JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 24, No. 6, pp. 1361-1371, Jun. 2023

Check for updates

계층적 KoBERT를 활용한 SNS 문맥 기반 이모티콘 추천

김 지 현[1] · 김 예 림[1] · 변 혜 원[2*]
[1]성신여자대학교 대학원 미래융합기술공학과 석사과정 [2*]성신여자대학교 AI융합학부 교수

SNS Context-based Emoji Recommendation Using Hierarchical KoBERT

Jee-Hyun Kim[1] · Ye-Rim Kim[1] · Hae-Won Byun[2*]

[1]Master's Course, Department of Convergence Technology Engineering, Sungshin Women's University, Seoul 02844, Korea
[2*]Professor, School of AI Convergence, Sungshin Women's University, Seoul 02844, Korea

[요 약]

이모티콘 추천은 수천 개의 이모티콘들 중에서 사용자가 원하는 적절한 이모티콘을 용이하게 찾도록 도와주는 중요한 태스크이다. 기존의 이모티콘 추천 방법들은 채팅 플랫폼을 대상으로 하며 사용자들이 많이 사용하는 감정 이모티콘 위주로 추천한다. 그러나 인스타그램 등 SNS 플랫폼에서는 감정 전달보다는 업로드한 짧은 게시글의 내용을 보완하거나 강조하는 용도로 이모티콘을 사용하는 경향이 있다. 이 연구에서는 SNS 플랫폼에서 한국어 게시글의 문맥을 파악하여 이모티콘을 추천하는 새로운 방법론을 제안한다. 이모티콘 추천 문제에 계층적 KoBERT를 도입하여 한국어 게시글의 문맥을 파악하고 이에 적합한 다양한 이모티콘을 추천한다. 314개 이모티콘 카테고리에 속하는 616개의 이모티콘 추천은 SNS 게시글의 함축적인 단문을 보다 정확하게 전달하는데 유용하다. 인스타그램 게시글을 수집하여 실제 세계를 반영하는 데이터셋을 구성하고 각 텍스트에 삽입되어 있는 이모티콘의 계층적 카테고리를 학습하기 위해 계층적 KoBERT 모델을 구축한다. 실험 결과에서 DNN, LSTM, Bi-LSTM, GRU 모델과 비교하여 계층적 KoBERT 모델이 이모티콘 추천에서 높은 성능을 보이는 것을 검증하였다.

[Abstract]

Emoji recommendation is an important task that assists users in finding appropriate emojis from thousands of candidates. Existing methods primarily focus on popular emojis related to user emotions in chat platforms. However, on SNS platforms, such as Instagram, emojis are often used to complement or emphasize the content of short uploaded posts rather than conveying emotions. This paper proposes a method for recommending emojis in Korean language posts on SNS platforms by understanding the context of the posts. We apply a hierarchical KoBERT model to capture the context of Korean posts and recommend a diverse range of emojis suitable for the content. We considered 616 emojis from 314 emoji categories for accurately conveying the context of SNS posts. We constructed the real-world dataset by collecting Instagram posts and developed the hierarchical KoBERT model to learn the hierarchical categories of emojis embedded within the texts. Experimental results validate the superior performance of the hierarchical KoBERT model in emoticon recommendation compared to DNN, LSTM, Bi-LSTM, and GRU models.

색인어 : 이모티콘 추천, 계층적 코벌트, 자연어 처리 모델, 텍스트 분류
Keyword : Emoticon Recommendation, Hierarchical KoBERT, Natural Language Processing Model, Text Classification

http://dx.doi.org/10.9728/dcs.2023.24.6.1361

Received 04 May 2023; **Revised** 16 May 2023
Accepted 19 May 2023

***Corresponding Author; Hae-Won Byun**
Tel: +82-2-920-7615
E-mail: hyewon@sungshin.ac.kr

Ⅰ. 서 론

모바일 기기의 발전으로 시간과 장소의 제약 없이 자신의 생각과 일상을 공유할 수 있는 SNS의 사용량이 지속적으로 증가하고 있다. SNS는 마이크로 블로그 형태의 서비스로서 단문과 몇 개의 이미지들을 사용하여 짧은 내용의 콘텐츠를 구성하여 업로드하는 형태이다. SNS의 대표적인 예시로는 인스타그램, 트위터 등이 존재한다. 카카오톡 등 채팅 서비스에서 주로 사용되어 온 이모티콘은 SNS 플랫폼에서도 활발하게 사용되고 있다. 2022년 9월에 출시된 유니코드 이모티콘 버전 15.0에서는 4,526개의 이모티콘을 제공하고 있다. 수 천개의 이모티콘은 표현의 다양성을 제공하는 장점이 있는 반면에, 텍스트에 적합한 이모티콘을 검색하고 선택하는 데 오랜 시간이 걸린다는 불편함이 있다.

SNS 플랫폼에서 이모티콘을 사용하는 방식은 채팅 플랫폼에서 이모티콘을 사용하는 양상과는 차이를 보이고 있다. 채팅에서는 주로 텍스트를 전달하는 사람의 희로애락 등의 감정을 표현하고자 이모티콘을 사용하는 반면에, 인스타그램 등의 SNS에서는 게시글의 내용을 보완 또는 강조하거나 시각적인 효과를 부가적으로 표현하고자 하는 데 그 보편적 사용의도가 있다[1],[2]. 예를 들어, 그림 1을 보면, 피자, 케이크, 야구 등의 단어를 시각적으로 강조하기 위하여 피자 모양, 케이크 모양의 이모티콘을 사용하고 있으며, '영롱하다'의 느낌을 보다 정확하게 전달하기 위하여 이를 시각화한 이모티콘을 사용하고 있다. 반면에, 채팅 서비스에서는 피자나 케이크 이모티콘 보다는 음식을 먹는 행복한 감정을 담은 표정 이모티콘을 전달하는 경향이 있다. 기존 연구는 대부분 채팅에서의 감정 이모티콘 추천에 집중되어 있다. 감정 기반으로 텍스트를 분류하고 30~100 개 정도의 이모티콘을 추천하므로 SNS 게시글의 다양한 내용과 문맥을 표현하는 데에는 한계가 있다[3]-[7].

