ISSN 1229-1889(Print)
ISSN 2287-9005(Online)
J Korean Soc Qual Manag Vol. 48, No.3: 421-431, September 2020
https://dx.doi.org/10.7469/JKSQM.2020.48.3.421

고객 감성 분석을 위한 학습 기반 토크나이저 비교 연구

김원준†

성결대학교 산업경영공학과 AI경영기술연구소

Comparative Study of Tokenizer Based on Learning for Sentiment Analysis

Kim, Wonjoon†

The Research Institute of AI Management Technology, Department of Industrial & Management Engineering, Sungkyul University

ABSTRACT

Purpose: The purpose of this study is to compare and analyze the tokenizer in natural language processing for customer satisfaction in sentiment analysis.

Methods: In this study, a supervised learning-based tokenizer Mecab-Ko and an unsupervised learning-based tokenizer SentencePiece were used for comparison. Three algorithms: Naïve Bayes, k-Nearest Neighbor, and Decision Tree were selected to compare the performance of each tokenizer. For performance comparison, three metrics: accuracy, precision, and recall were used in the study.

Results: The results of this study are as follows; Through performance evaluation and verification, it was confirmed that SentencePiece shows better classification performance than Mecab-Ko. In order to confirm the robustness of the derived results, independent t-tests were conducted on the evaluation results for the two types of the tokenizer. As a result of the study, it was confirmed that the classification performance of the SentencePiece tokenizer was high in the k-Nearest Neighbor and Decision Tree algorithms. In addition, the Decision Tree showed slightly higher accuracy among the three classification algorithms.

Conclusion: The SentencePiece tokenizer can be used to classify and interpret customer sentiment based on online reviews in Korean more accurately. In addition, it seems that it is possible to give a specific meaning to a short word or a jargon, which is often used by users when evaluating products but is not defined in advance.

Key Words: Sentiment Analysis, Natural Language Processing, Machine Learning, Tokenizer, Online-Customer Review

● Received 6 August 2020, 1st revised 10 September 2020, accepted 16 September 2020
† Corresponding Author(wjkim@sungkyul.ac.kr)

※ 본 논문은 성결대학교 A.I.경영기술연구소에서 대학혁신지원사업비의 지원을 받아 수행한 연구입니다.

1. 서 론

기업 간 기술 수준의 평준화 현상으로 인해서 개발자들은 제품의 성능, 기능적 특징 및 가격 등과 같은 측면에서의 차별화를 달성하기 어렵다는 것을 인식하고 있다. 기업들은 제품 및 서비스의 개발에 있어서 고객의 감성 만족도를 극대화할 수 있는 매력적 품질을 향상하는 것에 집중하고 있다(Kim et al., 2009). 따라서, 심미적 경험, 감성적 만족도와 같은 주관적이고 추상적으로 표현되는 고객의 감성적 니즈를 충족시키기 위한 제품 개발에 초점을 맞추고 있다. 고객이 추구하는 감성 품질을 정량적으로 해석하기 위해서 감성 분석이라는 개념이 제품 개발에 도입되었다. 감성 분석의 방법론들은 고객의 개념적이고 암묵적인 요구를 정량화 할 수 있기 때문에, 자동차, 휴대폰, TV, 의자 등과 같은 다양한 제품에 적용되어 왔다.

고객의 감성 품질을 명확히 파악하고 이를 제품 개발에 적용하기 위해 감성 분석 연구에서는 고객의 감정이나 제품에 대한 인상을 정량적으로 측정하기 위해서 감성 어휘를 수집하는 다양한 방법을 제안 및 개발하였다 (Henson et al., 2006). 가장 일반적으로 활용되는 방법은 문헌 조사 및 인터뷰이지만, 이러한 방식의 고객 감성 수집은 몇 가지 문제점이 발생할 여지가 있다. 먼저, 피험자의 주관적 기준, 성별, 나이, 지역 등의 차이로 인해 감성 표현 및 평가에 편향이 발생할 수 있다. 예를 들어, Kuwano et al. (2009)은 일본인과 독일인 간의 자동차 소리에 대한 감정 평가 결과가 다르다는 것을 확인하였다. 따라서 제품의 특성이 동일하더라도 문화적 차이에 따라 사람들이 표현하는 감성적 경험에 차이가 있음을 알 수 있었으며, Montefinese et al. (2014)는 애정 속성 및 수준을 나타내는 단어의 종류 별 개인간의 차이가 존재한다는 것을 밝혔다. 둘째로, 사용자를 대상으로 감성 평가를 수행할 때, 일반적으로 평가 실험은 적은 수를 대상으로 수행되기 때문에 다양한 연령대, 성별, 인종 그룹 등에서 얻은 결과를 종합적으로 파악하기 어렵다. 이로 인해, 실험 결과를 바탕으로 도출된 감성 모형의 일반화가 어렵고 설계자의 의도로 인한 편향 가능성도 배제할 수 없다.

이러한 한계점을 극복하기 위해서, 텍스트 마이닝 기술을 통한 감성 모델 개발과 관련된 다양한 연구가 진행되고 있다(Yang et al., 2020). 텍스트 마이닝은 자연어 처리 기술(natural language processing, NLP)을 통해서 구조화되지 않은 텍스트에서 의미 있고 유용한 정보를 처리하는 기술을 뜻하며, 자연어 처리란 인간이 일상 생활에서 사용하는 언어를 컴퓨터가 이해, 분석 및 해석할 수 있도록 하는 일련의 과정을 의미한다 (Collobert et al., 2011). 최근, 사용자가 제품을 사용하면서 느낀 다양한 경험들을 평가하고 의견을 공유할 수 있는 웹에서의 고객 리뷰를 통해 고객의 감성을 파악하고자 하는 다양한 시도들이 이뤄지고 있다 (Fang and Zhan, 2015). 웹 기반 텍스트 마이닝은 사용자 조사나 전문가 인터뷰와 같은 기존의 감성 분류 방법에 비해 비교적 많은 수의 데이터를 다룸으로써 결과의 일반화를 도출하는데 장점이 있다.

