidhuber (1997) proposed the concept of a Long Short-Term Memory (LSTM) cell. In the structure of LSTM, a long-term cell in a neural network learns a part to remember, a part to delete, and a part of reading as it passes through each stage. In addition, the current input vector and the previous short-term state are injected into different fully connected layers.

To evaluate the classification performance of the model derived through machine learning, four performance indicators: Accuracy, Precision, Recall, and F1 were used (See Eq. (1)-(4)). These indicators were calculated according to four criteria: true positive (T_P), true negative (T_N), False positive (F_P), and False negative (F_N). Evaluating the learning model's performance can be understood as the relationship between the actual data values and the values derived from the model. Accuracy represents the proportion of exactly fit across the entire case. In other words, it is calculated as the rate of classifying reviews that were "positive" as "positive" in actual data and classifying reviews that were "negative" as "negative." The recall rate is calculated as the ratio of what was classified as "positive" by the learning model among "positive" in the actual data. Finally, precision is the proportion of what the learning model classifies as "positive" for the actual review to be "positive."

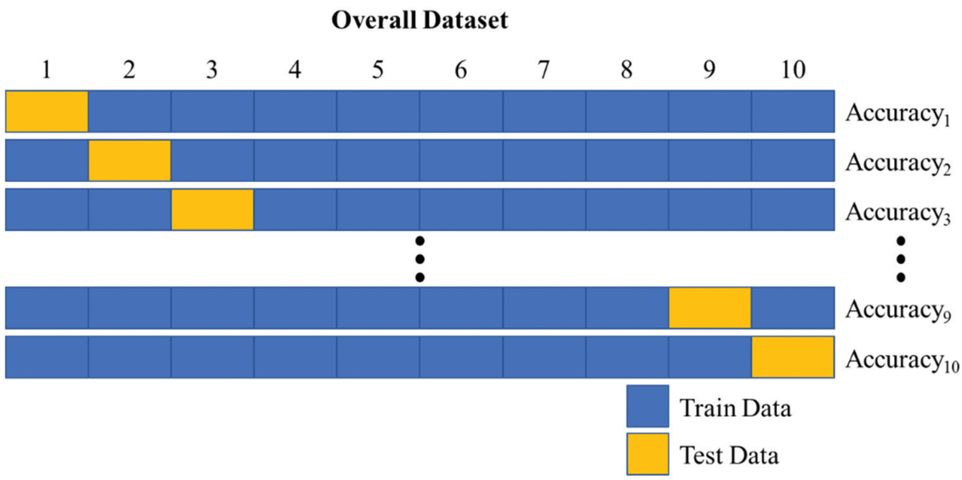

Figure 1. Illustrative of 10-fold cross-validation in this study.

$$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \tag{1}$$

$$Precision = \frac{T_p}{T_p + F_p} \tag{2}$$

$$Recall = \frac{T_p}{T_p + T_n} \tag{3}$$

$$F1 = 2\frac{precision \times recall}{precision + recall} \tag{4}$$

Verification of Performance Index

In this study, model verification was performed to verify the classification performance calculated through the learning model. Since a training model developed for a specific data set may not accurately classify other data sets, cross-validation of the proposed model is required. If the performance of the model is evaluated by performing cross-validation, since the entire data set is used for evaluation, there is an advantage of preventing overfitting that specific data is used for evaluation. In this paper, 10-fold cross-validation was used to verify the classification performance of each learning algorithm. 10-fold cross-validation divides the entire data set into ten subsets as shown in Figure 1 below, performs ten evaluations, calculates

