「정보시스템연구」 제31권 제4호
한국정보시스템학회
2022년 12월, pp. 01~23

http://dx.doi.org/10.5859/KAIS.2022.31.4.1

# 작성자 언어적 특성 기반 가짜 리뷰 탐지 딥러닝 모델 개발

신동훈* · 신우식** · 김희웅***

〈목 차〉



## Ⅰ. 서 론

온라인 플랫폼 시장이 발전하면서, 소비자가 정보탐색을 위해 제품이나 판매처에 대한 리뷰를 찾아보는 행동 양상이 확대되고 있다(강지우 등, 2017). 온라인 플랫폼에서는 상품과 서비스를 직접 볼 수 없기 때문에 대다수의 소비자는 앞선 구매자들의 후기를 활용하여 구매 여부를 결정한다(야오즈옌 등, 2020; 이호근, 곽현, 2013). 긍정적인 리뷰는 온라인 구전 마케팅의 형태로 더 많은 잠재적 소비자의 구매를 유도하지만, 부정적 리뷰는 반대로 소비자의 구매 전환율을 낮추어 기업의 이익에 부정적 영향을 미칠 수 있다. 소비자 입장에서뿐만 아니라 상품의 판매자도 사용자들의 후기를 통해 상품의 문제점을 파악하고 상품 개선에 반영할 수 있기 때문에 온라인 리뷰의 중요성이 더욱 증대되고 있다(강성안 등, 2021; 박지현, 김종권, 2017; 사공원 등, 2016).

온라인 리뷰가 여행 및 숙박, 온라인 소매업체, 엔터테인먼트 등 다양한 영역에서 큰 영향력을 미침에 따라 직접 경험하지 않고 허위로 후기를 남기는 가짜 리뷰 작성 행태가 문제로 떠오르고 있다. 주로 나타나는 행태로는 기업에

* 현대카드, dh_shin@yonsei.ac.kr(주저자)
** 연세대학교 정보대학원, woosik.shin@yonsei.ac.kr
*** 연세대학교 정보대학원, kimhw@yonsei.ac.kr(교신저자)

서 제품 및 서비스 판매 증진을 위해 전문 리뷰어를 고용하여 제품의 인기를 높이고 잠재적 소비자를 유치하고자 제품에 대한 긍정적 가짜 리뷰를 작성하는 것이 있다(Wang et al., 2020). 또는 악의적으로 경쟁 상대 또는 불특정 업체를 폄하하거나 거짓 정보를 리뷰로 등록하는 현상이 있다. 이러한 가짜 리뷰 행태는 꾸준히 증가하고 있어 심각한 사회적 문제로 대두되고 있다(이민철, 윤현식, 2020).

최근 국내 가짜 리뷰 관련 행태는 배달 및 이커머스 분야에서 두드러지고 있다. 이와 관련하여 배달의 민족, 요기요, 쿠팡 등의 플랫폼에서 가짜 리뷰가 건당 5,000원에서 8,000원 수준으로 거래되는 현황도 공개되었다.[1] 영국 경쟁시장청의 발표에 따르면, 영국에서 발생하는 약 36조원 상당의 온라인 쇼핑 거래가 리뷰에 영향을 받는다고 발표하였고, 가짜 리뷰로 인하여 이커머스 플랫폼 내 사업자들이 사업을 중단하는 피해 사례가 급증하고 있다고 밝혔다.[2] 악의적 가짜 리뷰를 경제적으로 환산할 경우 전 세계적으로 온라인 지출에 미치는 영향은 약 1,520억 달러에 달했다는 점을 통해 가짜 리뷰로 인한 사회적 피해가 심각하다는 것을 알 수 있다.[3] 이처럼 가짜 리뷰는 소비자, 공급자 양측에 부정적 영향을 미치며 온라인 플랫폼 생태계의 신뢰성을 심각하게 훼손시키고 있다. 따라서, 소비자들의 합리적 선택에 방해 요소로 작용하며 소상공인과 브랜드 업체들에 악영향을 미칠 수 있는 가짜 리뷰에 대한 탐지와 관리가 필요한 상황이다.

이에 가짜 리뷰 탐지와 관련한 다양한 선행연구들이 진행되어오고 있다. 선행연구는 가짜 리뷰를 작성하는 고객 특성을 파악하는 연구와 가짜 리뷰를 탐지하는 자동화 모델을 개발하는 연구로 구분된다. 초기에는 사회과학 분야에서 가짜 리뷰 작성자의 행동적, 언어적 특성을 파악하기 위해 리뷰에서의 어휘 빈도를 기반으로 다양한 통계적 수치를 추출하여 유의 변수를 파악하는 연구가 주를 이루었다(Ott et al., 2011; Ong et al., 2014). 이후, 머신러닝 기술이 발달하며 유의 변수를 파악하는 연구뿐만 아니라 컴퓨터 공학 분야에서는 리뷰 텍스트 데이터를 활용하여 가짜 리뷰를 탐지하는 자동화 모델을 개발하는 연구로 확대되어왔다(Ball and Elworthy, 2014; Li et al., 2011; Fontanarava et al., 2017). 최근에는 딥러닝을 활용한 자연어 처리 기술이 급속도로 발전하며, 임베딩 방법론을 기반으로 리뷰 데이터의 의미론적 특성을 활용한 탐지 모델이 등장하였다(Shukla et al., 2019). 하지만 여전히 가짜 리뷰 탐지 모델 구축을 위한 한계점도 존재한다. 먼저, 작성자 행동 및 언어적 특성을 파악하는 유의 변수 검증 연구는 대부분 활용한 데이터 수가 충분하지 않고 다양한 도메인에 일반화가 어려우며 데이터 수에 따라 유의 수준의 변동이 있다는 한계점이 존재하였다(Yoo and Gretzel, 2009). 이후에 연구된 머신러닝 기반의 탐지 모델 또한 작성자의 언어적 특성 변수만을 활용한 탐지 모델 구축은 성능적 한계를 보였다(Li et al., 2011). 딥러닝 기반의 자연어 처

1) https://www.hankyung.com/life/article/2020091864071
2) https://www.bbc.com/korean/news-54069025
3) https://www.weforum.org/agenda/2021/08/fake-online-reviews-are-a-152-billion-problem-heres-how-to-silence-them/

리 기술을 활용한 탐지 모델이 성능적 발전에 기여했지만 의미론적 특성이 가짜 리뷰의 모든 특성을 설명하기에는 한계가 존재한다고 지적되었다. 최근 가짜 리뷰들은 인간이 직접 분류할 때조차 진짜 리뷰와의 의미론적인 차이를 찾기 어려워졌기 때문이다. 그리고 선행 연구를 통해 가짜 리뷰를 작성하는 리뷰어들은 공통적 언어적 패턴을 보이기 때문에, 단순히 의미론적인 정보만을 활용하는 것이 아닌, 언어적 특성도 함께 반영해야 하는 필요성이 제기되었다(Dematis et al., 2018). 하지만, 이와 같이 사회과학 분야에서 주로 연구된 가짜 리뷰의 언어적 특성과 컴퓨터 공학 분야에서 적용된 비정형적 특성을 융합하는 접근은 가짜 리뷰 관련 연구에서는 부족한 실정이다.

따라서, 본 연구의 목적은 가짜 리뷰에서 드러나는 유의한 작성자 언어적 특성과 딥러닝 모델 기반의 의미론적 특성을 추출하고, 이 두 가지 특성을 결합한 탐지 모델을 개발하고자 한다. 작성자의 언어 패턴과 관련한 다양한 통계적 수치를 추출하고 리뷰의 의미론적 특성과 결합한 모델을 개발함으로써, 작성자의 언어적 특성이 가짜 리뷰의 특성을 설명하는 역할 뿐만 아닌 자동화된 탐지 모델의 중요 변수로 활용될 수 있다는 시사점을 도출하고자 한다. 또한 각 특성을 단일적으로 활용하여 모델을 구축한 선행연구와 대비하여 다양한 특성을 결합한 모델의 성능 향상 가능성을 제안하고자 한다.

본 연구는 전 세계 최대 온라인 리뷰 사이트인 Yelp 리뷰 데이터를 활용한다. 총 358,071개의 리뷰 데이터 중 36,475개의 리뷰가 플랫폼에서 가짜 여부로 표시되어있으며 비중은 약 10.2%이다. 리뷰의 언어적 패턴 관련 통계적 특성을 추출하기 위해 본 연구에서는 LIWC (Linguistic Inquiry and Word Count) 소프트웨어를 활용하였으며, 5가지 범주에 해당하는 24개의 변수를 추출하여 linguistic features로 활용하였다. 또한 의미론적 특성 추출을 위해 recurrent neural network (RNN) 계열의 모델과 Transformer 모델을 활용하여 semantic features를 추출하고 학습하여 모델 간 성능을 비교하고자 한다.

# Ⅱ. 개념적 배경

## 2.1 가짜 리뷰

온라인 플랫폼에서의 리뷰는 소비자의 행동과 구매 의도에 큰 영향력을 미치고 있으며 이에 따라 실제로 경험하지 않거나 진실과 다르게 리뷰를 작성하는 가짜 리뷰가 함께 증가하고 있다(Babić Rosario et al., 2016). 가짜 리뷰의 주 형태로는 의도적으로 부정적인 리뷰를 작성하는 행위와 일정한 보상을 제공하며 마케팅 목적으로 긍정적인 리뷰를 작성하도록 유도하는 방식이 있다(Dellarocas and Narayan, 2006). 실제 경험에 근거하지 않은 정보를 제공하는 가짜 리뷰는 소비자, 공급자 양측에 부정적 영향을 미치며 온라인 플랫폼 생태계의 신뢰성을 심각하게 훼손시킨다. 최근 사례로는 코로나바이러스 유행 기간에 백신 미접종자의 입장을 제한하였다는 이유로 해당 음식점들이 대규모의 악성 가짜 리뷰를 받아 영업에 큰 피해를 본 사례가 있다. 온라인 플랫폼 시장의 성장

에 있어 대규모로 생성되고 있는 가짜 리뷰를 즉각적으로 탐지하여 관리하는 시스템의 확립이 필요한 실정이다.