자연어 처리는 일반적으로 토큰화(tokenization), 정제(cleaning), 어간 및 표제어 추출(stemming and lemmatization)의 과정을 거치는데, 이 중 토큰화는 주어진 코퍼스(corpus)를 토큰이라는 단위로 나누는 과정이다. 기존의 한글 기반 온라인 리뷰를 대상으로 한 감성 분석 연구에서는 형태소 분석 토크나이저가 주로 사용되었다 (Lim and Kim, 2014). 하지만, 온라인 상품평에는 줄임말, 은어, 이모티콘 등과 같은 인터넷 신조어들이 많이 포함되어 있고 이러한 유행어들은 확산 속도가 빠르고 활용되는 기간은 짧다. 따라서, 온라인 상품평 전체에 대한 형태소 분석을 수행하는 것은 현실적인 제약이 따른다. 최근, 구글에서 공개된 서브워드 텍스트 토크나이저인 SentencePiece는 미등록 어휘(Out of vocabulary, OOV)에 대한 어휘 모델을 생성하기 위해 특정 언어에 의존하지 않고 텍스트에 입력된 문장을 기반으로 서브 워드 모델을 학습하는 방법을 제안하였다. 영어, 일본어, 중국어, 프랑스어 등의 고객 리뷰를 대상으로 감성 분석을 수행한 기존 연구에서 SentencePiece를 활용하여 정확도를 높인 결과들이 보고되었기 때

문에, 한국어 상품평을 기반으로 한 감성 분류의 정확도를 개선하기 위해 SentencePiece의 적용 가능성 여부를 검토할 필요가 있다.

기술적 평준화에 의한 기업간 경쟁이 격화되고 있는 상황에서 고객이 제품이나 서비스를 이용할 때 발생하는 다양한 형태의 경험을 파악하고 이러한 경험들을 통합하여 감성 품질을 체계적으로 관리하는 것은 기업 경영의 핵심 가치 중 하나로 주목받고 있다. 본 연구는 최근 제품 및 서비스 사용 경험의 교류가 활발히 일어나는 웹에서 고객의 경험을 명확히 분류할 수 있는 평가 방법을 제안하는 것을 목표로, 자연어 처리를 통한 고객의 감성 분류의 고도화를 위해 학습 기반의 토크나이저의 성능을 비교 평가하고 분류 성능을 최적화할 수 있는 알고리즘을 제안하고자 한다. 특히, 비언어적 표현, 줄임말, 은어 및 신조어 등이 혼재되어 있는 온라인 상품 리뷰의 특성을 고려하여 다른 나라 언어에 적용되어 활용된 비지도 학습 기반의 토크나이저인 SentencePiece가 한국어에서도 통용될 수 있는지에 대한 여부를 확인하고자 한다. 비교 평가를 위해서 한국어 토크나이저로 실효성이 입증된 지도 학습 기반의 토크나이저인 Mecab-Ko가 선정되었으며, 평가 데이터로는 TV 온라인 상품 리뷰를 활용하였다.

2. 이론적 배경 및 선행연구

2.1 감성 분석과 온라인 고객 리뷰

감성 분석(sentiment analysis)은 제품이나 서비스에 대한 인간 감성의 여러 수준을 자연어 처리 방법을 통해 파악하는 것이다 (Liu, 2012). 온라인 리뷰, 소셜 미디어, 멀티미디어 공유 플랫폼의 확산으로 인해서 디지털 형태로 이뤄진 방대한 양의 텍스트 데이터가 웹에 축적되게 되었다. 감성 분석은 자연어 처리에서 가장 활발한 연구 분야로 성장하였으며, 사용자의 의견은 제품 및 서비스를 이용하는 인간 활동의 산출물이자 다른 사람의 의사 결정에 영향을 미치는 주요 요소이기 때문에 마케팅, 경영, 사회 과학, 커뮤니케이션 등 전방위 영역으로 그 중요성이 확산되었다(Balbi et al., 2018).

본 연구에서는 온라인 고객 리뷰(electronic word-of-mouth, e-WOM)를 웹사이트에 게시된 제품의 리뷰로 정의한다(Rose et al., 2011). e-WOM은 제품 관련 문의에 대한 정보 외에도 제품의 품질, 가격 및 구매 후기 등과 같은 제품 경험과 관련된 정보를 제공한다. Amazon으로 대표되는 온라인 소매 시장은 고객이 자신의 가치와 경험을 자유롭게 표현할 수 있으며, 이러한 표현들은 다른 사람의 선택에 영향을 줄 수 있다 (Gruen et al., 2006). 따라서, 온라인 상점 관리자들은 e-WOM을 효과적이고 효율적으로 관리하기 위해 노력한다 (Litvin et al., 2008).