SNS 플랫폼에서의 이모티콘 추천은 사용자가 작성한 게시글의 문맥을 파악하는 것이 중요하다. 이 연구에서는 인스타그램 SNS 플랫폼에서 문맥 기반으로 이모티콘을 추천하는 새로운 시스템을 제안한다. 자연어 처리 분야에서 성능이 뛰어난 사전학습 모델 BERT를 도입하여 문맥 인지 이모티콘 추천 문제에 적용한다. 변화의 범위가 매우 큰 사용자 게시글의 문맥에 적합한 이모티콘을 다양하게 추천하기 위하여 이모티콘 추천 문제를 314개의 멀티 클래스를 가지는 한국어 텍스트 분류 문제로 정의하고 성능 향상을 극대화하기 위하여 계층적 KoBERT 모델을 구축한다. 616개의 이모티콘 데이터에 계층적 군집화를 적용하여 메인 카테고리와 서브 카테고리로 분류하고 텍스트 데이터의 라벨로 사용함으로써 대량의 텍스트 데이터를 자동 라벨링하는 장점을 가진다. 또한, DNN, LSTM, Bi-LSTM, GRU 등의 모델들과 성능을 비교 분석하여 본 연구에서 구축한 계층적 KoBERT 모델의 성능이 우수함을 보여 준다.

이 연구에서 기여하는 점은 다음과 같다.

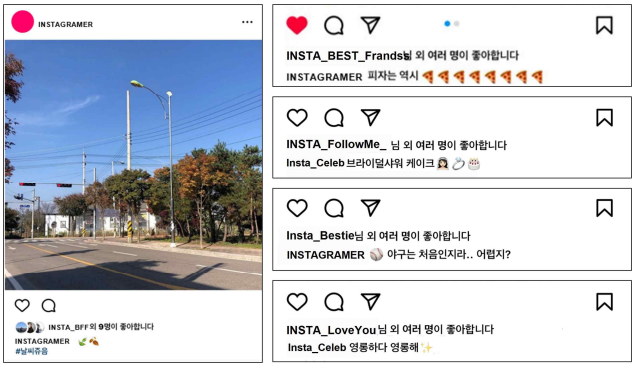
첫째, 이전의 연구들은 채팅창에서의 감정 기반 이모티콘 추천에 집중한 반면 본 연구에서는 SNS 플랫폼에서의 문맥 인지 이모티콘 추천 방법을 새롭게 제안한다.

둘째, 기존 연구들은 사용자들이 주로 사용하는 30~100 개 정도의 이모티콘을 추천하고 있으나, 본 연구에서는 314 여 개의 다양한 이모티콘 카테고리를 추천하여 문맥 인지 이모티콘 추천을 시도한다.

셋째, 계층적 KoBERT 기반으로 분류기를 생성하여 314 개의 멀티 클래스로 SNS 게시글을 분류하는 문제를 해결하고 이모티콘 추천 성능을 향상시켰다.

넷째, 이모티콘 계층적 군집화를 통해 학습용 대규모 텍스트 데이터를 자동 라벨링하는 방법을 제시한다.

본 논문의 구성은 2장에서 관련 연구들을 소개하고, 3장에서 시스템 구조, 4장에서 데이터 처리, 그리고 5장에서 모델 학습에 관하여 설명한다. 6장에서 실험 결과와 분석 내용을 제시하고, 마지막으로 7장에서 결론으로 마무리한다.



*This research is about Korean text processing.

그림 1. 인스타그램 이모티콘 사용 예시

Fig. 1. Instagram emoticon example

Ⅱ. 관련연구

2-1 이모티콘 추천

최근에 이모티콘의 사용행태를 분석하고 이모티콘을 추천하는 연구들이 몇 년간 진행되어왔다. 국내에서 한국어 데이터셋에 대한 이모티콘 추천은 주로 카카오 채팅 플랫폼에서 텍스트의 감정을 분석하여 해당 감정 카테고리에 속하는 이모티콘을 추천하는 방식이다. [5]는 카카오톡 채팅 대화문에서 희로애락 등 7가지 감정 유형을 추출하기 위하여 LSTM을 도입하고 이를 기반으로 이모티콘을 추천하였다. 영어 데이터셋에 대한 이모티콘 추천에 관한 연구는 주로 트위터 데이터셋을 사용하고 딥러닝 모델을 도입하여 감성 기반으로 텍스트를 분류하는 연구가 진행되었다[3],[6]-[10]. 대부분의 연구들에서 CNN과 LSTM 계열의 학습 네트워크를 도입하여 감정 기반으로 텍스트를 분류하고 해당 감정에 속하는 이모티콘을 추천한다. 채팅 플랫폼에서의 이모티콘 추천은 두

사람 간의 주고받은 여러 개의 대화문 히스토리를 대상으로 LSTM 모델을 사용하여 문맥을 파악하는 시도로 발전되었다[11]-[13]. 특히, [11]은 다자간의 대화문을 대상으로 연구를 확장하였다.