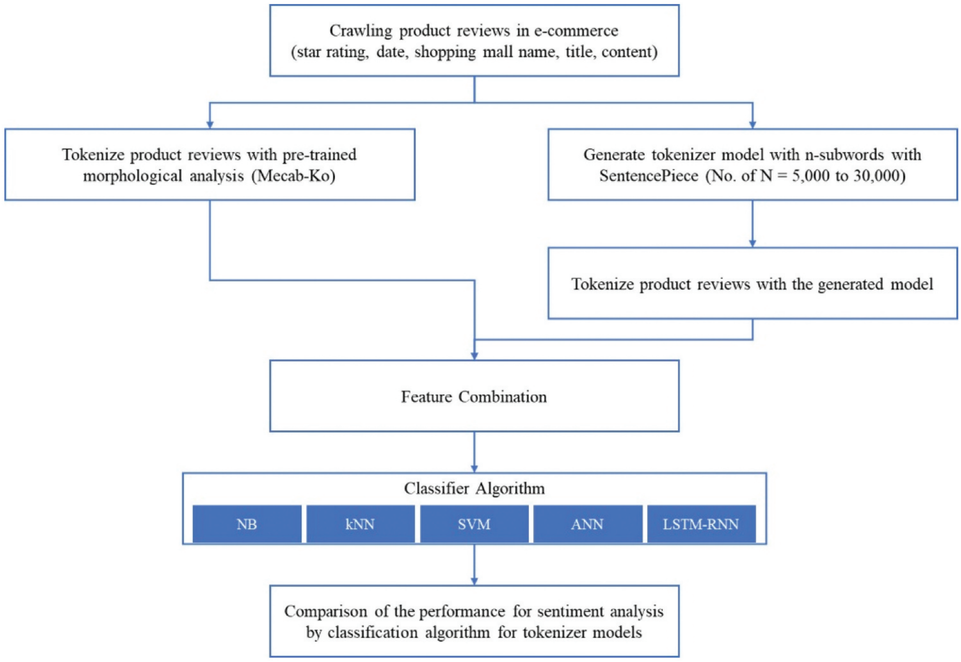

Figure 2. Illustrative of overall research flow.

the average value of the derived performance index, and evaluates the performance of the model.

The overall flow of this research is shown in Figure 2 below. First, product reviews and star ratings are crawled on Danawa (www.dananwa.com), the most extensive website providing product information in South Korea. Next, product reviews are tokenized with Mecab-Ko, a morpheme analyzer, and based on the number of tokens derived from this, k, the number of tokens of SentencePiece, is determined (5k to 30k in this study), and tokenization is performed. Finally, Mecab-Ko and SentencePiece were compared through four performance indicators (accuracy, F1-score, precision, and recall) for each of the five pre-selected classification algorithms.

Results

Comparative Analysis of Supervised/Unsupervised Learning Tokenizer by Classification Algorithm

The results of evaluating the performance of the positive and negative classification algorithm for customer sentiment analysis are shown in Tables 1–4 below.

Table 1. The results of accuracy for SentencePiece and Mecab-Ko each classification algorithm.

		No. of Tokens in SentencePiece					
Algorithm	Mecab-Ko	5000	10000	15000	20000	25000	30000
NB	0.880	0.903	0.900	0.907	0.906	0.921	0.914
kNN	0.884	0.875	0.904	0.912	0.922	0.933	0.924
SVM	0.948	0.950	0.951	0.948	0.948	0.948	0.949
ANN	0.948	0.947	0.960	0.956	0.948	0.948	0.943
LSTM-RNN	0.954	0.953	0.964	0.962	0.970	0.966	0.964

Table 2. The results of precision for SentencePiece and Mecab-Ko each classification algorithm.

		No. of Tokens in SentencePiece					
Algorithm	Mecab-Ko	5000	10000	15000	20000	25000	30000
NB	0.953	0.953	0.953	0.952	0.953	0.953	0.953
kNN	0.966	0.964	0.953	0.973	0.970	0.971	0.974
SVM	0.970	0.970	0.977	0.976	0.970	0.970	0.966
ANN	0.957	0.958	0.960	0.958	0.957	0.958	0.957
LSTM-RNN	0.968	0.955	0.967	0.966	0.973	0.971	0.967

Table 3. The results of recall for SentencePiece and Mecab-Ko each classification algorithm.

		No. of Tokens in SentencePiece					
Algorithm	Mecab-Ko	5000	10000	15000	20000	25000	30000
NB	0.919	0.945	0.942	0.949	0.949	0.965	0.973
kNN	0.912	0.903	0.946	0.933	0.948	0.949	0.956
SVM	0.976	0.974	0.981	0.979	0.976	0.975	0.975
ANN	0.990	0.991	0.990	0.989	0.990	0.989	0.991
LSTM-RNN	0.984	0.996	0.996	0.996	0.996	0.994	0.996

Table 4. The results of F1-score for SentencePiece and Mecab-Ko each classification algorithm.