제품 디자인 영역에서 고객의 경험과 감정의 관점에서 고객의 니즈를 충족시킬 수 있는 방안에 대한 관심이 높아지고 있다. e-WOM를 기반으로 다양한 관점에서 고객의 감성적 경험을 식별하기 위한 연구들이 많이 수행되었다 (Wang et al., 2010). 이러한 연구들은 e-WOM이 고객의 감성 경험, 선호도, 궁극적으로는 구매 의도와 행태에 미치는 영향 등을 조사한다 (Decker and Trusov, 2010). 최근에는 인공 신경망(artificial neural network, ANN), 지지기반 벡터(support vector machine, SVM) 등과 같은 학습 알고리즘을 사용하여 온라인 리뷰에서 고객의 감정을 분류하기 위한 연구들이 주로 수행되었다. Liu et al. (2013)은 온라인 리뷰 분석에서 부사 기반의 의견 기능 추출 방법을 사용하여 정확성을 향상시키는 추천 알고리즘을 개발했다. Jiang and Qi (2016)는 분류 알고리즘을 사용하여 온라인 리뷰에서 6개의 고객 감정을 분류했다.

2.2 SentencePiece

Kudo and Richardson (2018)에 의해 제안된 SentencePiece는 신경망 기반의 텍스트 생성 모델을 위한 비지도 텍스트 토크나이저 및 해독기이다. SentencePiece는 원시 문장에서 직접 학습 개념을 확장하여 2개의 하위 단어 분류 알고리즘인 byte-pair encoding(BPE)과 유니그램 언어 모델로 구현된다. BPE의 기본 원리는 텍스트 코퍼스에서 가장 많이 등장하는 문자열을 병합하여 문자열을 압축하고, 어휘 집합의 크기가 원하는 수준에 이르기까지 반복적으로 고 빈도 문자열들을 병합 및 추가하는 과정을 반복하는 것이다. SentencePiece는 형태소와 같은 언어별 사전 지식 없이 등장 빈도 기반의 단어 분리를 수행하기 때문에 언어의 형식이나 특징에 종속되지 않는 장점이 있다. SentencePiece는 네 가지 구성 요소: Normalizer, Trainer, Encoder 및 Decoder 로 구성된다. Normalizer는 의미적으로 동등한 유니코드 문자를 정규화 형태로 표준화하는 단계이고, Trainer는 정규화 된 코퍼스에서 서브워드 분할 모델을 학습하는 단계이다. Encoder는 입력 텍스트를 정규화하고 Trainer가 학습한 서브워드 모델을 사용하여 서브워드 순서로 토큰화하는 과정이고, 마지막으로 Decoder는 서브워드 순서를 정규화 된 텍스트로 변환하는 단계이다.

SentencePiece를 활용하여 웹 리뷰 기반 감성 분석과 관련된 기존 연구들을 살펴보면, Bérard et al. (2019)은 레스토랑의 리뷰를 영어-프랑스어로 번역하는 방법에 SentencePiece의 개념을 활용하였다. 그들은 감정 분석이나 도메인 별 번역 정확도를 기반으로 한 과업 특화 측정 지표들을 제안하였다. Su et al. (2020)은 감성 분석에서 기존에 주로 활용되던 BERT와 XLNet 같은 지도 사전-훈련 기반 모형을 보완하는 새로운 알고리즘인 XLNetCN을 제안하였으며, 레스토랑과 노트북 리뷰 데이터를 통해 제안된 모델의 우수성을 검증하였다. Bataa and Wu (2019)는 SentencePiece를 활용하여 전이 학습 기반 일본어 감성 분석을 수행하였다. 하지만, SentencePiece를 활용하여 한국어가 사용되는 상품평이나 리뷰에 대한 감성 분석을 수행한 연구는 아직 없다. 다양한 종류의 언어에서 감성 분석 시 SentencePiece의 우수성이 조사되었기 때문에, 한국어 기반 상품평과 리뷰에 대한 연구가 이뤄질 필요가 있다.

3. 연구방법

3.1 훈련 및 시험 데이터 세트

본 연구에서는 국내 최대 가격 비교 사이트인 다나와(www.danawa.com)에서 TV 카테고리에 해당하는 상품평 약 14만 건을 크롤링 하였다. 이 중, 내용이 중복되거나 상품평이 이모티콘과 같은 문자 이외의 형태로 이뤄졌거나 상품에 대한 별점이 없거나 별점만 있는 것, 단어수가 2개 미만인 것 등을 정제하여 최종적으로 133,535건의 상품평을 추출하였다. 또한, 본 연구에서는 웹에서 사용되는 다양한 형태의 줄임말, 은어, 비문과 같은 것들이 사용된 리뷰에서 고객의 감성을 보다 정확히 분류할 수 있는지에 대한 가능성을 파악하고자 했기 때문에 자연어 처리의 일반적인 전처리 과정을 생략했다.

다나와 사이트의 상품평 데이터를 이용하여 긍정, 부정을 나타내기 위한 레이블링(labelling)을 수행하기 위해서, 상품평 데이터에서 고객이 구매한 TV에 대해 평가한 평점 정보를 활용하였다. 평점은 별로 매겨지며 1개부터 5개까지 구매 고객이 평가를 매길 수 있다. 본 연구에서는 별 5개와 4개의 평점에 해당하는 상품평 데이터는 긍정으로, 별 1, 2, 3개의 평점에 해당하는 리뷰 데이터에 대해서는 부정으로 간주하여 레이블링을 수행하였다.