대화문 분석을 위해서 사전학습 모델을 도입한 방법들도 시도되었다. [14]는 LSTM에 사전학습 개념을 추가한 DeepMoji 모델을 제안하고 감정 검출과 냉소적인 글의 검출을 통해 64개의 공통 이모티콘을 추천하였다. [15]는 사전학습 기반 자연어 처리 모델인 BERT를 적용하여 영어와 일본어 등 다국어를 대상으로 이모티콘을 추천하는 연구를 발표하였다. [4]는 사용자의 트윗 히스토리에서 시간에 따라 변화하는 사용자의 동적인 선호도를 학습하는 BERT 모델의 설계를 통해 개인화된 이모티콘 추천 방법을 제안하였다. 이후의 이모티콘 추천 연구는 트위터 텍스트뿐만 아니라 시각적인 부가 정보 등을 추가로 사용하는 멀티모달 접근법을 시도하였다. [16]은 트위터의 텍스트, 업로드한 이미지 및 사용자 위치정보를 학습하는 mmGRU 모델을 제안하고 이모티콘 추천과 텍스트 내 이모티콘 위치도 함께 추천하였다. [17]은 텍스트의 문맥과 함께 사용자 선호도, 성별 및 텍스트 입력 시간 등 개인 성향을 함께 사용하는 행렬 분해 기법(matrix factorization method)을 제안하고 개인화된 이모티콘을 추천하였다.

2-2 BERT 모델

최근에 BERT를 기반으로 한 자연어 처리 연구가 다양하게 진행되고 있다. 영어 데이터셋을 대상으로 감정을 분석하는 연구 분야에서는 BERT 모델을 도입하여 성능을 향상시킨 실험 결과를 보고하고 있다[18],[19]. BERT 모델을 발전시킨 BERT 변형 모델들도 다양하게 제안되고 있다. [18]의 연구에서는 대상 정보를 추가한 Target-Dependent BERT (TD-BERT) 모델을 제안하였고 [19]의 연구에서는 지속적 학습 기능을 추가한 BERT-based Continual Learning (B-CL) 모델을 발표하였다. 이외에도, 중국어, 아랍어 등 다국어에 대한 텍스트 분류 연구가 진행되었다[20],[21]. 국내에서는 한국어에 특화된 KoBERT 모델을 활용한 연구가 주목받고 있다. 감정 분석을 비롯한 다양한 텍스트 분류 문제에 KoBERT를 적용한 연구들에서 KoBERT 모델의 우수성을 입증하고 있다[22],[23].

또한, 계층적 구조를 갖는 BERT에 대한 연구도 활발히 진행되고 있다. BERT 모델이 가지는 입력 시퀀스 길이 제한 문제를 해결하기 위하여 계층적 KoBERT를 도입하여 1단계에서 입력 텍스트를 세그먼트로 나누어 2단계로 전달한다[24]. 계층적 BERT는 인터넷 게시글 등에서 빈정대거나 비판하는 부정적인 댓글을 검출하는 응용에도 적용되었다[14]. 이외에도 토론 플랫폼 게시글이나 대화문과 같이 문맥 정보가 중요한 응용에서 Hierarchical BERT을 활용한 연구들이 발표되었다[25].

Ⅲ. 계층적 이모티콘 카테고리

수 천개의 다양하고 방대한 이모티콘은 일반적으로 그림 2와 같이 트리 형태의 계층적 이모티콘 카테고리 형태로 제공되고 있다. 유니코드 이모티콘의 경우, 이모티콘 카테고리는 메인 카테고리와 서브 카테고리로 구성된다. 예를 들어, ‘Smileys & Emoticon’ 이모티콘 메인 카테고리 하위에 ‘face-smiling’, ‘face-negative’, ‘face-unwell’ 등의 이모티콘 서브 카테고리가 있는 형태이다.

그림 2. 유니코드 이모티콘 메인 카테고리와 서브 카테고리
Fig. 2. Unicode emoticon main category and subcategory

유니코드 이모티콘의 서브 카테고리에 착안하여 이 연구에서는 32개의 메인 카테고리 하위에 314개의 서브 카테고리로 구성되는 계층적 카테고리를 그림 3과 같이 생성하였다.

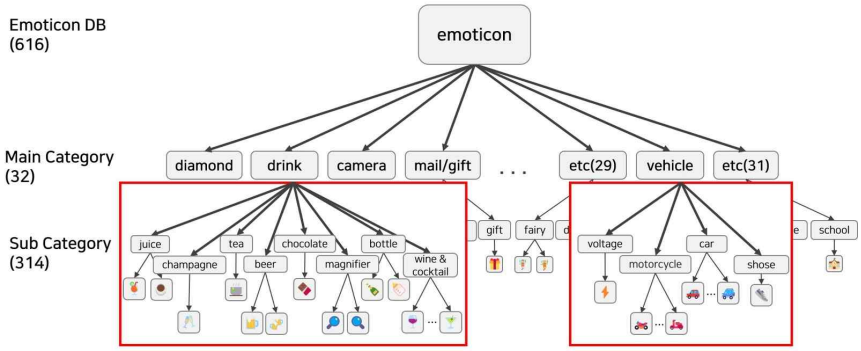


그림 3. 이모티콘 메인 카테고리와 서브 카테고리
Fig. 3. Emoticon main category and subcategory

인스타그램 게시글 119,148개를 수집하여 게시글에 포함되어 있는 이모티콘 데이터를 분석하고 의미있게 사용되고 있는 이모티콘 616개를 추출하였다. 계층적 클러스터링을 통해 이모티콘 카테고리를 2단계로 구성하였다. 상위 카테고리는 음식, 사람, 동물, 꽃 등의 큰 분류를 포함하며, 이 중 음식 카테고리 하위에 있는 서브 카테고리는 피자, 떡볶이, 컵 케익 등을 포함한다.