## 2.2 선행 연구

가짜 리뷰 관련 선행연구는 가짜 리뷰가 지니고 있는 언어 통계적 특성을 바탕으로 유의한 변수를 탐색하는 연구와 리뷰 텍스트 데이터를 활용하여 가짜 리뷰 판별 모델을 개발하는 연구로 구분될 수 있다. 가짜 리뷰 판별에 유의한 요인들을 탐색하는 연구는 리뷰 데이터로부터 작성자의 언어적 특성을 대표할 수 있는 형태소, 도메인 관련 어휘 빈도 기반 언어 통계적 수치를 추출하거나 감성 점수를 활용하였다. Ott et al.(2011)은 가짜 리뷰와 진짜 리뷰 사이의 품사 분포 차이를 활용하여 가짜 리뷰가 진짜 리뷰보다 동사가 더 많고, 명사가 더 적은 경향을 보이는 것을 보였다. Yoo and Gretzel(2009)은 호텔 리뷰 플랫폼의 리뷰 데이터셋을 활용하여 가짜 리뷰와 진짜 리뷰의 언어 구조를 비교하였으며, 어휘의 복잡성, 1인칭 대명사의 사용, 브랜드명 언급 및 감정의 측면에서 가짜 리뷰와 진짜 리뷰 간 차이가 있음을 보였다. Ong et al.(2014)은 텍스트의 정보성, 가독성 변수가 가짜 리뷰의 유의한 특성임을 검증하였고 유사하게 리뷰 내용의 세부 정도, 이해도, 작성 형식 등 인지적 지표 측면에서의 변수를 도출 및 검증한 연구도 존재한다 (Banerjee and Chua, 2014). 최근 연구에서는 LIWC를 통해 포괄적으로 추출한 감정, 명사 사용 빈도, 특정 단어 활용 빈도 등이 가짜 리뷰 판별에 어떻게 영향을 주는지를 파악하였다 (Moon et al., 2021).

관련 연구로 온라인 리뷰로부터 추출한 다양한 언어적 특성과 리뷰의 유용성과의 관계에 관한 연구도 존재한다. Peng et al.(2018)은 확률적 언어 용어 집합을 활용하여 온라인 리뷰의 요약 정보를 추출하고 리뷰 유용성에 대한 영향력을 보았다. Wang et al.(2019)은 음식점 리뷰 데이터를 활용하여 8가지 감정적 차원 변수와 linguistic style matching (LSM) 변수를 추출하고 이항 회귀 분석을 활용하여 음식점 리뷰의 유용성에 유의한 영향을 미치는 것을 보였다. 가짜 리뷰 또는 리뷰 유용성에 관련된 다양한 언어적 특성 변수를 활용한 연구들에 대한 모델과 실험 결과는 표 1과 같다.

다음은 자동화된 가짜 리뷰 탐지 모델 개발 연구이며, 초기에는 주로 감성 분석, 형태소 분석, 그리고 가독성 수치와 같은 텍스트 데이터의 통계적 수치를 주요 변수로 활용하여 가짜 리뷰 판별을 진행하였다. Ball and Elworthy (2014)은 형태소 분석을 통해 나온 분포를 수치화하여 변수로 활용하였고 PCA와 로지스틱 회귀를 활용하여 분류 모델을 구축하였다. Li et al.(2011)와 Fontanarava et al.(2017)은 형태소 분석뿐만 아니라, 리뷰 데이터에서 특정 어휘의 빈도를 기반으로 감성 점수, 신뢰성, 주관성 등의 변수를 추출하여 모델에 적용하였고 각각 Naïve Bayes, Random Forest를 통해 다양한 변수를 활용하여 성능 향상의 가능성을 보였다. 이후 딥러닝 기술의 발전에 따라 단순히 통계적 수치를 기반으로 한 모델 개발 대신 언어의 의미론적 특성을 학습할 수 있기 위해 임베딩 기반의 딥러닝 모델 개발 연구도 꾸준히 진행되어왔다. Ren and Ji (2017)는 리뷰의 의미론

<표 1> 가짜 리뷰 관련 선행 연구

| 목적 | | Reference | 활용 변수 | 이용 모델 | 실험 결과 (유의 변수 / 예측 성능) |
|---|---|---|---|---|---|
| 유의 변수 탐색 | | OTT et al. (2011) | POS | SVM | Past tense, Third person singular, Adverbs |
| | | Banerjee and Chua (2014) | readability, POS, writing style | Logistic Regression | Readability, Verb, Past tense, function words, self-reference words, perceptual words |
| | | Yoo and Gretzel (2009) | length of the review, complexity, Branding, sentiment | T-test | Branding, Positive & Negative Sentiment, Complexity |
| | | Ong et al. (2014) | informativeness, product usage experience, readability | T-test | Informativeness, Readability, Subjectivity level |
| | | Wang et al. (2019) | sentiment, linguistic style matching | Negative binomial regression | Readability, Consumer elite status, review elapsed days, Sentiment Features |
| | | Moon et al. (2021) | emotion, pronoun, Cognitive heuristics, time-orientation | stepwise logistic regression | Emotional exaggeration, more negative, tone, 1st-person singular, Present tense |
| 예측 모델 | (Statistical) Linguistic Features | Ball and Elworthy (2014) | POS | PCA, Logistic Regression | Accuracy = 0.717<br>Precision = 0.729<br>Recall = 0.712<br>F1-Score = 0.718 |
| | | LI et al. (2011) | sentiment, rating | Naïve Bayes | Precision = 0.517<br>Recall = 0.669<br>F1-Score = 0.583 |
| | | Fontanarava et al. (2017) | rating, textual feature, temporal feature | Random Forest | Precision = 0.776<br>Recall = 0.861<br>F1-Score = 81.6 |
| | | Bhargava et al. (2019) | POS, sentiment | SVM, Logistic Regression, LSTM, CNN, RNN | Accuracy = 0.813<br>Precision = 0.782<br>Recall = 0.844<br>F1-Score = 0.812 |
| | Semantic Features | Ren and Ji (2017) | textual word embedding | RNN | F1-Score = 82.3<br>Accuracy = 83.5 |
| | | Shukla et al. (2019) | word embedding | RNN | AUC = 0.724<br>Accuracy = 0.938 |
| | | Hajek et al. (2020) | word embedding, emotion indicator | DFNN, CNN | F1-Score = 0.901<br>Accuracy = 0.898 |

*Note*. POS(part-of-speech)는 품사를 나타내며 Noun, Adjective, Preposition, Verb 등이 있다. SVM(support vector machine), PCA(principal component analysis), LSTM(long short-term memory), CNN(convolutional neural network), RNN(recurrent neural network), DFNN(deep feed-forward neural network).

적 표현을 추출하기 위해 워드 임베딩을 활용하고 대표적으로 텍스트 분류에 활용되는 딥러닝 모델인 RNN과 CNN 기반의 모델을 구축하였다. 그 결과 로지스틱 회귀 모델과 비교하여 향상된 성능을 보였으며 리뷰 내용에 대한 의미론적 특성 추출의 중요성을 시사하였다. Shukla et al.(2019)은 의사 리뷰 데이터에 대해서 Word2Vec 기반의 워드 임베딩 방법론을 활용하여 RNN 모델을 구축하였다. 그 결과, Random Forest와 같은 기존 머신러닝 모델 대비하여 딥러닝 모델의 성능이 향상될 수 있다는 시사점을 입증했다. Bhargava et al.(2019)는 호텔 플랫폼에서의 가짜 리뷰 탐지 모델 개발을 위해, 다양한 통계적 변수를 기반으로 한 SVM, Naïve Bayes, Logistic Regression과 같은 전통적 분류 모델과 RNN, CNN, LSTM을 활용한 딥러닝 모델 간의 성능을 비교하였다. 그 결과, 딥러닝 모델이 전통적 머신러닝과 비교하여 더욱 높은 성능을 보인다는 점을 입증하였다. 이러한 연구 결과를 통해 가짜 리뷰 탐지를 위해 딥러닝 기반의 언어 모델이 적용되어야 하는 시사점을 도출할 수 있다. Hajek et al.(2020)은 리뷰의 의미론적 특성뿐만 아니라 리뷰 내에서의 숨겨진 감정 지표 또한 텍스트 분류에서 핵심적인 역할을 할 것으로 보며 이 두 가지 특성을 결합하여 변수로 활용하는 신경망 모델을 개발하였다. 해당 모델 또한 텍스트 데이터의 통계적 수치 변수를 활용한 머신러닝 모델보다 높은 성능을 보였고, 감정 지표 변수가 결합되어 더욱 정밀한 가짜 리뷰 탐지가 가능하게 된 것을 입증하였다.