3.2 토큰화(Tokenization)

본 연구에서는 한국어 토큰화의 성능을 비교 및 평가하기 위해서 두 가지 종류의 토큰화 방법: 한국어 형태소 분석 엔진인 Mecab-Ko와 SentencePiece 방법을 비교하였다. Mecab-Ko는 한국어 문장을 토큰화하기 위해 널리 사용되는 형태소 분석기 중 하나이며, SentencePiece는 신경망 기계 번역을 위해 개발된 새로운 토큰화 방법으로 (Kudo and Richardson, 2018), 데이터 중심적 접근 기반 비지도 방식 텍스트 토크나이저이다. 지도 기반 접근 방식과 달리, SentencePiece는 사전 토큰화된 데이터 집합이 필요하지 않으므로 여러 언어가 혼합적으로 사용된 텍스트나 사전에 없는 단어를 포함하는 텍스트 코퍼스에 더 적합하다. SentencePiece 알고리즘과 일반적인 형태소 분석 엔진과의 중요한 차이점 중 하나는 토큰 수를 미리 지정하는 것에 있다. Taniguchi et al. (2019)에 의하면, 토큰의 개수가 SentencePiece 알고리즘의 분류 정확도에 영향을 주기 때문에 적절한 토큰 개수를 선정하는 것이 중요하다. 본 연구에서는 먼저, Mecab-Ko을 통해 모든 문장을 토큰화하고 발생된 모든 토큰을 계산하여, 총 33,991개의 토큰을 어휘로 선택하였다. SentencePiece의 경우, 최적 k를 3000, 5000, 10000, 15000, 20000으로 설정하였다. 또한, Mecab-Ko와 SentencePiece는 입력 값으로 UTF-8로 인코딩된 텍스트를 사용해야 하며, 본 연구에서 타겟 사이트로 사용된 다나와에서 수집한 상품평 텍스트는 모두 UTF-8 인코딩이 되어 있었으므로, 별도의 인코딩 변환 없이 사용하였다. 전체 상품평 데이터셋에서의 형태소 수 별 상품평의 분포는 아래 그림 1과 같다.

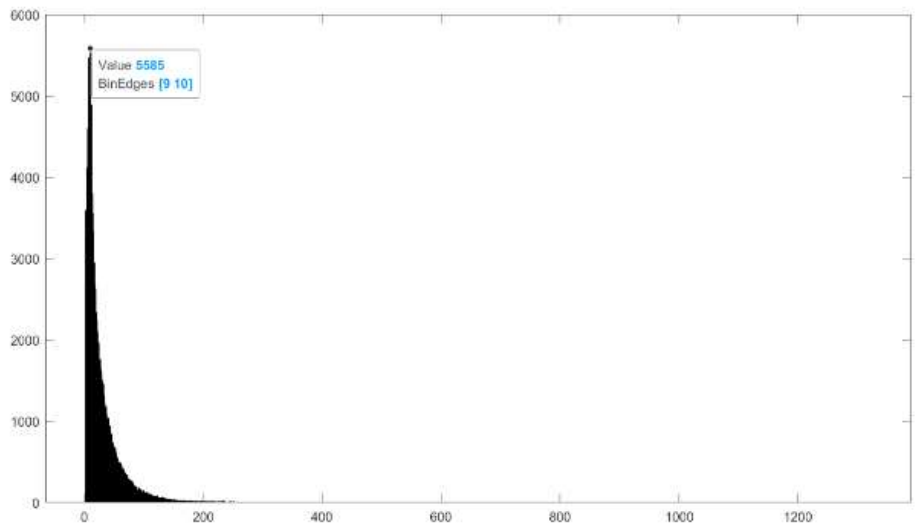

Figure 1. Number of reviews for morpheme distribution in the overall data

3.3 분류 기법 및 성능 평가 지표

본 연구에서는 한국어 상품평의 감성 분류에 대한 성능 평가를 위해서 세 가지 종류의 지도 학습 기반의 분류 알고리즘: 나이브 베이즈(Naïve Bayes, NB), k-최근접 이웃(k-Nearest Neighbor, kNN), 의사결정 나무(Decision Tree, DT)가 사용되었다. NB는 베이즈 정리(Bayes' Theorem)를 기반으로 하는 확률적 분류 알고리즘이다. 이 알고리즘은 클래스가 주어지면 모든 속성이 서로 독립적이라고 가정하며, 일반적으로 파라미터의 수를 줄이기 위해 널리 사용된다. NB는 베이즈 규칙, 조건부 독립성 가정 및 입력 데이터에 대한 분류 규칙으로 구성된다. NB는 다른 복잡한 그래픽 모형에 비해 분류에 필요한 파라미터 추정에 요구되는 데이터의 수가 적은 장점이 있다. k-NN은 데

이터 세트에서 학습 샘플과 테스트 샘플 사이의 거리를 계산하여 근접한 요소를 기반으로 분류하는 알고리즘으로 해석이 용이하고 계산 시간이 짧은 장점이 있다. DT 알고리즘은 데이터 세트에서 의사 결정 트리를 구조화하는 방법으로, 입력 변수의 공간을 나누고 분할된 부분에서 종속 변수의 값을 예측한다. 일반적으로 이 알고리즘은 성장, 가지 치기, 타당성 평가, 해석 및 예측 순으로 이뤄지며, Kamiński et al. (2018)에 따르면 DT 알고리즘은 결정, 기회 및 종료 노드의 세 가지 유형으로 구성된다.