계층적 이모티콘 카테고리는 다음과 같이 표기한다.

$$C=\{c_{i,j}\ i=1\sim 31,\ j=1\sim \text{각 카테고리의 서브 카테고리 개수}\} \quad (1)$$

Ⅳ. 시스템 구조

이 연구에서 제안하는 이모티콘 추천 시스템은 크게 (1) 데이터 수집 및 전처리, (2) 컨텍스트 인지 이모티콘 클러스터링 (3) 계층적 KoBERT 모델 학습, (4) 추천 시스템으로 구성되며 전체적인 시스템 구조는 그림 4와 같다.

첫 번째, 데이터 수집 및 전처리 단계는 잡음 제거, 이모티콘/텍스트 분할, 이모티콘-텍스트 변환, 그리고 이모티콘-텍스트 유사도 검사의 4개의 부분으로 구성된다. 잡음 제거 부분에서는 인스타그램 게시글을 수집한 후 노이즈와 중복 문장을 제거한다. 이모티콘-텍스트 변환 부분에서는 문장으로부터 텍스트와 이모티콘을 분할하고, 이모티콘-텍스트 변환 부분에서는 이모티콘 사전을 이용하여 이모티콘을 텍스트로 변환한다. 이후 이모티콘 유사도 검사 부분에서 여러 개의 이모티콘들 중 텍스트와 유사도가 가장 높은 이모티콘 한 개를 선정한다.

두 번째, 컨텍스트 인지 이모티콘 클러스터링 단계에서는 이모티콘 벡터를 대상으로 계층적 K-means 클러스터링을 적용한다. 거리 기반으로 그룹 간 비유사도를 최소화하는 방식으로서 그룹 내에 컨텍스트가 유사한 텍스트들이 모이는 효과가 있다. 이 과정은 자율 학습으로서 레이블이 없는 이모티콘 데이터에 계층적 레이블(메인 카테고리, 서브 카테고리)을 자동으로 달아주는 역할을 수행한다.

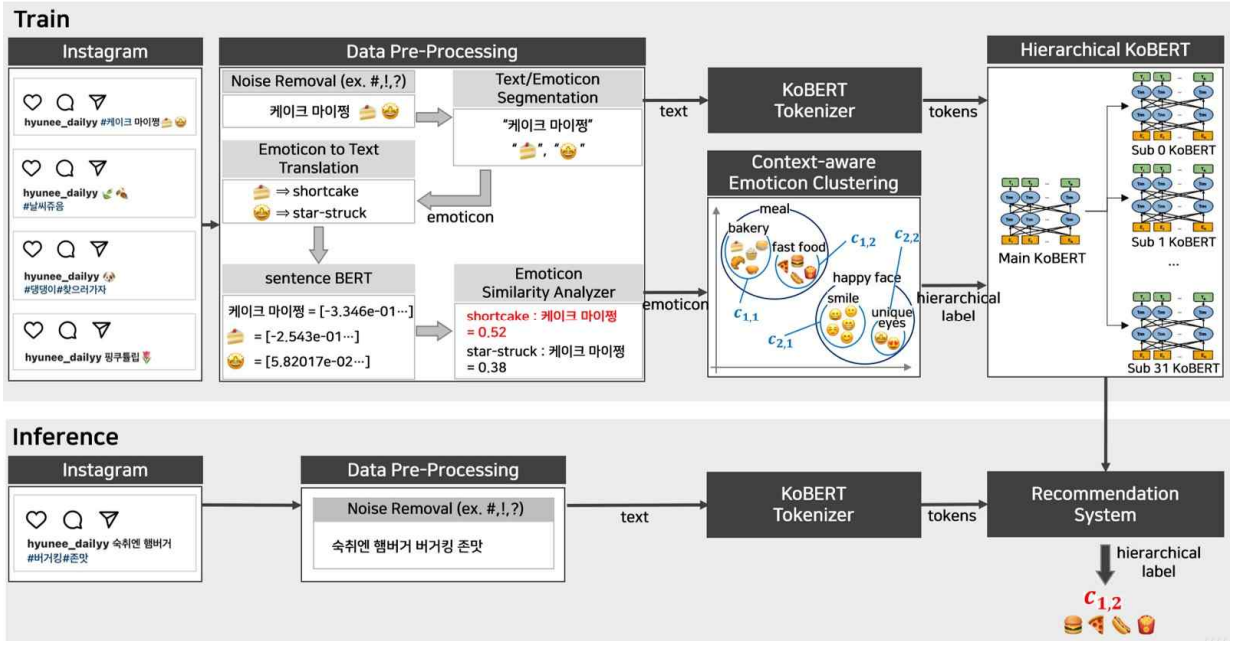
세 번째, 계층적 KoBERT 모델 학습 단계에서는 (텍스트, 이모티콘 레이블) 데이터셋을 사용하여 텍스트를 이모티콘 레이블로 분류하는 학습을 진행한다. 전 단계에서 생성한 계층적 이모티콘 레이블을 학습하기 위하여 계층적 KoBERT 모델을 구축하였다.

네 번째, 이모티콘 추천 단계에서 SNS 게시글에 대해 계층적 이모티콘 카테고리를 추론하여 해당 서브 카테고리에 있는 이모티콘들을 추천한다.

Ⅴ. 데이터 처리

5-1 데이터 수집 및 전처리

계층적 KoBERT 모델을 학습하기 위하여 다양한 이모티콘을 사용하는 대량의 텍스트 데이터가 필요하다. 이를 위하여 파이썬 기반의 크롤러를 사용하여 #일상 또는 #일기 등의 해시태그를 키워드로 인스타그램 게시물 119,148개를 수집하였다. 수집된 게시물에서 이모티콘을 포함한 문장들로 데이터셋을 구축한 후, 데이터 임베딩을 위한 전처리 과정으로서 이



*This research is about Korean text processing.