하지만 선행 연구 대부분의 경우 활용 데이터 수가 적고 데이터 수의 차이에 따라 유의 변수의 차이도 함께 발생하는 문제가 존재한다. 가짜 리뷰와 진짜 리뷰의 특성을 탐색한 연구에서도 단순히 리뷰의 구조적 특성을 기준으로 가짜 리뷰와 진짜 리뷰를 정확하게 탐지하는 데 한계가 있음을 지적하고 있다. 그뿐만 아니라, 리뷰의 의미론적 특성을 반영하여 딥러닝 언어 모델을 구축한 연구에서도 사람이 단순히 리뷰만 보고 소비자의 실제 리뷰인지 가짜 리뷰인지 분류하기 어려울 정도이기 때문에 단순히 의미적인 이해만으로는 정확하게 가짜 리뷰를 탐지하는 데 어려움이 존재한다고 밝혔다. 또한, TF-IDF를 활용한 방법론은 단순히 토큰화된 리뷰의 형태소의 종류와 개수만을 고려하며, Doc2Vec, Word2Vec 기반의 임베딩 방법론은 리뷰 내용의 유사도나 전체적인 의미만을 반영하기 때문에 가짜 리뷰만이 지니고 있는 특성을 모두 반영하는데 한계점이 존재한다. 즉 선행연구에서 구축된 자동화 모델들은 가짜 리뷰 작성자의 주요 행동 특성에 대한 고려가 부족한 실정이다.

따라서 본 연구에서는 미국의 온라인 리뷰 플랫폼인 Yelp에서 제공하는 약 35만 개의 데이터셋을 활용하여 선행 연구의 한계점을 보완하고자 한다. 또한 LIWC를 활용하여 리뷰 텍스트 데이터로부터 작성자의 언어적 특성을 대표하는 다양한 통계적 특성을 추출하고 딥러닝 모델을 통해 추출한 문맥적 의미를 함께 결합하여 이 두 가지 특성을 함께 학습할 수 있는 모델을 개발하고자 한다.

# Ⅲ. 연구 방법

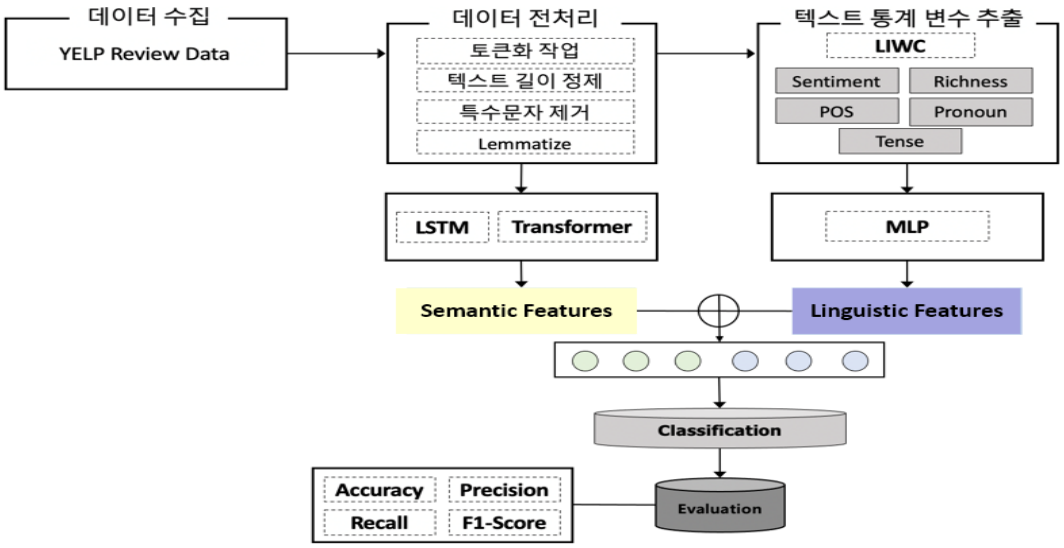
## 3.1 연구 절차

본 연구의 절차는 데이터 수집, 전처리 및 텍스트 통계 변수 추출, 딥러닝 모델 학습 및 성능 평가로 진행된다. 우선, 본 연구에서는 가짜 리뷰 탐지 모델 개발을 위해 Yelp 리뷰 데이터를 활용한다. Yelp는 최대 온라인 소비자 리뷰 사이트로 레스토랑, 호텔, 전자기기, 약국 등 다양한 업종에서의 서비스에 대한 소비자의 리뷰 데이터를 보유하고 있어 특정 도메인에 구애받지 않아 리뷰 관련 연구 대상으로 광범위하게 사용된다(Kc and Mukherjee, 2016). 또한, Yelp는 최소 100자 이상의 리뷰를 요구하기 때문에 다른 플랫폼에서의 리뷰 데이터에 비해 신뢰성이 높으며 획득할 수 있는 가치가 크다(Mukherjee et al., 2013; Rayana and Akoglu, 2015). 본 연구에서 활용한 Yelp 리뷰 데이터는 2004년부터 2015년까지의 Yelp 자체적 필터링을 통해 가짜 여부가 라벨링 되어있는 데이터셋이다. 총 358,071개의 데이터 중 가짜 리뷰의 비중은 36,475개이며 약 10.2%를 차지한다.

다음으로 전처리 작업을 진행하고 전처리 된 텍스트 데이터를 기반으로 LIWC를 활용해 5가지 범주(감성 (Sentiment), 풍부성(Richness), 형태소(POS), 대명사(Pronoun), 시제(Tense))에 해당하는 변수를 추출한다. 이 변수들은 linguistic features로 활용되고 텍스트 원본 데이터는 다차원 임베딩 과정을 통해 semantic features로 활용된다. 이러한 두 가지 특성들은 딥러닝 모델에 투입되어 학습되고 최종적으로 성능 평가가 진행되며 전체 과정을 도식화하면 그림 1과 같다.

## 3.2 데이터 전처리 및 변수 추출

리뷰 데이터에 대한 전처리를 진행하기 위해,




<그림 1> 연구 절차

토큰화 작업, 텍스트 길이 정제, 특수문자 및 표제어 추출(lemmatize) 작업을 거쳤다. 가장 먼저 토큰화 작업을 위해, Sentencepiece패키지를 활용하였다. Sentencepiece는 글자의 빈도를 기반으로 새로운 단어 묶음을 학습하는 방식의 토큰화 기법으로 사전 기반 토큰화 기법에서 발생하는 out of vocabulary 문제를 방지하고 전처리가 되지 않은 텍스트 데이터에도 사용할 수 있다는 장점이 있다(Kudo and Richardson, 2018). 리뷰 데이터는 은어, 유행어, 속어 등 다양한 형태의 단어가 존재하는 특성을 보이고 있으므로 본 방법을 통해 더욱 정교한 토큰화 작업을 진행할 수 있다. 이후, 텍스트 길이 정제 작업을 거쳤으며 리뷰의 약 90%가 250 이하의 토큰을 포함하고 있기 때문에 모델 학습의 용이성을 위해 250 토큰을 기준으로 truncate 및 padding 작업을 진행했다. 또한 별도로 존재하는 특수문자와 같은 노이즈 데이터를 제거하는 작업을 거쳤다. 또한 토큰화 결과, 데이터의 약 90%가 길이 250 이하의 데이터로 분포되어 있으며 추후 학습의 용이성을 위해 해당 길이를 기준으로 pruning 및 padding 작업을 진행하였다. 추가로 본 연구에서 제안하고자 하는 모델 학습을 위해 train, validation, test set의 비중은 60:20:20을 기준으로 분할하여 학습과 테스트를 진행하였다.