		No. of Tokens in SentencePiece					
Algorithm	Mecab-Ko	5000	10000	15000	20000	25000	30000
NB	0.936	0.949	0.947	0.951	0.950	0.959	0.963
kNN	0.938	0.933	0.949	0.953	0.959	0.960	0.965
SVM	0.973	0.972	0.979	0.977	0.973	0.972	0.970
ANN	0.973	0.974	0.975	0.973	0.973	0.973	0.974
LSTM-RNN	0.976	0.975	0.981	0.981	0.984	0.982	0.981

The analysis reveals that with respect to accuracy and recall, the unsupervised learning-based tokenizer SentencePiece outperformed the supervised learning-based tokenizer Mecab-Ko when applied to the NB, for all the selected numbers of tokens. However, when the number of tokens was 5000, the kNN, ANN, and LSTM-RNN demonstrated lower accuracy than Mecab-Ko. The results from SentencePiece, when using the five algorithms, showed that accuracy was dependent on the number of tokens, with significant variations observed in the kNN algorithm. Additionally, LSTM-RNN exhibited the highest accuracy among all algorithms for all tokens, while NB had the lowest accuracy, except when the number of tokens was 5000. In the case of Mecab-Ko, LSTM-RNN still demonstrated the highest accuracy, while NB had the lowest.

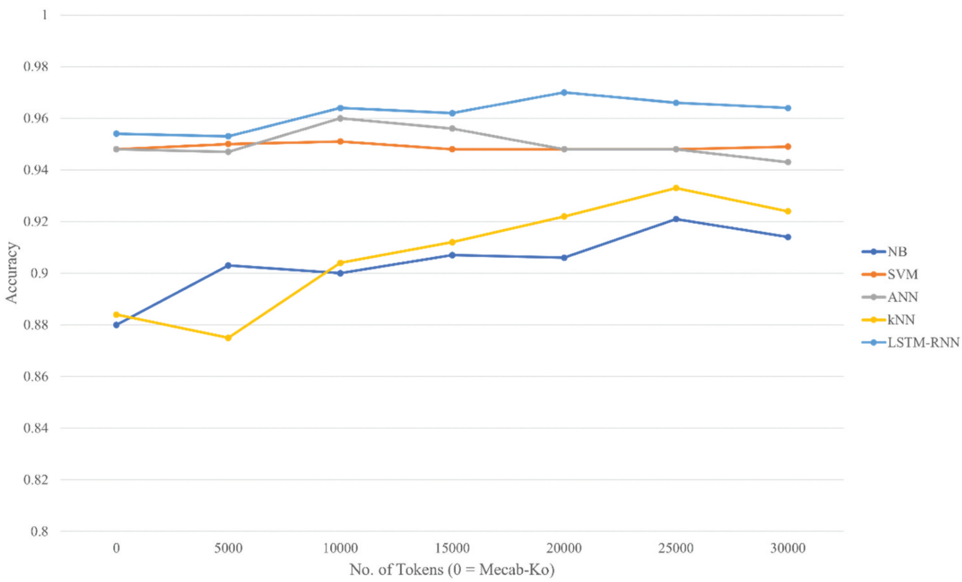

Figure 3. Result of the accuracy for token number of SentencePiece of each learning algorithm.