그림 4. 시스템 구성도 (** 한국어 텍스트 처리에 관한 연구이므로 한국어 데이터 표기)

Fig. 4. System overview

모티콘을 제외한 특수기호를 제거하고 다수의 문장은 줄 바꿈을 기준으로 분리하였다. 각 문장에서 이모티콘을 분리하여 (텍스트, 이모티콘)' 쌍으로 구성된 데이터셋을 구축하였다.

이 때, 하나의 문장이 2개 이상의 이모티콘을 포함할 수 있다. 본 연구에서는 문장과 이모티콘 간 코사인 유사도를 계산하여 유사도가 가장 높은 한 개의 이모티콘만 선택한다. 표 1의 이모티콘 사전을 사용하여 이모티콘을 텍스트로 변환하고, Sentence BERT를 구현한 문장 트랜스포머(Sentence Transformer) 패키지를 사용하여 이모티콘 변환 텍스트와 문장을 각각 벡터화한다. 이모티콘 벡터와 문장 벡터 간의 유사도를 계산하여 한 개의 이모티콘만 남기고 나머지 이모티콘들은 삭제한다.

전처리 과정을 거쳐 구축한 데이터는 총 172,149건으로 학습 데이터 120,489건, 검증 데이터 34,338건, 평가 데이터 17,322건으로 대략 7:2:1로 나누어 학습과 성능 평가를 실시하였다.

표 1. 이모티콘 사전
Table 1. Emoticon dictionary

Emoticon	Emoticon name
	crying cat
	rolling on the floor laughing
	couple with heart
	cloud with snow
	woman dancing
	spouting whale

5-2 컨텍스트 인지 이모티콘 계층적 군집화

데이터 전처리 과정에서 이모티콘은 이모티콘 사전을 이용하여 텍스트로 변환되고 Sentence BERT를 도입하여 이모티콘 벡터로 변환되었다. 이모티콘 계층적 군집화 단계에서는 616개의 이모티콘 벡터를 대상으로 컨텍스트가 유사한 이모티콘들을 계층적으로 클러스터링하여 메인 그룹과 각 메인 그룹에 속하는 서브 그룹을 생성한다. Sentence BERT를 이용하여 추출한 이모티콘 벡터를 대상으로 클러스터링을 수행함으로써 컨텍스트가 유사한 이모티콘들끼리 군집화하는 역할을 한다.

이모티콘 벡터 데이터셋 x1, x2, ..., x616에 대해서 탑다운 방식의 계층적 K-means 군집화를 적용한다. 1단계 K-means 군집화를 수행하여 메인 그룹 32개(Gi, i=1,2,..,32)를 생성하고, 2단계에서 각 Gi에 속한 데이터들에 대해서 재귀적으로 K-means 군집화를 수행하여 서브 그룹 314개 (Gij, i=1,2,....,31, j=1,2,...,# of subclusters)를 생성한다. 이모티콘 계층적 군집화 결과물인 Gij는 텍스트 데이터를 자동으로 라벨링하는데 사용된다.

K-means 클러스터링 과정에서 임계치는 이모티콘 군집화 결과의 품질에 큰 영향을 미친다. K값이 커지면 상이한 컨텍스트를 가지는 이모티콘들이 동일한 그룹으로 분류되고, K값이 작아지면 과다한 상세 분류가 되어 계층적 KoBERT 학습 성능을 저하시키는 결과를 초래한다. 본 연구에서는 인공지능 전공 석사 과정에 있는 2명의 연구원이 참여하여 그룹화 결과 각 그룹에 있는 데이터들이 유사한 컨텍스트를 가지는지 판단하는 역할을 수행하였다. 연구원들은 주어진 데이터셋

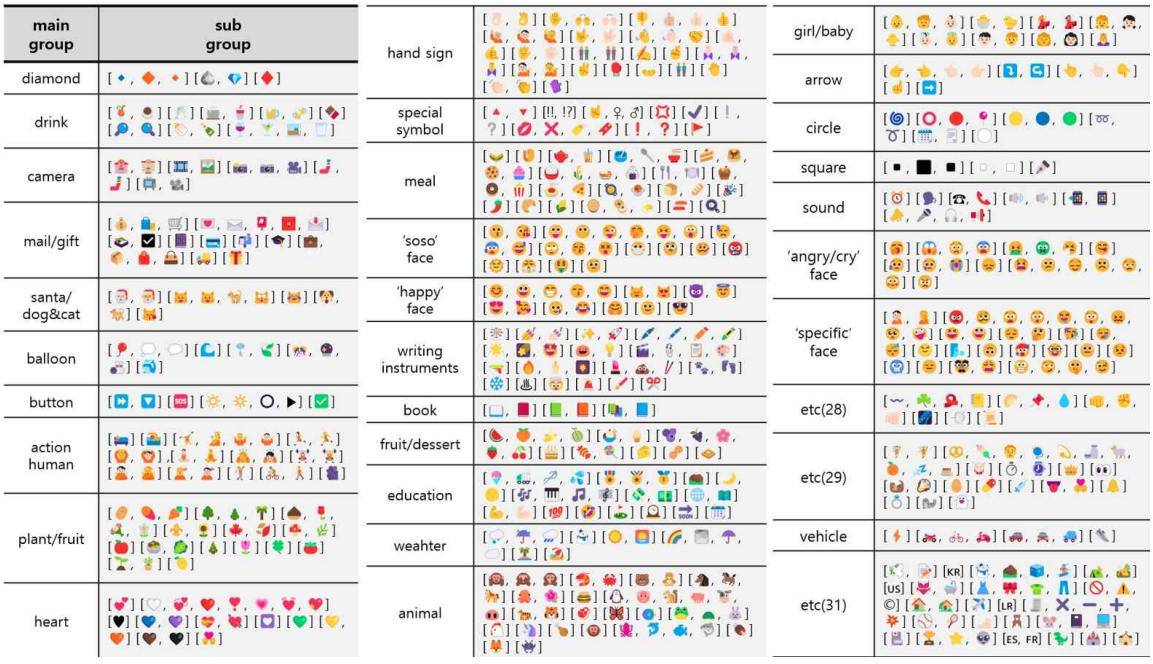


그림 5. 이모티콘 메인 카테고리와 서브 카테고리
Fig. 5. Emoticon main category and sub category

에 다양한 임계치를 적용한 결과 그룹의 적절성을 평가하고 논의를 통해 군집화 품질을 최적화하는 임계치를 경험적으로 도출하였다. 향후 연구에서 임계치를 자동으로 결정하는 최적화 알고리즘이 필요하다.