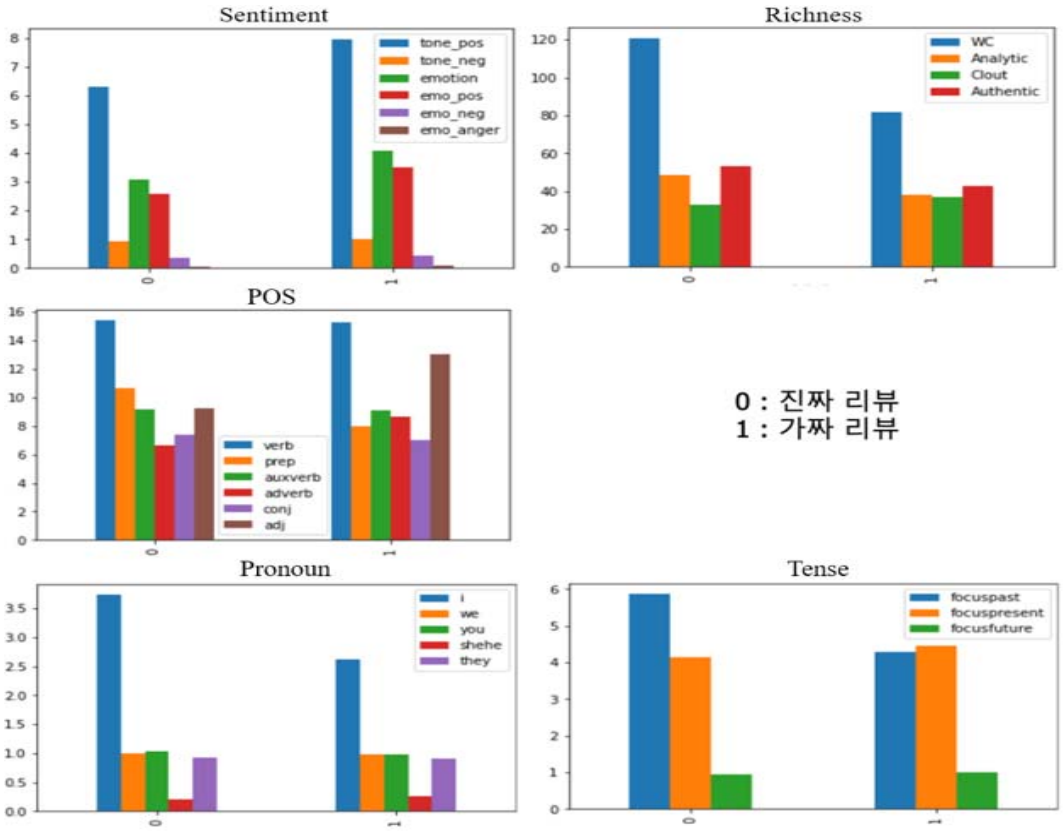
리뷰 데이터에서 작성자 언어적 특성을 대표하는 통계적 변수를 추출하기 위해 LIWC 2015 소프트웨어를 활용하였다. LIWC는 텍스트 마이닝 관련 연구에서 텍스트 데이터의 특성을 추출하기 위해 활발하게 활용되며 공식적으로 활용되고 있다(Pennebaker et al., 2015). 선행연구를 바탕으로 크게 5가지의 범주로 24개의 변수를 추출하였다. 첫 번째로 감성 범주이며 해당 범주에는 긍정, 부정 수치와 분노, 불안, 슬픔 변수로 구성되어 있다. LIWC의 긍정 관련 감정 사전은 620개의 단어로 구성되어 있으며 사랑, 달콤, 좋은 등의 대표적인 단어가 포함되어 있다. 분노 사전은 미움, 짜증 등의 대표적인 단어를 포함하여 230개의 단어로 구성되어 있다. 불안 사전에는 걱정, 두려운 등의 단어를 포함하여 116단어가 수록되어 있으며, 슬픔 사전은 울다, 슬퍼하다, 슬프다 등의 단어를 포함하여 136개의 단어가 수록되어 있다. 이를 바탕으로 감성 범주에서는 positive tone(tone_pos), negative tone(tone_neg), positive emotion (emo_pos), negative emotion(emo_neg), overall emotion(emotion), anger(emo_anger) 등 총 6개 변수를 추출하였다. 두 번째는 풍부성 범주이며, 리뷰 텍스트에 대한 총 단어 수, 분석적 사고, 신뢰성, 표현성, 진실성 변수로 구성된다. 분석적 사고, 신뢰성 그리고 진실성은 관사, 전치사 및 접속사의 사용 패턴을 연결하여 글의 분석적 특성, 자기 자신에 대한 적극적인 표현 특성 등을 수치로 추출된다. 풍부성 범주의 변수는 총 4개로 word count(WC), 분석적 사고(Analytic), 신뢰성(Clout), 진실성 (Authentic)을 포함한다. 세 번째는 형태소 범주이며 빈도를 점수로 나타낸 동사(verb), 전치사 (prep), 조동사(auxverb), 부사(adverb), 접속사 (conj), 형용사(adj) 등 총 6개의 변수를 추출하였다. 네 번째는 대명사 범주로 I, We 등의 인칭 대명사 활용 빈도를 점수로 나타낸다. 구체적으로 1인칭 단수(I)와 복수(we), 2인칭(you), 3인칭 단수(shehe), 3인칭 복수(they) 등 총 5개의 변수를 추출하였다. 마지막으로 시제 범주는

각각의 리뷰가 과거(focuspast), 현재(focuspresent), 미래(focusfuture) 시제 어휘를 활용한 정도를 점수로 나타낸다. 5개 범주에서 추출한 24개의 언어적 특성 변수들에 대한 기초 통계량은 표 2와 같다.

다음으로 해당 변수들에 대한 가짜 리뷰와 진짜 리뷰 간의 분포 비중을 분석하였으며, 변수 분포 차이는 그림 2와 같다. 감성 범주(Sentiment)에서는 가짜 리뷰의 긍정 관련 변수인 tone_pos, emo_pos 점수가 상대적으로 높게 나타나고 있다. 이는 기존 문헌에 밝힌 가짜 리뷰의 과장된 긍정 표현이 나타난 것으로 볼 수 있다(Banerjee, 2022). 풍부성 범주(Richness)에서는 글의 풍부성을 나타내는 활용 단어 수(Word Counts), 분석성(Analytic), 진실성(Authentic) 점수가 상대적으로 낮은 경향을 보여, 내용의 구체성 및 풍부성이 진짜 리뷰 대비 낮다는 특징을 파악할 수 있다. 형태소(POS) 범주에서는 형용사와 부사 빈도가 높고 전치사 빈도가 낮은 경향을 보여, 감성 범주에서와 마찬가지로 과장된 평가 위주의 내용이 가짜 리뷰에 반영되어 있으며 진짜 리뷰는 구체적인 장소 관련 내용이 더 많이 포함되어 있음을 확인할 수 있다. 대명사 범주(Pronoun)에서는 자신을 드러내는 'I' 단어 활용이 가짜 리뷰에서 낮은 경향을 보이는 점을 통해, 실제 경험을 기반으로 한 리뷰가 아닌 허구의 내용이 반영되어 있음을 간접적으로 알 수 있다. 마지막으로 시제 범주(Tense)에서는 과거형 표현 사용 빈도가 상대적으로 낮은 점을 통해, 진짜 리뷰 대

<표 2> Linguistic Features 기초 통계량

| Category | Variables | mean | std | min | max |
|---|---|---|---|---|---|
| Sentiment | tone_pos | 6.30 | 5.51 | 0.00 | 100 |
| | tone_neg | 0.92 | 1.78 | 0.00 | 100 |
| | Emotion | 3.10 | 3.89 | 0.00 | 100 |
| | emo_pos | 2.59 | 3.73 | 0.00 | 100 |
| | emo_neg | 0.37 | 1.09 | 0.00 | 100 |
| | emo_anger | 0.05 | 0.37 | 0.00 | 33.33 |
| Richness | WC | 120.51 | 108.47 | 1.00 | 5,294 |
| | Analytic | 48.65 | 26.07 | 1.00 | 99 |
| | Clout | 32.59 | 27.39 | 1.00 | 99 |
| | Authentic | 53.40 | 31.10 | 1.00 | 99 |
| POS | verb | 15.37 | 5.04 | 0.00 | 100 |
| | prep | 10.65 | 4.12 | 0.00 | 57.14 |
| | auxverb | 9.14 | 3.83 | 0.00 | 66.67 |
| | adverb | 6.61 | 3.76 | 0.00 | 66.67 |
| | conj | 7.39 | 3.15 | 0.00 | 66.67 |
| | adj | 9.22 | 5.23 | 0.00 | 100 |
| Pronoun | I | 3.73 | 2.95 | 0.00 | 100 |
| | we | 1.00 | 1.66 | 0.00 | 33.33 |
| | you | 1.03 | 1.67 | 0.00 | 40 |
| | shehe | 0.21 | 0.67 | 0.00 | 33.33 |
| | they | 0.93 | 1.44 | 0.00 | 33.33 |
| Tense | focuspast | 5.86 | 4.25 | 0.00 | 66.67 |
| | focuspresent | 4.15 | 3.35 | 0.00 | 50 |
| | focusfuture | 0.94 | 1.52 | 0.00 | 50 |




<그림 2> 진짜 리뷰와 가짜 리뷰의 linguistic features 분포 비교

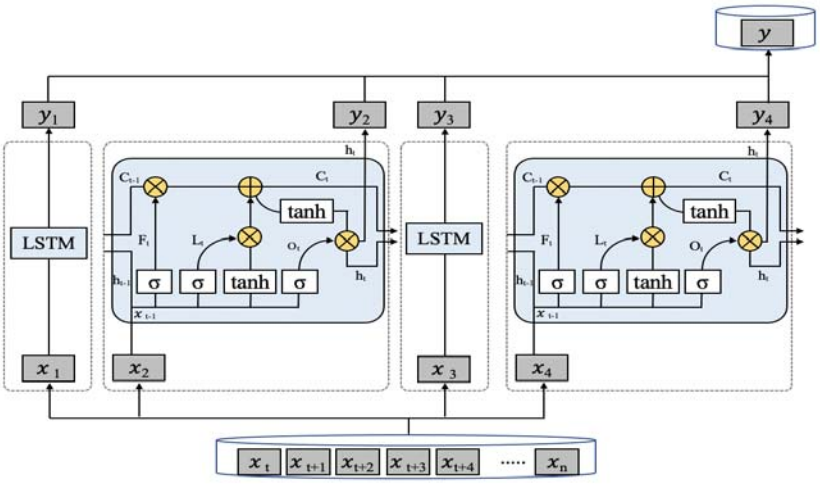
비 가짜 리뷰가 과거의 실제 경험에 의존한 내용이 아니라는 특성을 볼 수 있다.

## Ⅳ. 가짜 리뷰 탐지 모델

### 4.1 MLP / RNN(LSTM) / Transformer

다층 퍼셉트론(multi-layer perceptron, MLP)은 퍼셉트론으로 이루어진 층(layer)을 여러 겹으로 쌓은 형태로 정방향 인공신경망(feed-forward deep neural network, FDNN)이라고 부르기도 한다. 이러한 MLP 구조는 딥러닝의 가장 기본적인 구조이지만, 순차적 정보를 반영하는 데 한계점이 존재한다. 이러한 단점을 보완하기 위해 시퀀스(sequence)가 있는 데이터를 모델링할 수 있는 RNN이 개발되었다. 기존 MLP 구조와 다른 점은 기억하는 상태 정보(hidden state)를 가지고 있다는 점으로 순환 구조를 가지고 있다(Connor et al., 1994). 새로운 입력이 투입될 때마다 네트워크를 구성하는 파라미터, 즉 내부 기억을 업데이트하는 과정이 반복되고 모든 입력이 처리되면 네트워크 내부에는 전체 시퀀스의 요약 정보가 추출된다.