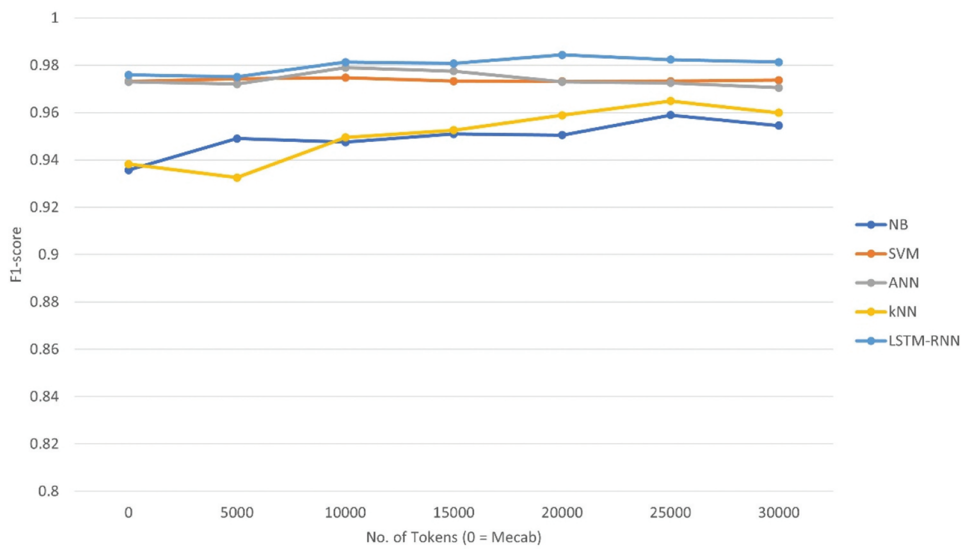

Figure 4. Result of the F1-score for token number of SentencePiece of each learning algorithm.

The result of comparing the accuracy & F1-score of each algorithm according to the number of tokens is shown in Figures 3–4 below.

Performance Verification of Classification Analysis by Tokenizer

In order to verify the statistical significance of the classification accuracy between morpheme analysis and SentencePiece, this study compared the SentencePiece results and morpheme analysis results for the specific number of tokens with the highest accuracy for each algorithm using the paired-sample t-test. In the verification, ANN and LSTM-RNN, which showed high performance among the five algorithms, were targeted. As mentioned in Section 4.1, when the number of tokens of SentencePiece is 20,000, the value of the classification performance index of the two algorithms was the highest. Furthermore, ANN and LSTM-RNN were selected for performance comparison in this study because they are derived algorithms based on neural networks. For the verification experiment, training and test data were randomly selected from the entire data set at a ratio of 90:10 and repeated 100 times for each algorithm.

Two algorithms for each tokenizer before performing the paired-sample t-test: the Kolmogorov-Smirnov test and the Shapiro-Wilk test to check the normality of the performance indicator data of ANN and LSTM-RNN. As a result of the test, the p-value in all data is greater than 0.05, so the data can be considered to have normality. As a result of the paired sample t-test, it was found that there were statistically significant differences ($p\text{-value} < 0.05$) in all three performance indicators in two algorithms: ANN and LSTM-RNN.

Discussion

The purpose of this study is to provide clear information on affective quality to companies or related stakeholders by proposing a method for upgrading the classification of affective experiences in customers expressed and exchanged on the web. It is expected to help companies make decisions by accurately identifying information and feedback on the quality of products and services provided to customers. This study compared the performance of SentencePiece, a sub-word tokenizer designed for neural network-based text processing, and Mecab-Ko, a morpheme-based tokenizer, when classifying customer emotions based on product review data from a price comparison site. This study confirmed that SentencePiece, an unsupervised learning tokenizer, is suitable for affective analysis studies using Korean product review data. SentencePiece can be applied to an end-to-end system for sentiment analysis based on review data because it can tokenize sub-words and convert the text of sentences into a time series of token IDs.

In this study, the representative affective factors of customers expressed in online reviews were selected as "positive" and "negative," and a total of four products, which a product that people use a lot in daily life and can receive enough affective experience, were selected as the target product. Three algorithms were adopted to confirm the classification performance, and the tokenizer's classification performance was examined in terms of accuracy, precision, recall, and F1 score.

According to van der Heijden, Abnar, and Shutova (2020), it was confirmed that the Bert model using the word piece embedding method showed optimized performance in various languages such as English, Dutch, Spanish, etc. In this study, it was also confirmed that the model using the SentencePiece tokenizer showed better performance than the model using the morpheme analyzer.

From the perspective of research for sentiment analysis, it was confirmed that ANN showed superior performance compared to other algorithms among machine learning models (Kalarani and Selva Brunda 2019). There was almost no performance difference between ANN and SVM in Korean customer review data. Since SVM shows a high performance when the number of classes is small (Tian et al. 2017), it seems necessary to expand the results of this study in the future to compare the performance in various Korean review datasets.