그림 5는 이모티콘 계층적 군집화 결과를 보여준다. 1단계 32개의 메인 카테고리와 2단계 314개의 서브 카테고리로 구성된 트리 구조의 군집화 결과이다.

5-3 자동 데이터 라벨링

전처리 과정을 통해(텍스트, 이모티콘) 데이터셋을 구축하고 계층적 이모티콘 카테고리를 생성한다. 모든 이모티콘은 계층적 이모티콘 카테고리에 속해 있으므로 모든 게시글은 (텍스트, 계층적 이모티콘 카테고리)로 자동 레이블링 가능하다. 여기에서 계층적 이모티콘 카테고리는 메인 카테고리 번호와 서브 카테고리 번호로 구성되며, $C=\{c_{0,0}, c_{0,1},\cdots, c_{0,k}, c_{1,1}, c_{1,2}, \ldots, c_{m,n}\}$으로 표현 가능하다. 이때 m은 메인 카테고리 개수 32개를 의미하고, n은 1단계에 있는 메인 카테고리 하위에 있는 서브 카테고리의 총 개수이다. 이와 같은 표기를 이용하여 172,149개의 모든 문장에 대해 해당 이모티콘 카테고리(메인 카테고리 번호, 서브 카테고리 번호)로 레이블을 자동 생성한다(표 2 참조). 여기서 생성한 레이블은 계층적 KoBERT 네트워크의 1단계에서 메인 카테고리 번호를 학습하고 2단계에서 서브 카테고리 번호를 학습하는 데 사용된다.

표 2. 실험 데이터 구성 예시
Table 2. Text data for experiment

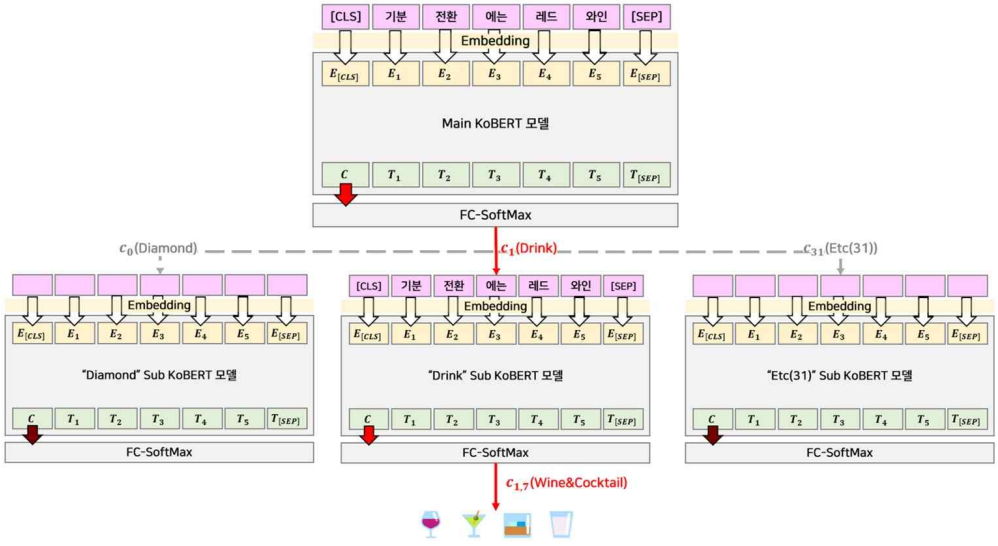
Main cat.	Sub cat.	Text Data	Label
Drink	Juice	시원 한 에이드 한잔 서비스	$c_{1,1}$
		제주 여행 펍	$c_{1,1}$
		레인보우는 내가 꼭 먹겠다고 벼르고 있었지	$c_{1,1}$
	Champaign	월요병 없는 월요일을 위하여	$c_{1,2}$
		생일축하해 유니	$c_{1,2}$
		안 놀아줘서 인생 첫 혼술 해본다	$c_{1,2}$

*This research is about Korean text processing.

Ⅵ. 계층적 KoBERT 모델

이 연구에서는 BERT 모델의 한국어 버전인 KoBERT 모델을 활용한다. BERT(Bidirectional Encoder Representations from Transformers)는 트랜스포머 모델의 인코더만을 활용하여 사전 학습된 모델로서 다양한 자연어 처리 태스크에서 뛰어난 성능을 보여주고 있다. KoBERT 모델의 양방향 인코딩 기능은 문장 내 앞뒤 문맥 정보를 모두 고려하여 한국어 텍스트의 중의성 해소와 문맥 이해력을 향상시키는 장점이 있다. 따라서, 다양한 표현과 감정을 함축적으로 포함한 SNS 단문을 파악하는 데 유리하다.