본 연구에서 활용되는 모델인 LSTM은 기존 RNN의 장기 의존성 문제(long term dependency)




<그림 3> LSTM 구조

를 보완하기 위해 새롭게 개발되었다(Hochreiter and Schmidhuber, 1997). 장기 의존성 문제는 여러 시퀀스를 거칠수록 새로운 정보를 학습하고 과거의 정보를 손실하는 상태를 나타내며 이는 RNN 네트워크에서 과거 데이터를 누적으로 학습하면서 발생하는 문제이다. LSTM은 이러한 한계점을 개선하여 앞선 시퀀스의 정보도 지속해서 학습할 수 있도록 개발되었다(그림3 참조).

LSTM에는 Forget Gate, Input Gate, Memory Cell, Update Gate 4개의 상호작용하는 게이트가 존재하며 이를 통해 각 시퀀스에서의 Cell State에서 버릴 과거 정보를 조절하여 필요한 과거 정보만을 보존하여 학습하게 된다. 결국 LSTM 네트워크 구조의 핵심은 4가지 게이트를 활용하여 hidden state를 기억하고 잊어버리는 과정을 거치게 된다. 이를 통해 기울기 소실(vanishing gradient) 문제를 방지하고 효과적으로 역전파 학습이 가능해지도록 한다(Sak et al., 2014).

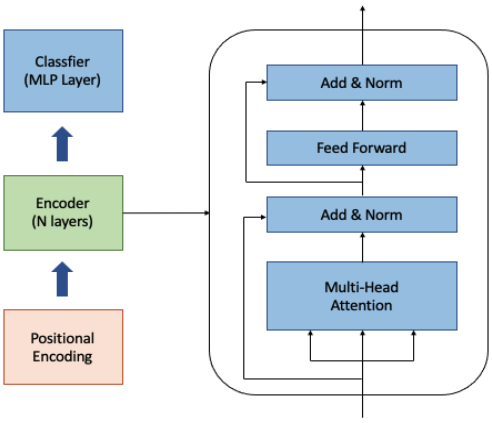
이러한 RNN 기반 모델이 발전하면서 Seq2Seq(Sequence-to-Sequence) 방식의 기계 번역과 같은 시퀀스 변환 모델이 함께 발전했다. Seq2Seq 방식은 encoder-cecoder 구조로 구성되어 있으며 encoder가 입력받은 시퀀스를 고정된 길이의 벡터로 압축하여 decoder로 전달하게 된다. 하지만 encoder에서 입력된 시퀀스 정보를 벡터로 요약하기 때문에 정보가 소실되는 문제가 발생하였고 attention이 도입되었다(Bahdanau et al., 2014). 하지만 RNN 기반의 attention 모델은 병렬 연산이 되지 않아 많은 시간이 소요된다는 단점과 긴 시퀀스에서의 종속성을 학습하기 어렵다는 한계가 여전히 존재하였다.

이러한 한계점을 보완하기 위해 RNN 구조를 모두 attention 구조로 변환한 Transformer 모델이 활용되었다(Vaswani et al., 2017). Transformer는 encoder와 decoder 각 여러 층으로 구성되어 층마다 encoder에서의 self-attention, decoder에서의 self-attention, 그

리고 encoder와 decoder 간의 cross-attention 연산이 진행된다. Transformer의 핵심인 attention 연산은 query, key, value 벡터를 활용하게 된다. 입력으로 받아 내적 연산을 통해 가중치가 계산되고 전체 맥락 중에서 특정 토큰과 연관이 있는 정보를 추출한다. 이를 통해 긴 시퀀스에서도 RNN 계열 모델보다 장거리 종속성 학습 면에서 강점을 지닌다. attention 과정은 단일 연산으로 이루어지는 것이 아닌 multi-head attention으로 구성되며, 이는 서로 상이한 위치에서 추출되는 다양한 표현 정보를 동시에 반영할 수 있도록 해준다. 이를 통해 하나의 attention 정보에 집중하는 것이 아닌, 다양한 Attention 결과를 융합적으로 반영할 수 있다는 점에서 언어 정보를 더욱 다양하게 학습할 수 있다.

MLP는 딥러닝 알고리즘의 출발점으로 다양하게 활용이 가능하지만 모델의 복잡도가 증가하면 그와 동시에 과적합 발생 가능성도 증가하는 한계점이 있다. 또한 RNN 구조는 시퀀스 정보가 반영하여 텍스트 및 시계열 정보 처리가 가능하지만, 순차적 학습을 통한 긴 학습 시간이 필요하고 긴 시퀀스 정보를 효과적으로 반영하기 어려운 한계점이 존재한다. 그리고 Transformer 구조는 self-attention을 기반으로 텍스트의 contextual 정보를 파악할 수 있어, RNN 대비 긴 시퀀스 정보 반영이 용이하다. 하지만 모델의 규모가 크기 때문에 충분한 컴퓨팅 자원과 학습시간이 필요하다는 단점 또한 존재한다.

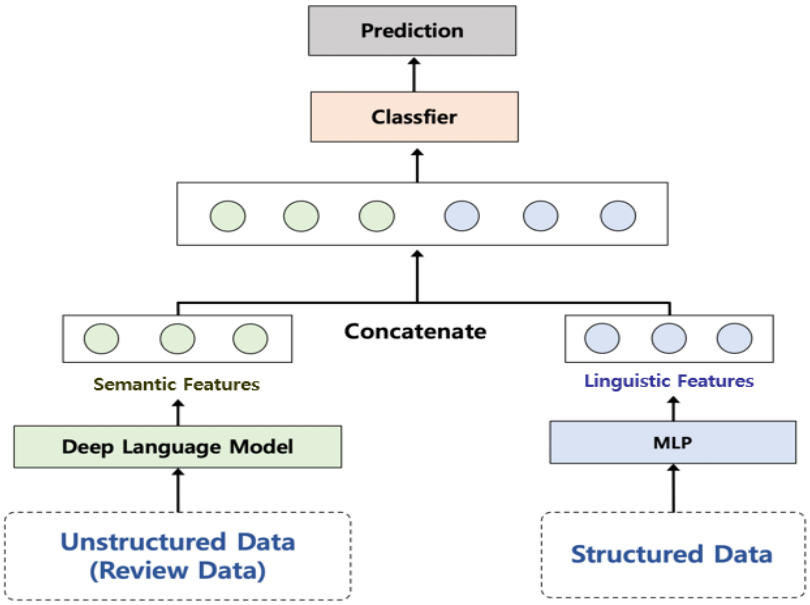
본 연구에서는 linguistic features는 MLP 모델에 입력되고 semantic features는 LSTM 또는 Transformer 모델에 입력되어 새로운 특성을 학습하고 추출하게 된다. MLP는 총 5개의 hidden layer로 구성하고 LSTM은 2개의 층(layer), Transformer는 4개의 층을 쌓은 구조로 성능 비교를 진행한다. 또한 임베딩 차원과 은닉층 차원은 동일하게 128로 고정하고 이후 최적의 성능을 보이는 혼합 모델의 파라미터 최적화가 진행된다. 또한 본 연구는 시퀀스 생성이 아닌 텍스트 분류에 해당하기 때문에, 그림 4와 같이 Transformer 구조에서 시퀀스를 생성하는 decoder 부분을 제외한 encoder 부분만을 활용하여 분류할 수 있도록 모델을 구성하였다.




<그림 4> 텍스트 분류를 위한 Transformer Encoder 구조

## 4.2 모델 구성

본 연구에서 작성자의 행동 특성을 나타내는 linguistic features와 리뷰 데이터의 의미론적 특성을 나타내는 semantic features를 결합한 딥러닝 모델을 구축하고자 한다. 이처럼 2가지 상이한 특성의 데이터를 융합하기 위해 linguistic features는 MLP 네트워크, semantic features는




<그림 5> Linguistic, Semantic Features 결합 딥러닝 모델 도식

시퀀스 모델링을 위해 LSTM, Transformer 네트워크에 입력되고 학습되어 새로운 특성을 추출하게 된다. linguistic features는 LIWC를 통해 추출한 5가지 범주에 해당하는 총 24개의 변수가 활용되며 semantic features는 토큰화 작업이 된 텍스트 데이터를 학습 모델에 맞게 학습되어 새로운 임베딩 값을 갖게 된다. 이후 추출된 특성은 단순 결합되어 가짜 리뷰를 예측하기 위해 최종 분류기 역할의 추가적인 MLP 네트워크에 입력된다(그림 5).

본 연구의 목적은 작성자의 행동 특성이 가짜 리뷰 분류에 얼마나 큰 영향력을 지니고 있는지 입증하는 것에 있으므로 이를 위해 먼저 개별 특성을 활용한 단일 모델을 구축하고 성능 비교를 진행한다. 이후, 두 가지 특성을 결합한 모델과의 추가적인 성능 비교를 진행하여, 두 특성의 결합을 통해 더 높은 성능의 분류가 가능한지 실험을 통해 입증한다. 그리고 최종적으로, 가장 좋은 성능의 결합 모델의 하이퍼 파라미터를 조정하여 학습을 진행해 최적의 성능을 지닌 가짜 리뷰 탐지 모델을 구축하고자 한다.