In the research for sentiment classification using customer reviews, it has been confirmed through many studies that LSTM-RNNs guarantee high performance (Monika, Deivalakshmi, and Janet 2019; Singh et al. 2022). It has been reported that LSTM-RNN shows excellent performance compared to other algorithms in sentiment analysis for Korean (Eom, Yun, and Byeon 2022; Kim and Song 2022), and the same results were obtained in this study. Therefore, LSTM-RNN can be regarded as a suitable classification algorithm for Korean-based sentiment analysis. Recently, methods that improve RNN or combine CNN and RNN, such as Bidirectional RNN and Bi-LSTM, are being used in sentiment analysis (Basiri et al. 2021; Colón-Ruiz and Segura-Bedmar 2020). Therefore, it is expected that in the future Korean sentiment analysis research, it will be possible to research to improve classification performance by improving the algorithm based on RNN.

According to Taniguchi, Konomi, and Goda (2019), the SentencePiece algorithm, unlike other morpheme classifiers, requires specifying the number of tokens, which significantly impacts classification results. In fact, in their study, it was confirmed that there is a difference in predictive performance when the number of tokens is 1500, 3500, and 5500. It was confirmed that similar results were obtained in this study. All algorithms showed differences in performance index according to k, and in general, performance improved as the number of k increased, but performance decreased after a certain level. This is because overfitting occurs when the number of tokens is large compared to the size of the entire document. Therefore, through this study, it can be seen that it is necessary to perform the task of finding an appropriate k when performing sentiment analysis based on product reviews written in Korean.

Conclusion

This study aims to develop a tokenizer and classification algorithm to evaluate sentiment on Korean product reviews effectively. Customer opinions for sentiment analysis were collected through the comparison shopping website in Korea. For sentiment analysis using NLP, two types of tokenizers were compared and evaluated with five classification algorithms.

It was considering why the SentencePiece-based tokenizer showed superior performance compared to the morpheme analysis-based tokenizer in the problem of sensibility classification in online product reviews. In the case of online product reviews, typos, inscriptions, and nonstandard words are compared to news articles or patent documents. Because there are many, the result of morpheme analysis, such as Mecab-Ko, seems to be lower. In addition, SentencePiece, which combines and learns mode expressions based on a text corpus of more than a specific volume, groups meaningful expressions based on the same syllable sequence that people mainly use in product reviews for particular products.

Therefore, it seems that through the SentencePiece tokenizer, it is possible to give specific meanings to terms related to abbreviations or slang words, which are frequently used by users above a certain level in evaluating products but are not defined in advance. In addition, in the present era, where customer expectations and requirements for products and services are increasingly complex and there are many alternatives, it has become an essential element of corporate management for companies to respond to customer needs immediately. By expanding the results obtained through this study, it is expected to be used to classify positive and negative emotions about products and services and classify various factors that can subdivide affective quality.

The limitations of this study and future research tasks are as follows. First, this study utilized review data for smartphone to confirm the performance of customer sentiment classification. However, since the sensibility of each product is often different in existing studies, it is necessary to conduct research from a comprehensive perspective on various products in the future. Second, there is a need to try to improve the classification accuracy of the tokenizer by additionally utilizing deep learning-based performance evaluation algorithms such as CNN, MobileNet, etc. that have not been used in the study.

Acknowledgements

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT: Ministry of Science and ICT) (No. 2020R1G1A1003384).

Disclosure Statement

No potential conflict of interest was reported by the authors.

Funding

The work was supported by the National Research Foundation of Korea [2020R1G1A1003384]

ORCID

Sanghyun Choo http://orcid.org/0000-0002-8884-3437
Wonjoon Kim http://orcid.org/0000-0001-5177-8072

References

Alluhaibi, R., T. Alfraidi, M. A. Abdeen, and A. Yatimi. 2021. A comparative study of Arabic part of speech taggers using literary text samples from Saudi novels. *Information* 12 (12):523. doi:10.3390/info12120523.

Arora, M., and V. Kansal. 2019. Character level embedding with deep convolutional neural network for text normalization of unstructured data for twitter sentiment analysis. *Social Network Analysis and Mining* 9 (1):1–14. doi:10.1007/s13278-019-0557-y.

Auxier, B., and M. Anderson. 2021. Social media use in 2021. *Pew Research Center* 1:1–4.

Balbi, S., M. Misuraca, and G. Scepi. 2018. Combining different evaluation systems on social media for measuring user satisfaction. *Information Processing & Management* 54 (4):674–85. doi:10.1016/j.ipm.2018.04.009.