이 연구에서는 그림 6에서 볼 수 있듯이, 계층적 이모티콘 카테고리를 학습하기 위해 메인 모델과 서브 모델로 구성된 계층적 KoBERT 모델을 구축하였다. 입력 데이터가 제공되면, 메인 KoBERT에서 KoBERT Tokenizer를 통해 데이터를 토큰화하고 임베딩 층을 거쳐 문맥 정보를 추출한다. 이 단계에서 얻어진 문맥 정보는 완전 연결 신경망에 입력되어



*This research is about Korean text processing.
그림 6. 이모티콘 추천을 위한 계층적 KoBERT 모델
Fig. 6. Hierarchical KoBERT model for emoticon recommendation

메인 카테고리를 결정한다. 메인 KoBERT 모델을 실행하여 결정한 메인 카테고리에 따라 서브 KoBERT 모델을 선택하고, 해당 서브 KoBERT 모델은 서브 카테고리 중 하나를 결정하는 분류 작업을 수행한다. 이와 같은 계층적 분류 과정을 거치면서, 모델은 다양한 SNS 문맥에 가장 적절한 이모티콘 카테고리를 선택하는 능력을 학습하게 된다. 계층적 접근방식은 모델의 정확도를 향상시키는 데 중요한 역할을 하며 이모티콘이 지속적으로 추가되고 변화하는 상황에서 새로운 이모티콘이 도입될 때마다 계층적 KoBERT의 서브 모델만을 학습시키면 되므로, 신속한 대응이 가능하다는 장점이 있다.

6-1 계층적 KoBERT 학습

데이터셋에 있는 모든 문장은 1개의 이모티콘을 가지고 있고 이모티콘은 계층적 이모티콘 카테고리에 속해 있다. 계층적 카테고리는 32개의 메인 카테고리와 각 메인 카테고리 하위에 3~12개 사이의 서브 카테고리로 구성되어 있다. 메인 KoBERT는 SNS 게시글을 32개의 메인 카테고리로 분류하는 역할을 하고, 전체 데이터셋을 사용하여 학습한다. 메인 KoBERT 하위에 32개의 서브 KoBERT를 구성하여 2차 분류를 진행한다. 32개의 서브 KoBERT는 각 서브 카테고리에 속하는 데이터셋을 사용하여 학습하고 서브 모델에 따라 클래스 개수와 데이터셋 개수는 각각 다르다. 각 서브 모델의 클래스 개수는 5~20 범위 내에 있고 모든 서브 모델의 클래스들의 총개수는 314개이다. 학습에 사용된 하이퍼 파라메터는 표 3과 같다.

표 3. 하이퍼 파라메터
Table 3. Hyperparameters

Parameters	Value
Max length	64
Classes	32
Batch size	64
Optimizer	AdamW
Epochs	1,000
Early stopping	20
Learning rate	5e-5

6-2 모델 연결 및 추천

이모티콘 추천 시스템은 1개의 메인 KoBERT 모델과 하위에 32개의 서브 KoBERT 모델로 구성된다. SNS에 사용자가 입력한 게시글은 먼저 메인 KoBERT 모델로 전달되어 메인 카테고리로 1차 분류되고, 이후 해당 서브 KoBERT 모델에서 2차 세부 분류를 거쳐 서브 카테고리가 결정된다. 이모티콘 추천 시스템은 총 314개의 서브 그룹 중 서브 KoBERT 모델의 출력인 서브 카테고리에 속하는 이모티콘을 추천한다. 그림 6을 보면, 계층적 KoBERT 모델의 추론 과정에서, "기분 전환에는 레드와인"이라는 문장이 입력되면 메인 KoBERT 모델에 임베딩 되고 메인 KoBERT 모델에서 Drink라는 메인 클래스로 1차 분류된다. 이후에 메인 클래스 하위에 Drink 클래스를 학습해 놓은 Drink 서브 KoBERT 모델에서 juice, champaign, wine… 등으로 2차 분류가 진행된다. 즉, 사용자가 입력한 게시글은 $c_{i,j}$레이블을 가진 카테고리로 최종 분류되고 $c_{i,j}$ 카테고리에 속하는 이모티콘들을 무작위로 추천한다.

VII. 실험

7-1 실험 설계

계층적 KoBERT 모델을 활용한 이모티콘 추천 시스템의 성능을 측정하기 위해 비교 대상 모델을 선정하고 성능 비교 실험을 수행한다. 계층적 구조를 가진 데이터셋을 사용하여 계층적 KoBERT 모델을 학습시킨 후 추천 성능을 비교한다. 비교 대상 모델로서 텍스트 분류 분야에서 널리 사용되고 있는 DNN, LSTM, Bi-LSTM, 그리고 GRU 모델들을 선정하고 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), 그리고 F1-Score를 측정 비교한다.

• DNN (Deep Neural Network): 가장 기본적인 신경망 중의 하나로서 여러 개의 은닉층을 가지며 비선형성을 이용하여 고차원의 복잡한 데이터셋에서 패턴을 찾아내는 데 사용된다.

• LSTM (Long Short-Term Memory): 이전 계산 결과에 관한 메모리를 도입하여 길이가 긴 시퀀스 정보를 기억하고 전파할 수 있는 신경망으로서 순차적인 데이터를 학습하는 데 장점을 가진다.

• Bi-LSTM (Bidirectional LSTM): Bi-LSTM은 양방향 LSTM으로 시간적 의존성이 있는 양방향 데이터를 처리하는 데 주로 사용된다. 순방향과 역방향의 양방향으로 입력 시퀀스를 처리하는 방식으로서 텍스트 분류 분야에서 널리 사용된다.

• GRU (Gated Recurrent Unit): 순환 신경망(RNN)의 일종으로서 LSTM과 유사하고 리셋 게이트와 업데이트 게이트의 상호작용을 통해 학습한다. LSTM보다 파라메터 수가 적어서 학습 속도가 빠른 장점을 가진다.