기계 학습을 통해 도출된 모델의 분류 성능을 평가하기 위해서, 세 가지 성능 지표: 정확도(Accuracy), 재현율(Recall), 정밀도(Precision)가 활용되었다. 학습 모델의 성능을 평가하는 것은 실제 데이터 값과 모델을 통해 도출된 값 간의 관계로 파악할 수 있다. 정확도는 전체 케이스에서 정확하게 맞춘 비율을 나타낸다. 즉, 실제 데이터에서 '긍정'이었던 리뷰를 '긍정'으로 분류하고, '부정'이었던 리뷰를 '부정'으로 분류한 비율로 계산된다. 재현율이란, 실제 데이터에서 '긍정'이었던 것 중에서 학습 모델이 '긍정'으로 분류한 것의 비율로 계산된다. 마지막으로, 정밀도는 학습 모델이 '긍정'이라고 분류한 것 중에서 실제 리뷰가 '긍정'인 것의 비율이다.

3.4 분류 성능 검증

본 연구에서는 학습 모델을 통해 계산된 분류 성능을 검증하기 위해서 모델 검증을 수행하였다. 특정 데이터 셋을 대상으로 개발된 훈련 모델이 다른 데이터 셋을 정확하게 분류하지 못할 수도 있기 때문에 제안된 모델에 대한 교차 검증(cross validation)이 필수적으로 요구된다. 교차 검증을 수행하여 모델의 성능을 평가하면 전체 데이터 셋을 평가에 활용하기 때문에 특정 데이터가 평가에 활용되는 과적합(Overfitting)을 방지할 수 있는 장점이 있다. 본 논문에서는 각 학습 알고리즘의 분류 성능을 검증하기 위해서 10-fold cross validation이 활용되었다. 10-fold cross validation은 아래 그림 2와 같이 전체 데이터 셋을 10개의 subset으로 나눠서 10번의 평가를 수행하고 도출된 성능 지표의 평균값을 계산하여 모델의 성능을 평가하는 과정을 거친다.

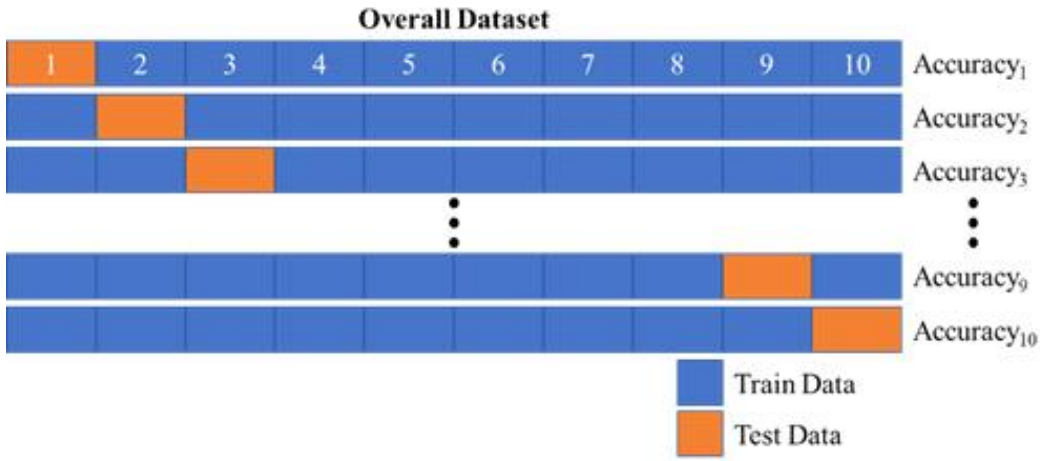

Figure 2. 10-fold cross validation method using this study

4. 연구결과

4.1 분류 알고리즘 별 지도 비지도 학습 토크나이저 비교 분석

고객 감성 분석을 위한 긍정, 부정 분류 알고리즘의 성능을 평가한 결과는 아래 표 1-3과 같다. 다나와 사이트의 TV 리뷰 데이터에 대한 분류 평가 결과, NB에서는 정확도와 재현율의 측면에서 비지도 학습 토크나이저인 SentencePiece는 선정된 모든 토큰 수에서 지도 학습 토크나이저인 Mecab-Ko 보다 높은 값을 나타냈다. 정밀도에서는 토큰 수가 2만 개인 경우를 제외하고 모두 Mecab-Ko보다 높거나 같았다. kNN에서는 정확도의 측면에서 비지도 학습 토크나이저인 SentencePiece의 토큰 수가 1,500~5,000개 일 때, 지도 학습 토크나이저인 Mecab-Ko 보다 높은 값을 나타냈으며, 3000개에서의 정확도 값이 0.956으로 가장 높았다. 정밀도에서는 토큰 수가 15,000, 25,000, 30,000개인 경우를 제외하고 모두 Mecab-Ko보다 높은 값을 기록했다. 마지막으로 DT에서는 세 가지 성능 지표: 정확도, 정밀도, 재현율 모두에서 비지도 학습 토크나이저인 SentencePiece는 선정된 토큰 수 3,000, 5,000개에서 가장 높은 값을 나타냈다. 또한, NB와 kNN과는 달리 토큰 수 30,000개 이상에서 알고리즘의 성능이 급격히 떨어지는 것이 확인됐다. 토큰 수에 따른 각 알고리즘 별 정확도 성능 비교의 결과는 아래 그림 3과 같다.