Basiri, M. E., S. Nemati, M. Abdar, E. Cambria, and U. R. Acharya. 2021. ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis. *Future Generation Computer Systems* 115:279–94. doi:10.1016/j.future.2020.08.005.

Bataa, E., and J. Wu. 2019. An investigation of transfer learning-based sentiment analysis in japanese. *arXiv preprint arXiv:1905.09642*.

Bérard, A., I. Calapodescu, M. Dymetman, C. Roux, J. L. Meunier, and V. Nikoulina. 2019. Machine translation of restaurant reviews: New corpus for domain adaptation and robustness. *arXiv preprint arXiv:1910.14589*.

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12:2493–537.

Colón-Ruiz, C., and I. Segura-Bedmar. 2020. Comparing deep learning architectures for sentiment analysis on drug reviews. *Journal of Biomedical Informatics* 110:103539. doi:10.1016/j.jbi.2020.103539.

Decker, R., and M. Trusov. 2010. Estimating aggregate consumer preferences from online product reviews. *International Journal of Research in Marketing* 27 (4):293–307. doi:10.1016/j.ijresmar.2010.09.001.

Duarte, P., S. C. E Silva, and M. B. Ferreira. 2018. How convenient is it? Delivering online shopping convenience to enhance customer satisfaction and encourage e-WOM. *Journal of Retailing and Consumer Services* 44:161–69. doi:10.1016/j.jretconser.2018.06.007.

Eom, G., S. Yun, and H. Byeon. 2022. Predicting the sentiment of South Korean twitter users toward vaccination after the emergence of COVID-19 Omicron variant using deep learning-based natural language processing. *Frontiers in Medicine* 9. doi:10.3389/fmed.2022.948917.

Fang, X., and J. Zhan. 2015. Sentiment analysis using product review data. *Journal of Big Data* 2 (1):5. doi:10.1186/s40537-015-0015-2.

Farha, I. A., and W. Magdy. 2021. A comparative study of effective approaches for Arabic sentiment analysis. *Information Processing & Management* 58 (2):102438. doi:10.1016/j.ipm.2020.102438.

Gruen, T. W., T. Osmonbekov, and A. J. Czaplewski. 2006. eWOM: The impact of customer-to-customer online know-how exchange on customer value and loyalty. *Journal of Business Research* 59 (4):449–56. doi:10.1016/j.jbusres.2005.10.004.

Hashimoto, J., A. Mutoh, K. Moriyama, A. Yokogoshi, E. Yoshida, T. Matsui, and N. Inuzuka (2021, October). Classification of buzzwords by focusing on time trends using twitter data. In 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE), Kyoto, Japan, (pp. 342–44). IEEE.

Henson, B., C. Barnes, R. Livesey, T. Childs, and K. Ewart. 2006. Affective consumer requirements: A case study of moisturizer packaging. *Concurrent Engineering* 14 (3):187–96. doi:10.1177/1063293X06068358.

Hochreiter, S., and J. Schmidhuber. 1997. Long short-term memory. *Neural computation* 9 (8):1735–80. doi:10.1162/neco.1997.9.8.1735.

Jabreel, M., N. Maaroof, A. Valls, and A. Moreno. 2021. Introducing sentiment analysis of textual reviews in a multi-criteria decision aid system. *Applied Sciences* 11 (1):216. doi:10.3390/app11010216.

Jain, P. K., W. Quamer, V. Saravanan, and R. Pamula. 2022a. Employing BERT-DCNN with sentic knowledge base for social media sentiment analysis. *Journal of Ambient Intelligence and Humanized Computing* 1–13. doi:10.1007/s12652-022-03698-z.

Jain, P. K., G. Srivastava, J. C. W. Lin, and R. Pamula. 2022b. Unscrambling Customer Recommendations: A Novel LSTM Ensemble Approach in Airline Recommendation Prediction Using Online Reviews. *IEEE Transactions on Computational Social Systems* 9 (6):1777–84. doi:10.1109/TCSS.2022.3200890.

Jain, P. K., E. A. Yekun, R. Pamula, and G. Srivastava. 2021. Consumer recommendation prediction in online reviews using Cuckoo optimized machine learning models. *Computers & Electrical Engineering* 95:107397. doi:10.1016/j.compeleceng.2021.107397.