## Ⅴ. 결 과

각 모델을 활용하여 가짜 리뷰 탐지 성능을 평가하였다. 본 연구는 정확도(Accuracy; Accuracy $=\frac{True\ Positive(TP)+True\neg ative(TN)}{TP+TN+FalsePositive(FP)+False\neg ative(FN)}$), 정밀도(Precision; Precision=$\frac{TP}{TP+FP}$), 재현율(Recall; Recall=$\frac{TP}{TP+FN}$), F1(F1=$2\times\frac{Precisio\times Recall}{Precisio+Recall}$)을 활용하여 모델 간 성능 비교를 진행하였다. 우선적으로 linguistic

features만을 활용한 단일 모델, semantic features만을 활용한 단일 모델, 그리고 최종적으로 두 가지 특성을 결합한 결합 모델을 학습하여 테스트 결과를 비교하였다. 텍스트 데이터를 학습하는 semantic features 모델은 LSTM과 Transformer를 각각 활용하여 비교하였고 그 결과는 표 3과 같다. 성능 비교 결과, 두 가지 특성을 활용한 MLP + Transformer 모델의 정확도, 정밀도, 재현율, F1이 0.9531, 0.8715, 0.633, 0.7334로 제일 좋은 탐지 성능을 보였다. 또한 본 연구에서 제안하는 두 가지 특성을 모두 활용한 모델들(i.e., MLP+LSTM과 MLP+ Transformer)이 단일 모델보다 높은 성능을 보였으며, semantic feature를 활용한 모델에 대해서는 LSTM보다 Transformer를 적용한 모델이 성능이 더 높은 것을 확인할 수 있다.

이후 linguistic features와 semantic features를 결합한 모델의 성능 향상을 위해 하이퍼 파라미터 최적화 작업을 진행했다. 구체적인 하이퍼 파라미터 탐색 조건은 다음 표 4와 같으며, cross-entropy를 기반으로 최종 하이퍼 파라미터를 선정하였다. Optimizer, batch size, learning rate, activation function 또한 반복 실험을 통해 optimizer는 adaptive moment assesment(ADAM), batch size는 128, learning rate는 0.0001, activation function은 rectified linear unit(ReLU)로 설정하였다. 이처럼 최적의 하이퍼 파라미터를 설정하고 성능을 평가한 결과, 최적화 이전 대비 재현율은 6% 이상 향상되었고, 이에 따라 F1은 약 4%가 향상되었다(표 5).

<표 3> Fake Review 탐지 모델 성능 비교표

| Feature | Model | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|---|
| Linguistic | MLP | 0.9302 | 0.7204 | 0.5136 | 0.5997 |
| Semantic | LSTM | 0.8624 | 0.6149 | 0.5075 | 0.5561 |
| | Transformer | 0.8950 | 0.6287 | 0.5521 | 0.5879 |
| Linguistic + Semantic | MLP + LSTM | 0.9531 | 0.8715 | 0.6330 | 0.7334 |
| | MLP + Transformer | 0.9680 | 0.8802 | 0.6810 | 0.7679 |

<표 4> 하이퍼 파라미터 탐색 조건 및 모델 최종 성능

| | Hyperparameter | Search Space | Selected Value |
|---|---|---|---|
| Linguistic Modality | Number of Hidden Layer | [4,5,6] | 5 |
| | Number of Neuron | [128,256,512] | 256 |
| Semantic Modality | Network Type | [LSTM, Transformer] | Transformer |
| | Embedding Dimension | [128,256,512] | 256 |
| | Number of Hidden Layer | [2,3,4] | 3 |
| | Number of Neuron | [128,256,512] | 256 |
| Classifier | Number of Shared Layer | [1,2,3] | 2 |
| | Number of Neuron | [128,256,512] | 128 |

<표 5> 모델 최종 성능

| Feature | Model | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|---|
| Linguistic + Semantic | MLP + Transformer | 0.9762 | 0.8947 | 0.7249 | 0.8008 |

# Ⅵ. 토의 및 시사점

## 6.1 결과 토의

예측 성능 비교 결과, linguistic feature를 활용한 단일 모델이 semantic feature를 활용한 단일 모델보다 분류 성능이 높았다. 이를 통해 의미론적 특성뿐만 아니라 작성자의 언어적 특성 변수가 가짜 리뷰 판별의 성능 향상에 있어 중요하다는 것을 도출할 수 있었다. 또한, 가짜 리뷰는 의미론적인 접근만으로는 분류에 한계가 있으며, 이는 가짜 리뷰가 의미론적 특성에서는 진짜 리뷰와 유사해지고 있다는 점을 시사한다.

Semantic feature를 활용한 모델을 구체적 비교해보면, Transformer 기반 모델의 재현율이 LSTM 기반 모델 대비 더 높은 탐지 성능을 보인다는 점을 확인할 수 있다. 이를 통해 Transformer 기반의 텍스트 분류기가 기존의 LSTM 기반 모델보다 더 시퀀스 정보를 잘 반영한다는 점을 입증하는 결과로 볼 수 있다. 또한, Transformer가 attention 방식의 학습을 기반으로 장기 시퀀스에 대해 더욱 효과적으로 학습할 수 있다는 특성을 입증하기 위해, 각 단일 학습 모델을 기반으로 긴 시퀀스의 데이터 샘플을 별도로 추출하여 추가적 테스트를 진행하였고 결과는 표 6과 같다. 데이터 샘플은 token 기준으로 100 이상의 시퀀스인 데이터를 기존 테스트 셋에서 추출하였으며 기존 성능 측정 방식과 동일하게 4가지 성능 평가 지표를 활용하였다. 전체 테스트 셋에서의 실험 결과보다 Transformer의 성능이 더 명확하게 구분되는 점을 통해, 실제로 장기 시퀀스에서의 더욱 강점을 지니고 있음을 알 수 있다. 즉 self-attention을 통해 가짜 리뷰만이 지니고 있는 장기 시퀀스에서의 중요 토큰을 파악하고 학습함으로써, 모든 시퀀스 정보를 누적으로 학습하는 LSTM 대비 더 효과적인 분류를 할 수 있다는 시사점을 도출할 수 있다.

본 연구의 실험 결과를 통해 linguistic features와 semantic features를 융합한 모델들이 단일 모델 대비 성능이 크게 향상되었음을 확인하였다. 이는 작성자의 언어적 특성이 리뷰의 의미론적 특성과 결합되었을 때, 가짜 리뷰가 지닌 특성을 더욱 잘 학습하고 분류할 수 있다는 점을 유추할 수 있다. 이는 텍스트의 의미론적 시퀀스 정보와 작성자 언어적 특성 정보

<표 6> 장기 시퀀스 샘플에 대한 LSTM과 Transformer의 성능 비교

| Model | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| LSTM | 0.8724 | 0.6127 | 0.5263 | 0.5662 |
| Transformer | 0.8963 | 0.6475 | 0.5821 | 0.6131 |

가 상호간에 보충적인 정보를 새롭게 생성하여 분류에 활용되고 있음을 의미한다(Hu et al., 2019).

기존 선행 연구에서는 사용자의 언어적 행동 특성을 파악하기 위해, 어휘의 구성 분포를 통해 정보성, 가독성 등의 변수, 감정적 차원의 변수, 형태소 분석 그리고 어휘 분포 등의 통계적 수치를 활용하여 온라인 리뷰가 지닌 유의한 특성을 분석하였다(Ong et al., 2014; Peng et al., 2018; Wang et al., 2019). 하지만, 이는 리뷰의 특성을 설명하는 역할에 그쳤으며, 이러한 통계적 수치만을 활용한 탐지 모델을 구축하였을 경우는 텍스트의 의미론적 정보를 반영하지 못하여 성능적 한계가 존재해왔다. 본 연구에서의 실험 결과는 작성자의 언어적 특성 변수가 딥러닝을 기반으로 추출된 의미론적 특성과 결합되었을 때 가짜 리뷰 판별 성능이 향상되는 것을 보인다. 이는 언어적 특성이 의미론적 특성과 결합되어 표면적으로 드러나지 않은 정보를 새롭게 학습하고 반영할 수 있다는 점을 시사한다. 즉, 작성자의 언어적 특징이 단순히 가짜 리뷰의 특성에 대한 설명력을 부여한다는 장점만이 아닌, 예측 모델의 변수로도 중요하게 작용하고 있다는 것을 의미하기도 하며 본 연구에서 그 효과를 입증했다고 볼 수 있다.

## 6.2 한계점 및 향후 연구

본 연구에서는 크게 세 가지 연구 한계점이 존재한다. 먼저, 데이터의 시기적 한계점이 존재한다. 본 연구에서 활용한 데이터는 2004년부터 2015년까지의 데이터이다. 하지만 온라인 리뷰는 유행에 민감하게 반응하며 특성도 급격하게 변화한다는 점에서 실무 활용을 위해서는 최신 리뷰 데이터로 지속적으로 학습해야 한다. 따라서 향후 연구에서는 신규 데이터를 수집하고 추출된 통계적 변수들을 반영한 분류 모델을 테스트해 볼 필요가 있다.

다음으로 본 연구에서는 영어로 구성된 리뷰 데이터만을 활용하였다. 따라서 국내에서의 실무적 활용성에 대한 한계점이 존재한다. 향후 가짜 여부가 분류된 양질의 한글 기반 데이터가 확보 가능하다면 본 연구에서 제안하는 특성 결합 방법론을 적용한 새로운 연구가 가능해질 것으로 예상된다. 영어와 달리 한글은 자음과 모음이 음절을 이루고 다시 음절들이 모여 어절을 구성하는 체계를 가지고 있으므로 전처리와 언어적 특성 추출에 있어 한글 특성을 반영한 방법을 적용할 필요가 있다.