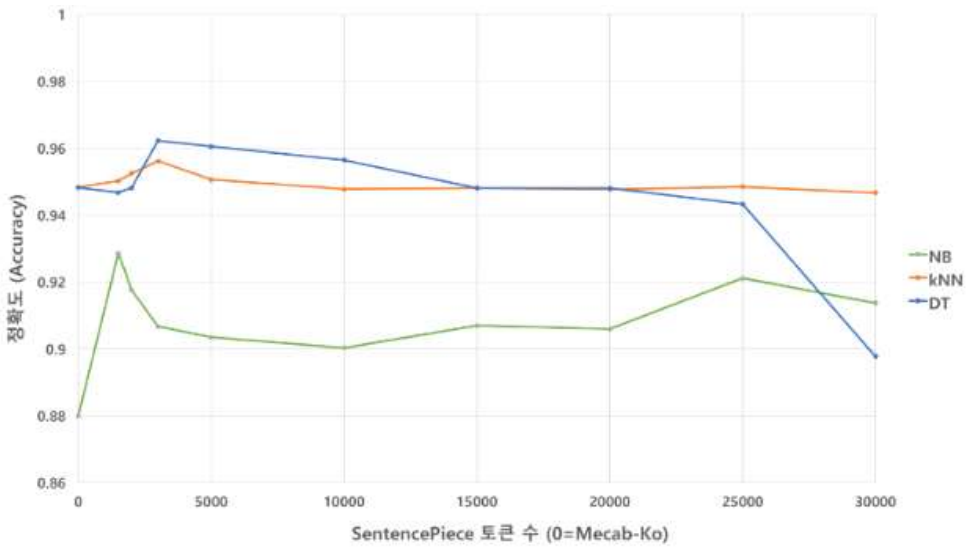

Figure 3. Result of the accuracy of learning algorithms for each token number of SentencePiece

Table 1. Results of performance metrics for SentencePiece and Mecab-Ko using Naïve Bayes

NB	Mecab-Ko	Number of Tokens in SentencePiece								
		1500	2000	3000	5000	10000	15000	20000	25000	30000
Accuracy	0.880	0.929	0.918	0.907	0.903	0.900	0.907	0.906	0.921	0.914
Precision	0.953	0.953	0.971	0.964	0.953	0.953	0.953	0.952	0.953	0.953
Recall	0.919	0.973	0.943	0.938	0.945	0.942	0.949	0.949	0.965	0.956

Table 2. Results of performance metrics for SentencePiece and Mecab-Ko using kNN

kNN	Mecab-Ko	Number of Tokens in SentencePiece								
		1500	2000	3000	5000	10000	15000	20000	25000	30000
Accuracy	0.948	0.950	0.952	0.956	0.951	0.948	0.948	0.948	0.949	0.947
Precision	0.957	0.958	0.961	0.964	0.960	0.958	0.957	0.958	0.957	0.956
Recall	0.990	0.991	0.990	0.991	0.990	0.989	0.990	0.989	0.991	0.989

Table 3. Results of performance metrics for SentencePiece and Mecab-Ko using DT

DT	Mecab-Ko	Number of Tokens in SentencePiece								
		1500	2000	3000	5000	10000	15000	20000	25000	30000
Accuracy	0.948	0.947	0.948	0.962	0.960	0.956	0.948	0.948	0.943	0.898
Precision	0.970	0.970	0.971	0.978	0.977	0.976	0.970	0.970	0.966	0.967
Recall	0.976	0.974	0.975	0.982	0.981	0.979	0.976	0.975	0.975	0.924

4.2 토크나이저 별 분류 성능 검증 분석

형태소 분석과 SentencePiece간의 분류 정확도의 통계적 유의성을 검증하기 위해서, 본 연구에서는 대응표본 t-검정을 사용하여 각 알고리즘 별 가장 높은 정확도를 보인 특정 토큰 수에 대한 SentencePiece 결과와 형태소 분석 결과를 비교하였다. 검증에서는 세 가지 알고리즘 중에서 높은 성능을 보인 kNN과 DT를 대상으로 하였다. 4.1장에서 언급된 바와 같이 SentencePiece의 토큰 수가 3,000개일 때의 두 알고리즘의 분류 성능 지표의 값이 가장 높았기 때문에, 이 때의 성능 지표들: 정확도, 재현율, 정밀도를 Mecab-Ko의 성능 지표와 비교하였다. 검증 실험을 위해서 전체 데이터 셋에서 훈련 및 시험 데이터를 90:10의 비율로 무작위 선정하여 각 알고리즘 별로 총 100회 반복 시행하였다. 대응표본 t-검정을 수행하기 앞서 토크나이저 별 두 개의 알고리즘: kNN, DT의 성능 지표 데이터의 정규성을 확인하기 위해서 콜모고르프-스미르노프(Kolmogorov-Smirnov) 검정과 샤피로-윌크(Shapiro-Wilk) 검정을 시행하였으며, 그 결과는 아래 표 4와 같다. 검정 결과, 모든 데이터에서의 p 값이 0.05보다 크므로 데이터는 정규성을 갖는다고 볼 수 있다. Mecab-Ko와 SentencePiece에 대한 대응표본 t-검정 결과는 아래 표 5와 같다. 분석 결과, 두 가지 알고리즘: kNN, DT에서의 세 가지 성능 지표 모두 통계적으로 유의한 차이가 있는 것으로 나타났다. 특히, kNN에서의 recall을 제외한 모든 알고리즘과 성능 지표에서 SentencePiece가 더 높은 것으로 밝혀졌다.