Kaewpitakkun, Y., K. Shirai, and M. Mohd (2014, December). Sentiment lexicon interpolation and polarity estimation of objective and out-of-vocabulary words to improve sentiment classification on microblogging. In Proceedings of the 28th Pacific Asia conference on language, information and computing, Phuket,Thailand, (pp. 204–13).

Kalarani, P., and S. Selva Brunda. 2019. Sentiment analysis by POS and joint sentiment topic features using SVM and ANN. *Soft Computing* 23 (16):7067–79. doi:10.1007/s00500-018-3349-9.

Kim, W. 2021. A study on the subjective feeling affecting tactile satisfaction of leather in automobile: A structural equation modeling approach. *International Journal of Industrial Ergonomics* 84:103167. doi:10.1016/j.ergon.2021.103167.

Kim, W., B. Jin, S. Choo, C. S. Nam, and M. H. Yun. 2019a. Designing of smart chair for monitoring of sitting posture using convolutional neural networks. *Data Technologies and Applications* 53 (2):142–55. doi:10.1108/DTA-03-2018-0021.

Kim, W., T. Ko, I. Rhiu, and M. H. Yun. 2019b. Mining affective experience for a kansei design study on a recliner. *Applied Ergonomics* 74:145–53. doi:10.1016/j.apergo.2018.08.014.

Kim, W., Y. Lee, J. H. Lee, G. W. Shin, and M. H. Yun. 2018a. A comparative study on designer and customer preference models of leather for vehicle. *International Journal of Industrial Ergonomics* 65:110–21. doi:10.1016/j.ergon.2017.07.009.

Kim, W., D. Park, Y. M. Kim, T. Ryu, and M. H. Yun. 2018b. Sound quality evaluation for vehicle door opening sound using psychoacoustic parameters. *Journal of Engineering Research* 6 (2):176–190.

Kim, S., and J. Song. 2022. Semantic analysis via application of deep learning using naver movie review data. *The Korean Journal of Applied Statistics* 35 (1):19–33.

Kim, Y. M., Y. Son, W. Kim, B. Jin, and M. H. Yun. 2018. Classification of children's sitting postures using machine learning algorithms. *Applied Sciences* 8 (8):1280. doi:10.3390/app8081280.

Kitsios, F., M. Kamariotou, P. Karanikolas, and E. Grigoroudis. 2021. Digital marketing platforms and customer satisfaction: Identifying eWOM using big data and text mining. *Applied Sciences* 11 (17):8032. doi:10.3390/app11178032.

Kudo, T., and J. Richardson. 2018. Sentencepiece: A simple and language independent sub-word tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808 06226*. doi:10.48550/arXiv.1808.06226.

Lee, Y., W. Kim, J. H. Lee, Y. M. Kim, and M. H. Yun. 2020. Understanding the relationship between user's subjective feeling and the degree of side curvature in smartphone. *Applied Sciences* 10 (9):3320. doi:10.3390/app10093320.

Lim, J., and J. Kim. 2014. An Empirical Comparison of Machine Learning Models for Classifying Emotions in Korean Twitter. Journal of Korea Multimedia Society 17 (2):232–239. doi:10.9717/kmms.2014.17.2.232.

Lin, Y., J. Li, L. Yang, K. Xu, and H. Lin. 2020. Sentiment analysis with comparison enhanced deep neural network. *IEEE Access* 8:78378–84. doi:10.1109/ACCESS.2020.2989424.

Lipton, Z. C., J. Berkowitz, and C. Elkan. 2015. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506 00019*. doi:10.48550/arXiv.1506.00019.

Litvin, S. W., R. E. Goldsmith, and B. Pan. 2008. Electronic word-of-mouth in hospitality and tourism management. *Tourism Management* 29 (3):458–68. doi:10.1016/j.tourman.2007.05.011.

Liu, B. 2012. "Sentiment analysis and opinion mining. "*Synthesis Lectures on Human Language Technologies* 5 (1):1–167.

Monika, R., S. Deivalakshmi, and B. Janet (2019, December). Sentiment analysis of US airlines tweets using LSTM/RNN. In 2019 IEEE 9th International Conference on Advanced Computing (IACC), Chennai, India, (pp. 92–95). IEEE.