마지막으로 본 연구에서는 데이터의 한계로 인하여 작성자의 언어적 특성을 파악하기 위해 텍스트 데이터의 어휘 분포를 기반으로 추출한 linguistic features만을 활용하였다. 하지만, 리뷰 참여자의 리뷰 패턴, 인구 통계학적 변수, 리뷰 플랫폼 유입 경로 등을 활용한 접근은 부족하였다. 향후 리뷰 작성자의 식별이 가능하고 이들의 인구 특성 정보와 플랫폼 내 활동 정보가 포함된 데이터가 확보된다면 가짜 리뷰 판별 모델 개발에 새로운 특성 정보로 유용하게 활용될 수 있을 것이다.

## 6.2 학술적 시사점

본 연구의 학술적 시사점은 크게 세 가지로 정리할 수 있다. 첫째, 작성자 언어적 특성 변수를 활용하여 가짜 리뷰 판별 모델의 성능을 향

상시켰다. 이를 통해 다양한 언어적 특성이 가짜 리뷰 분류를 위한 주요 변수로 활용 가능하다는 시사점을 도출하였다. 선행연구에서 입증된 가짜 리뷰 작성자의 언어적 특성들은 가짜 리뷰의 특성을 설명하는 역할로써 활용되거나 탐지 모델 구축에 있어서 단일적으로 활용되어 성능적 한계가 존재하였다(Li et al. 2020; Ball and Elworthy, 2014). 반면, 본 연구는 기존에 연구되었던 가짜 리뷰가 지니고 있는 유의한 특성과 다양한 분류 모델 구축 방법론을 활용하여 언어적 특성과 의미론적 특성을 결합한 혼합 딥러닝 모델을 구축하였다. 선행 연구에서 제시한 가짜 리뷰는 명확하게 구분되는 작성자의 행동 양상이 존재한다는 시사점을 기반으로 언어적 특성을 대표하는 linguistic features를 추출하였고 모델에 반영하여 성능이 향상되는 결과를 도출하였다. 나아가 기존 연구는 불충분한 데이터로 인해 상이한 도메인에 언어적 특성을 일반화하기 어려웠으며 데이터 규모에 따라 유의성도 변하는 한계가 존재하였으나 (Banerjee and Chua, 2014 ; Ong et al., 2014) 본 연구는 다양한 업종을 포함하고 있는 대규모 리뷰 데이터셋을 바탕으로 포괄적인 작성자 언어적 특성 정보를 활용하여 일반화 가능한 가짜 리뷰 판별 모델을 제시하였다.

둘째, 본 연구의 판별 모델은 가짜 리뷰 자동화 탐지를 위해 여러 성질의 데이터를 종합적으로 이용하고자 하는 기본 연구로써 활용될 수 있다. 그동안 선행연구를 통해 여러 성질의 데이터를 이용하는 것의 효과는 입증되어 왔다 (Kooti et al., 2016; Hu et al., 2019). 하지만 가짜 리뷰 탐지 모델 관련 최근 연구는 임베딩 기반의 의미론적 특성만을 변수로 활용한 딥러닝 기반 연구가 주를 이루었고 추가적인 가짜 리뷰 탐지의 성능 향상에 한계점을 지니고 있었다. 이에 본 연구는 기존 사회 과학 연구에서 파악된 가짜 리뷰 작성자의 언어적 특징을 컴퓨터 공학 영역에 접목하여 자동화된 가짜 리뷰 탐지 모델 개발의 성능향상에 기여하였다(Li et al., 2020). 즉, 기존 딥러닝 모델에서 활용된 임베딩 방법론과 더불어 언어적 특성 변수를 함께 추출하여 이 두 가지 특성을 결합하였고, 각각의 특성에서는 탐지하지 못했던 가짜 리뷰의 숨겨진 특성을 학습하고 탐지할 가능성을 포착하였다. 본 연구의 특성 결합 접근 방식은 가짜 리뷰 분야에서 여러 성질의 변수들을 종합적으로 활용했다는 점에서 학술적인 시사점을 지니고 있다.

마지막으로는, 텍스트 분류 문제에서 주로 활용된 RNN 계열 모델과 더불어 최신 자연어 처리 모델들의 기반인 Transformer를 활용한 실험을 통해, 긴 시퀀스의 리뷰 데이터 분류에 더욱 적합하다는 시사점을 도출한 점을 들 수 있다. 기존 딥러닝 기반의 가짜 리뷰 탐지 모델 개발 연구는 모두 텍스트 분류 문제에서 보편적으로 활용되어왔던 RNN, CNN 기반 모델을 활용하였다(Bhargava et al., 2019; Hajek et al., 2020). 반면 최신 딥러닝 아키텍처를 가짜 리뷰 탐지에 적용한 연구가 부족하였다. 본 연구에서는 가짜 리뷰 탐지 분야에 더 진보된 자연어 처리 모델인 Transformer를 적용하고 그 효과에 대해 입증하였다는 점에서 기여점이 있다.

## 6.3 실무적 시사점

본 연구의 실무적 시사점은 크게 3가지로 볼

수 있다. 첫째, 본 연구의 가짜 리뷰 탐지 모델은 가짜 리뷰를 사전에 탐지를 가능하게 하여 서비스 업체나 리뷰 플랫폼 사용자들에게 발생하는 가짜 리뷰 피해 방지에 활용될 수 있다는 점에서 실무적 시사점을 지니고 있다. 고객의 제품 구매에 있어 온라인 후기의 영향력이 증대함에 따라 가짜 리뷰로 인한 피해 또한 급속도로 확대되고 있으며 악의적 가짜 리뷰를 경제적으로 환산할 경우 전 세계적으로 온라인 지출에 미치는 영향은 약 1,520억 달러에 달했다. 또한, 중국 업체의 비중이 약 63%에 달하는 아마존에서의 중국 국경 간 전자상거래 업계가 입는 경제적 피해는 최소 18조원이 넘을 것으로 집계되었다는 점을 통해 최근까지 지속적으로 가짜 리뷰로 인한 피해가 늘어나고 있다는 점을 알 수 있다. 이에 본 연구에서 제안한 가짜 리뷰 탐지 방법을 국내 실정에 맞게 적절히 활용한다면, 플랫폼 사용자들에게 발생하는 경제적 피해를 줄이는 효과를 기대할 수 있다. 국내 리뷰의 대부분이 한글로 구성되어 있고 한국어의 어휘 분석은 형태론적 변형 현상과 더불어 어절을 구성하고 있는 각 형태소를 인식해야 하는 특성 때문에 다른 언어에 비해 분석이 복잡한 특성이 존재한다(박현정 등, 2018). 따라서, 제시된 탐지 방법을 효과적으로 적용하기 위해서는 한글에 적합한 전처리 기법과 한글 데이터로부터 언어적 특성 추출을 위한 사전 및 방법론 개발이 선행될 필요가 있다.

둘째, 본 연구에서의 일반화할 수 있는 데이터를 통해 입증된 가짜 리뷰의 특성들을 기반으로 자동화된 모델 구축뿐만 아니라, 리뷰를 활용하는 사용자들에게 가짜 리뷰의 특성을 공지하여 사용자들의 정보 침해를 최소화할 수 있다. 리뷰를 기반으로 특성 제품이나 서비스를 탐색할 때, 리뷰를 선별적으로 활용할 수 있는 지침을 통해서 더욱 합리적인 리뷰 활용을 유도할 수 있을 것으로 예상된다.

마지막으로, 리뷰 플랫폼이나 온라인 상거래 환경에서 뿐만이 아닌, SNS와 뉴스 분야에서도 가짜 콘텐츠의 피해가 증대하고 있으며 본 연구 결과가 이와 같은 분야에서도 활용될 수 있을 것으로 기대할 수 있다. 가짜 리뷰 뿐만 아니라 뉴스 분야에서 또한 가짜 뉴스로 인한 전 세계적 피해 규모는 연간 30조 원에 달했으며 현재까지 지속적으로 확대되고 있다.[4] 본 연구에서 제시된 방법론은 가짜 뉴스만이 지닌 언어적 특성과 의미론적 특성을 결합하는 새로운 가짜 뉴스 탐지 모델을 개발하는 것에 활용될 수 있다.

## 참고문헌

강성안, 김동연, 류민호. “텍스트 마이닝을 이용한 부동산 서비스 앱 리뷰 분석,” 정보시스템연구, 제 30권, 4호, 2021, pp. 227-245.

강지우, 김동욱, 송이현, 이석범, 이범진, 정윤경, “음식점 가짜 리뷰 판별을 위한 기계학습 방법 비교,” 한국정보과학회 학술발표논문집, 2017, pp. 1980-1982.

박지현, 김종권, “국내 쇼핑 사이트 적용을 위한 리뷰 스팸 탐지 방법의 성능 평가,” 정

4) http://hri.co.kr/board/reportView.asp?firstDepth=1&secondDepth=1&numIdx=27886&isA=1

보과학회논문지, 제44권, 4호, 2017, pp. 339-343.

박현정, 송민채, 신경식. “CNN 을 적용한 한국어 상품평 감성분석: 형태소 임베딩을 중심으로.” 지능정보연구 제24권, 2호, 2018, pp. 59-83.

사공원, 하성호, 박경배, “온라인 후기에 내재된 고객의 감성분석과 LQI 차원별 호텔서비스 품질 평가,” 정보시스템연구, 제25권, 3호, 2016, pp. 217-245.

야오즈옌, 박영기, 홍태호, “리뷰어의 속성이 온라인 리뷰 유용성에 미치는 영향에 관한 연구,” 정보시스템연구, 제 29권, 2호, 2020, pp. 173-195

이민철, 윤현식, “머신러닝을 활용한 가짜리뷰 탐지 연구: 사용자 행동 분석을 중심으로,” 지식경영연구, 제21권, 3호, 2020, pp. 177-195.