Table 4. Results of normality test for each performance metric of SentencePiece and Mecab-Ko

Type		Kolmogorov-Smirnova			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
kNN_Accuracy	SentencePiece	.073	100	.200	.994	100	.918
	Mecab-Ko	.065	100	.200	.977	100	.084
kNN_Precision	SentencePiece	.056	100	.200	.992	100	.839
	Mecab-Ko	.061	100	.200	.988	100	.535
kNN_Recall	SentencePiece	.065	100	.200	.980	100	.135
	Mecab-Ko	.078	100	.139	.984	100	.257

Type		Kolmogorov-Smirnova			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
DT_Accuracy	SentencePiece	.050	100	.200	.985	100	.333
	Mecab-Ko	.068	100	.200	.981	100	.149
DT_Precision	SentencePiece	.064	100	.200	.985	100	.331
	Mecab-Ko	.068	100	.200	.984	100	.270
DT_Recall	SentencePiece	.058	100	.200	.983	100	.218
	Mecab-Ko	.080	100	.116	.981	100	.171

Table 5. Results of paired t-test for the performance metrics of SentencePiece and Mecab-Ko

Performance Metric	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval	
						Lower	Upper
kNN_Accuracy	26.419	99	.000	.00693	.00027	.00640	.00746
kNN_Precision	32.437	99	.000	.00749	.00025	.00700	.00798
kNN_Recall	-2.531	99	.013	-.00042	.00017	-.00075	-.00009
DT_Accuracy	30.564	99	.000	.00921	.00030	.00861	.00982
DT_Precision	8.147	99	.000	.00189	.00023	.00143	.00234
DT_Recall	35.899	99	.000	.00761	.00021	.00718	.00803

5. 결 론

본 연구의 목적은 웹에서 발현 및 교류되는 고객 감성 경험의 분류 고도화를 위한 방법을 제안하여 기업 혹은 관련 이해 관계자에게 감성 품질에 대한 정보를 명확하게 제공하는 것이다. 이를 통해, 고객에게 제공되는 제품 및 서비스 품질에 대한 정보와 피드백을 정확하게 파악함으로써 기업의 의사 결정에 도움을 줄 것으로 기대된다. 본 연구는 국내 가격 비교 사이트에서의 제품 리뷰 데이터를 기반으로 고객의 감성을 분류할 때 신경망 기반 텍스트 처리를 위해 설계된 서브워드 토크나이저인 SentencePiece와 형태소 분석 토크나이저인 Mecab-Ko의 성능을 비교 및 분석하였다. 본 연구를 통하여 비지도 학습 토크나이저인 SentencePiece가 한국어 상품 리뷰 데이터를 활용한 감성 분석 연구에 적합하다는 것을 확인하였다. SentencePiece는 하위 단어에 대한 토큰화뿐만 아니라 문장의 텍스트를 토큰 ID의 시계열로 정방향, 역방향 변환이 가능하기 때문에 리뷰 데이터 기반 감성 분석에 대한 엔드 투 엔드 시스템에 적용할 수 있다.

본 연구에서는 온라인 리뷰에서 발현되는 고객의 대표 감성을 '긍정', '부정'으로 선정하였으며, 사람들이 일상에서 많이 사용하고 감성적 경험을 충분히 받을 수 있는 제품인 TV를 타겟 제품으로 선정하였다. 분류 성능을 확인하기 위해서, 세 가지 알고리즘이 채택되어 토크나이저의 분류 성능을 정확도, 정밀도, 재현율의 관점에서 검토하였다.

성능 평가와 검증 실험을 통해서 지도 학습 기반의 토크나이저인 Mecab-Ko보다 비지도 학습 기반의 토크나이저인 SentencePiece가 더 나은 성능을 보이는 것을 확인하였다. 도출된 결과의 강건성(robustness)를 확인하기 위해서 두 가지 형태의 토크나이저의 성능 평가 결과에 대해 독립표본 t-검정을 수행하였다. 연구 결과, kNN과 DT 알고리즘에서 SentencePiece 토크나이저의 분류 성능이 높은 것을 확인하였으며, DT가 세 가지 분류 알고리즘 중에 근

소하게 높은 정확도를 보이는 것을 확인할 수 있었다. 따라서, 본 연구에서 도출된 결과를 토대로 한국어 기반 온라인 상품평에서의 고객 감성을 보다 정확하게 분류하고 해석하는데 활용될 수 있을 것이다.

온라인 상품평에서의 감성 분류 문제에서 형태소 분석 기반 토크나이저에 비해 SentencePiece 기반 토크나이저의 성능이 우수한 결과를 보인 이유를 고찰해보면, 먼저, 온라인 상품평의 경우에는 뉴스 기사나 특허 문서에 비해 오타, 비문, 비표준어가 많기 때문에 Mecab-Ko와 같은 형태소 분석의 결과가 더 낮게 나온 것으로 보인다. 또한, 일정 볼륨 이상의 텍스트 코퍼스를 기반으로 최빈 표현들을 묶어서 학습하는 SentencePiece는 사람들이 특정 제품에 대한 상품평에서 주로 사용하는 동일한 음절 시퀀스에 기반하여 의미 있는 표현들을 그룹화 한다. 따라서, SentencePiece 토크나이저를 통해, 일정 수준 이상의 사용자들이 상품에 대한 평가에서 많이 사용하지만 사전에 정의되어 있지 않은 줄임말이나 은어 등과 관련된 용어에 대해서도 특정한 의미부여가 가능한 것으로 보인다. 또한, 제품 및 서비스에 대한 고객의 기대와 요구 사항이 점점 복잡해지고 많은 대체재가 존재하는 현 시대의 상황에서 기업이 고객의 니즈에 즉각적으로 반응하는 것은 기업 경영의 필수적인 요소로 자리잡았다. 본 연구를 통해 얻어진 결과를 확장하여 제품 및 서비스에 대한 긍정적, 부정적 감정에 대한 분류뿐만 아니라 감성 품질을 세분화할 수 있는 다양한 요소들을 분류하는데 활용할 수 있을 것으로 기대된다.