Moon, S., S. Park, D. Park, W. Kim, M. H. Yun, and D. Park. 2019. A study on affective dimensions to engine acceleration sound quality using acoustic parameters. *Applied Sciences* 9 (3):604. doi:10.3390/app9030604.

Moon, S., S. Shin, S. Kim, M. Jung, and J. Y. Jang (2022, June). Characteristic comparison of Korean unstructured dialogue corpora by morphological analysis. In 2022 IEEE 4th International Conference on Artificial Intelligence Circuits and Systems (AICAS), Incheon, Korea, (pp. 1–4). IEEE.

Park, D., S. Park, W. Kim, I. Rhiu, and M. H. Yun. 2019. A comparative study on subjective feeling of engine acceleration sound by automobile types. *International Journal of Industrial Ergonomics* 74:102843. doi:10.1016/j.ergon.2019.102843.

Polignano, M., P. Basile, M. De Gemmis, G. Semeraro, and V. Basile (2019). Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets. In 6th Italian Conference on Computational Linguistics, CLiC-it 2019, Bari, Italy, (Vol. 2481, pp. 1–6). CEUR.

Pota, M., F. Marulli, M. Esposito, G. De Pietro, and H. Fujita. 2019. Multilingual POS tagging by a composite deep architecture based on character-level features and on-the-fly enriched word embeddings. *Knowledge-Based Systems* 164:309–23. doi:10.1016/j.knosys.2018.11.003.

Rose, S., N. Hair, and M. Clark. 2011. Online customer experience: A review of the business - to - consumer online purchase context. *International Journal of Management Reviews* 13 (1):24–39. doi:10.1111/j.1468-2370.2010.00280.x.

Ryu, T., B. Son, and W. Kim. 2020. Analysis of perceived exertion and satisfaction in the opening and closing of tailgates of SUVs. *International Journal of Industrial Ergonomics* 80:103033. doi:10.1016/j.ergon.2020.103033.

Singh, C., T. Imam, S. Wibowo, and S. Grandhi. 2022. A deep learning approach for sentiment analysis of COVID-19 reviews. *Applied Sciences* 12 (8):3709. doi:10.3390/app12083709.

Son, Y., and W. Kim. 2023. Development of methodology for classification of user experience (UX) in online customer review. *Journal of Retailing and Consumer Services* 71:103210. doi:10.1016/j.jretconser.2022.103210.

Su, J., S. Yu, and D. Luo. 2020. Enhancing aspect-based sentiment analysis with capsule network. *IEEE Access* 8:100551–61. doi:10.1109/ACCESS.2020.2997675.

Taniguchi, Y., S. I. Konomi, and Y. Goda 2019. "Examining language-agnostic methods of automatic coding in the community of inquiry framework." In 16th International Conference on Cognition and Exploratory Learning in Digital Age IADIS Press, Cagliari, Italy, 19–26.

Tian, Y., M. Sun, Z. Deng, J. Luo, and Y. Li. 2017. A new fuzzy set and nonkernel SVM approach for mislabeled binary classification with applications. *IEEE Transactions on Fuzzy Systems* 25 (6):1536–45. doi:10.1109/TFUZZ.2017.2752138.

van der Heijden, N., S. Abnar, and E. Shutova (2020, April). A comparison of architectures and pretraining methods for contextualized multilingual word embeddings. In Proceedings of the AAAI Conference on Artificial Intelligence, Newyork, USA, (Vol. 34, No. 05, pp. 9090–97).

Vidal, L., G. Ares, and S. R. Jaeger. 2018. Application of social media for consumer research. In *Methods in consumer research*, ed. G. Ares and P. Varela, vol. 1, 125–55. Woodhead Publishing. doi:10.1016/B978-0-08-102089-0.00006-6.

Yang, L., Y. Li, J. Wang, and R. S. Sherratt. 2020. Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning. *IEEE Access* 8:23522–30. doi:10.1109/ACCESS.2020.2969854.

Yoo, C. W., G. L. Sanders, and J. Moon. 2013. Exploring the effect of e-WOM participation on e-loyalty in e-commerce. *Decision Support Systems* 55 (3):669–78. doi:10.1016/j.dss.2013.02.001.

Zou, J., Y. Han, and S. S. So. 2008. Overview of artificial neural networks. *Artificial Neural Networks*, vol. 458, 14–22.