이호근, 곽현, “온라인 소비자 리뷰의 효과에 영향을 미치는 요인에 대한 고찰,” 정보화정책, 제20권, 3호, 2013, pp. 3-17.

Babić Rosario, A., Sotgiu, F., De Valck, K., and Bijmolt, T. H., “The Effect of Electronic Word of Mouth on Sales: A Meta-analytic Review of Platform, Product, and Metric Factors,” *Journal of Marketing Research,* Vol. 53, No. 3, 2016, pp. 297-318.

Bahdanau, D., Cho, K., and Bengio, Y, “Neural Machine Translation by Jointly Learning to Align and Translate,” arXiv preprint arXiv:1409.0473, 2014, Available:https://doi.org/10.48550/arXiv.1409.0473.

Ball, L., and Elworthy, J., “Fake or Real? The Computational Detection of Online Deceptive Text,” *Journal of Marketing Analytics,* Vol. 2, No. 3, 2014, pp. 187-201.

Banerjee, S., and Chua, A. Y., “Understanding the Process of Writing Fake Online Reviews,” In Ninth International Conference on Digital Information Management (ICDIM), 2014, pp. 68-73.

Banerjee, S., “Exaggeration in Fake vs. Authentic Online Reviews for Luxury and Budget Hotels,” *International Journal of Information Management,* Vol. 62, 2022, 102416.

Bhargava, R., Baoni, A., and Sharma, Y., “Composite Sequential Modeling for Identifying Fake Reviews,” *Journal of Intelligent Systems,* Vol. 28, No, 3, 2019, pp. 409-422.

Connor, J. T., Martin, R. D., and Atlas, L. E.,“ Recurrent Neural Networks and Robust Time Series Prediction,” *IEEE Transactions on Neural Networks,* Vol. 5, No. 2, 1994, pp. 240-254.

Dellarocas, C., and Narayan, R., “What Motivates Consumers to Review a Product Online? A Study of the Product-Specific Antecedents of Online Movie Reviews,” *Statistical Science,* C, Vol. 21, 2006, pp. 277-285.

Dematis, I., Karapistoli, E., and Vakali, A., "Fake Review Detection via Exploitation of Spam Indicators and Reviewer Behavior Characteristics," In International Conference on Current Trends in Theory and Practice of Informatics, 2018, pp. 581-595.

Fontanarava, J., Pasi, G., and Viviani, M., "Feature Analysis for Fake Review Detection through Supervised Classification," In 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2017, pp. 658-666.

Hajek, P., Barushka, A., and Munk, M., "Fake Consumer Review Detection Using Deep Neural Networks Integrating Word Embeddings and Emotion Mining," *Neural Computing and Applications*, Vol. 32, No. 23, 2020, pp. 17259-17274.

Hochreiter, S., and Schmidhuber, J., "Long Short-term Memory," *Neural Computation,* Vol. 9, No. 8, 1997, pp. 1735-1780.

Hu, D., Wang, C., Nie, F., and Li, X., "Dense Multimodal Fusion for Hierarchically Joint Representation," IEEE International Conference on Acoustics, Speech and Signal Processing, 2019.

Kc, S., and Mukherjee, A., "On the Temporal Dynamics of Opinion Spamming: Case Studies on Yelp," In Proceedings of the 25th International Conference on World Wide Web, 2016, pp. 369-379.

Kooti, F., Lerman, K., Aiello, L. M., Grbovic, M., Djuric, N., and Radosavljevic, V., "Portrait of an Online Shopper: Understanding and Predicting Consumer Behavior," Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, 2016, pp. 205-214.

Kudo, T., and Richardson, J., "Sentencepiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing," arXiv preprint arXiv:1808.06226, 2018, Available: https://doi.org/10.48550/arXiv.1808.06226.

Li, F. H., Huang, M., Yang, Y., and Zhu, X., "Learning to Identify Review Spam," In Twenty-second International Joint Conference on Artificial Intelligence, 2011.

Li, L., Lee, K. Y., Lee, M., and Yang, S. B., "Unveiling the Cloak of Deviance: Linguistic Cues for Psychological Processes in Fake Online Reviews," *International Journal of Hospitality Management,* Vol. 87, 2020, 102468.

Moon, S., Kim, M. Y., and Iacobucci, D., "Content Analysis of Fake Consumer Reviews by Survey-based Text Categorization," *International Journal of Research in Marketing,* Vol. 38, No.

2, 2021, pp. 343-364.

Mukherjee, A., Venkataraman, V., Liu, B., and Glance, N., "Fake Review Detection: Classification and Analysis of Real and Pseudo Reviews," Technical Report, UIC-CS-03-2013, 2013.

Ong, T., Mannino, M., and Gregg, D., "Linguistic Characteristics of Shill Reviews," *Electronic Commerce Research and Applications,* Vol. 13, No. 2, 2014, pp. 69-78.

Ott, M., Choi, Y., Cardie, C., and Hancock, J. T., "Finding Deceptive Opinion Spam by Any Stretch of the Imagination," arXiv preprint arXiv:1107.4557, 2011, Available: https://doi.org/10.48550/arXiv.1107.4557.

Peng, H. G., Zhang, H. Y., and Wang, J. Q., "Cloud Decision Support Model for Selecting Hotels on TripAdvisor. com with Probabilistic Linguistic Information," *International Journal of Hospitality Management,* Vol. 68, 2018, pp. 124-138.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K., The Development and Psychometric Properties of LIWC2015, Austin, TX: University of Texas at Austin, 2015.

Rayana, S., and Akoglu, L., "Collective Opinion Spam Detection: Bridging Review Networks and Metadata," In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data ,Mining, 2015, pp. 985-994.

Ren, Y., and Ji, D., "Neural Networks for Deceptive Opinion Spam Detection: An Empirical Study," *Information Sciences,* Vol. 385, 2017, pp. 213-224.

Sak, H., Senior, A., and Beaufays, F., "Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition," arXiv preprint arXiv:1402.1128, 2014, Available: https://doi.org/10.48550/arXiv.1402.1128.

Shukla, Aishwarya, Wang, Weiguang, Gao, Guodong (Gordon), and Agarwal, Ritu, "Catch Me If You Can - Detecting Fraudulent Online Reviews of Doctors Using Deep Learning," 2019, Available at SSRN: http://dx.doi.org/10.2139/ssrn.3320258.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I., "Attention is All You Need," *Advances in Neural Information Processing Systems,* Vol. 30, 2017.

Wang, J., Kan, H., Meng, F., Mu, Q., Shi, G., and Xiao, X., "Fake Review Detection Based on Multiple Feature Fusion and Rolling Collaborative Training," *IEEE Access,* Vol. 8, 2020, pp. 182625-182639.

Wang, X., Tang, L. R., and Kim, E., "More than words: Do Emotional Content and Linguistic Style Matching Matter on Restaurant Review Helpfulness?," *International Journal of Hospitality Management,* Vol. 77, 2019, pp. 438-447.

Yoo, K. H., and Gretzel, U., "Comparison of Deceptive and Truthful Travel Reviews," *In Information and Communication Technologies in Tourism,* 2009, pp. 37-47.

**신 동 훈(Shin, Dong Hoon)**

연세대학교 정보대학원에서 비즈니스 빅데이터 분석을 전공하고, 정보시스템학 석사 학위를 취득하였다. 현재 현대카드에서 데이터 사이언티스트로 재직 중이다. 주요 연구 분야는 빅데이터분석, 머신러닝, 딥러닝 등이다.

**신 우 식 (Shin, Woo Sik)**

연세대학교 정보대학원에서 비즈니스 빅데이터 분석 트랙 박사과정이다. 현재 디지털서비스 연구실에서 빅데이터 분석 관련 연구를 진행 중이다. 주요 연구분야는 Digital Marketing, Cyberbullying, Business Analytics 등이다.

**김 희 웅 (Kim, Hee Woong)**

National University of Singapore 정보시스템학과에서 근무한 후, 현재 연세대학교 정보대학원 교수로 재직 중이다. 주요 연구분야는 디지털 비즈니스, 정보시스템 관리 및 활용 등이다. 관련 연구들은 MIS Quarterly, Information Systems Research, Journal of Management Information Systems 등에 60여 편의 논문이 게재되었다. MIS Quarterly의 편집위원으로 활동하고 있다.

<Abstract>

# Development of a Deep Learning Model for Detecting Fake Reviews Using Author Linguistic Features

Shin, Dong Hoon · Shin, Woo Sik · Kim, Hee Woong

**Purpose**

This study aims to propose a deep learning-based fake review detection model by combining authors' linguistic features and semantic information of reviews.

**Design/methodology/approach**

This study used 358,071 review data of Yelp to develop fake review detection model. We employed linguistic inquiry and word count (LIWC) to extract 24 linguistic features of authors. Then we used deep learning architectures such as multilayer perceptron(MLP), long short-term memory(LSTM) and transformer to learn linguistic features and semantic features for fake review detection.

**Findings**

The results of our study show that detection models using both linguistic and semantic features outperformed other models using single type of features. In addition, this study confirmed that differences in linguistic features between fake reviewer and authentic reviewer are significant. That is, we found that linguistic features complement semantic information of reviews and further enhance predictive power of fake detection model.

***Keyword***: Fake Review, Predictive Modeling, Deep Learning, Linguistic Feature, Semantic Feature

* 이 논문은 2022년 7월 29일 접수, 2022년 8월 26일 1차 심사, 2022년 11월 11일 게재 확정되었습니다.