본 연구의 한계점 및 추후 연구 과제는 다음과 같다. 첫째, 본 연구는 고객 감성 분류의 성능을 확인하기 위해서 TV에 대한 리뷰 데이터를 활용하였다. 하지만, 기존 연구에서 제품 별 감성은 다르게 나타나는 경우가 많았기 때문에, 추후에는 다양한 제품에 대한 종합적 관점에서의 연구를 수행할 필요가 있다. 둘째, 연구에서 활용하지 않은 ANN, SVM과 같은 기계 학습 기반 성능 평가 알고리즘을 추가적으로 활용하여 토크나이저의 분류 정확도를 개선하는 시도를 할 필요성이 있다.

REFERENCES

Balbi, S., Misuraca, M., and Scepi, G. 2018. Combining Different Evaluation Systems on Social Media for Measuring User Satisfaction. Information Processing & Management 54(4):674-685.

Bataa, E., and Wu, J. 2019. An Investigation of Transfer Learning-based Sentiment Analysis in Japanese. arXiv Preprint arXiv:1905.09642.

Bérard, A., Calapodescu, I., Dymetman, M., Roux, C., Meunier, J. L., and Nikoulina, V. 2019. Machine Translation of Restaurant Reviews: New Corpus for Domain Adaptation and Robustness. arXiv Preprint arXiv:1910.14589.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. 2011. Natural Language Processing (almost) from Scratch. Journal of Machine Llearning Research 12:2493-2537.

Decker, R., and Trusov, M. 2010. Estimating Aggregate Consumer Preferences from Online Product Reviews. International Journal of Research in Marketing 27(4):293-307.

Fang, X., and Zhan, J. 2015. Sentiment Analysis Using Product Review Data. Journal of Big Data 2(1):5.

Gruen, T. W., Osmonbekov, T., and Czaplewski, A. J. 2006. eWOM: The Impact of Customer-to-customer Online Know-how Exchange on Customer Value and Loyalty. Journal of Business Research 59(4):449-456.

Henson, B., Barnes, C., Livesey, R., Childs, T., and Ewart, K. 2006. Affective Consumer Requirements: A Case Study of Moisturizer Packaging. Concurrent Engineering 14(3):187-196.

Jiang, S., and Qi, J. 2016. Cognitive Detection of Multiple Discrete Emotions from Chinese Oonline Reviews. In 2016 IEEE First International Conference on Data Science in Cyberspace (DSC) 137-142.

Kamiński, B., Jakubczyk, M., and Szufel, P. 2018. A Framework for Sensitivity Analysis of Decision Trees. Central European Journal of Operations Research 26(1):135-159.

Kim, J. Y., Kim, H. J., & Kim, C. M. (2009). The Influence of Service Elements on Customers' Emotion and Loyalty-Focused on Specialty Coffee Shop Customers. Culinary Science and Hospitality Research, 15(1), 271-286.

Kudo, T., and Richardson, J. 2018. Sentencepiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. arXiv Preprint arXiv:1808.06226.

Kuwano, S., Namba, S., Takehira, O., and Fastl, H. 2009. Subjective Impression of Copy Machine Noises: An Examination of Physical Metrics for the Evaluation of Sound quality. In Proc. Inter-Noise 2009 Ottawa, Canada.

Lim, J. S., and Kim, J. M. 2014. An Empirical Comparison of Machine Learning Models for Classifying Emotions in Korean Twitter. Journal of Korea Multimedia Society 17(2):232-239.

Litvin, S. W., Goldsmith, R. E., and Pan, B. 2008. Electronic Word-of-mouth in Hospitality and Tourism Management. Tourism Management 29(3):458-468.

Liu, B. 2012. Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies 5(1):1-167.

Liu, H., He, J., Wang, T., Song, W., and Du, X. 2013. Combining User Preferences and User Opinions for Accurate Recommendation. Electronic Commerce Research and Applications 12(1):14-23.

Liu, Y., Jin, J., Ji, P., Harding, J. A., and Fung, R. Y. 2013. Identifying Helpful Online Reviews: a Product Designer's Perspective. Computer-Aided Design 45(2):180-194.

Montefinese, M., Ambrosini, E., Fairfield, B., and Mammarella, N. 2014. The Adaptation of the Affective Norms for English Words (ANEW) for Italian. Behavior Research Methods 46(3):887-903.

Rose, S., Hair, N., and Clark, M. 2011. Online Customer Experience: A Review of the Bbusiness-to-consumer Oonline Purchase Context. International Journal of Management Reviews 13(1):24-39.

Su, J., Yu, S., and Luo, D. 2020. Enhancing Aspect-Based Sentiment Analysis With Capsule Network. IEEE ACCESS 8:100551-100561.

Taniguchi, Y., Konomi, S. I., and Goda, Y. 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 19-26.

Wang, H., Lu, Y., and Zhai, C. 2010. Latent Aspect Rating Analysis on Review Text Data: a Rating Regression Approach. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 783-792.

Yang, L., Li, Y., Wang, J., and Sherratt, R. S. 2020. Sentiment Analysis for E-commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning. IEEE Access 8:23522-23530.

저자소개

김원준 성결대학교 산업경영공학과 교수로 재직중이며, 주요 관심분야는 감성 공학, 가상 현실, 데이터 기반 사용자 경험 분석, 심층 학습 등이다.