7-2 실험 결과 및 분석

첫 번째 실험은 계층적 KoBERT 모델에서 메인 모델의 성능을 측정한다. 메인 KoBERT 모델은 SNS 게시글을 이모티콘 메인 카테고리 32개 중 하나로 분류한다. 이를 위해 메인 KoBERT 모델과 비교군 모델인 DNN, LSTM, Bi-LSTM, GRU 4개 모델의 정확도, 정밀도, 재현율, 그리고 F1-Score를 그림 7과 같이 비교하였다. 메인 KoBERT 모델의 정확도는 0.489의 성능을 보여주고 있으며, DNN 대비 0.271, LSTM 대비 0.247, Bi-LSTM 대비 0.228, 그리고 GRU 대

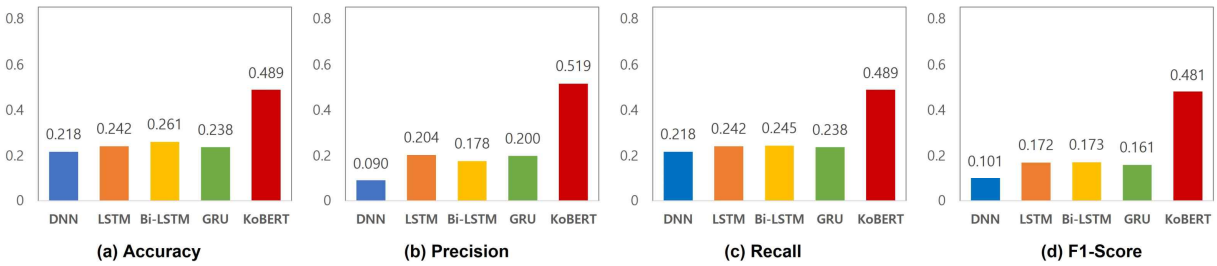


그림 7. 계층적 학습모델의 메인 모델 성능 비교 (DNN, LSTM, Bi-LSTM, GRU, KoBERT)
Fig. 7. Comparison of main model performances in hierarchical learning model(DNN, LSTM, Bi-LSTM, GRU, KoBERT)

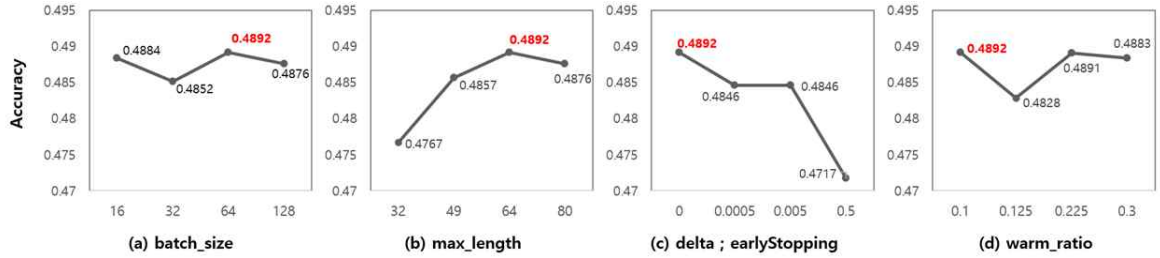


그림 8. 하이퍼 파라메터 튜닝 실험
Fig. 8. Effects of hyper parameters on model accuracy

비 0.251 정도 성능이 향상되었다. 메인 모델의 정밀도는 0.519, 재현율은 0.489, F1-Score는 0.481로 모든 성능지표에서 다른 모델과 비교하여 높은 성능을 보여주고 있다. 그림 8에서는 메인 KoBERT 모델 학습에서 미세 조정한 하이퍼 파라메터 값이 이모티콘 추천 정확도에 미치는 영향을 보여주고 있다.

두 번째 실험에서는 서브 KoBERT 모델의 성능을 평가하였다. 서브 KoBERT 모델은 메인 KoBERT 모델 하위에 있는 32개의 모델로서 서브 카테고리의 분류를 수행한다. 그림 9는 32개의 서브 모델별로 KoBERT 모델과 비교군 4개 모델(DNN, LSTM, Bi-LSTM, 그리고 GRU)의 성능을 비교한 결과를 보여주고 있다. 각 서브 KoBERT 모델마다 훈련 데이터셋의 크기와 클래스 개수가 달라서 성능도 각기 다른 것을 확인할 수 있다. 중요한 점은 다른 모델들과 비교하여 서브 KoBERT 모델이 정확도를 포함한 모든 성능지표에서 높은 수치를 보였다.

세 번째 실험에서는 계층적 KoBERT 모델의 성능을 검증하기 위하여 메인 모델과 서브 모델을 연결하고 전체 테스트 데이터에 대해서 최종 이모티콘 추천 성능을 표 4와 같이 비교 분석하였다. 5개 모델의 최종 연결 정확도를 비교한 결과, KoBERT 모델이 가장 높은 성능을 보였고 DNN 모델이 가장 낮은 성능을 보였다.

이모티콘 추천에 관한 기존 연구들은 추천하는 이모티콘 카테고리 개수가 30~100개 정도로 적은 상황에서 Top-5 정확도 0.4를 상회하는 정도이다. 이 연구에서는 314개의 대규모 이모티콘 카테고리를 추천하는 데 있어서 계층적 접근법을 사용하여 0.538로 Top-5 정확도를 향상시켰다. 이모티콘 추천 실험을 통해 계층적 KoBERT 모델은 복잡하고 다양한 SNS 게시글의 문맥을 파악하여 적합한 이모티콘을 추천하는 데 있어서 다른 모델에 비해 보다 좋은 성능을 보이는 것을 확인할 수 있다.

표 4. 다른 모델 대비, 계층적 KoBERT 모델의 성능 비교
Table 4. Performance of hierarchical KoBERT model compared to different models

	Accuracy@T1	Accuracy@T3	Accuracy@T5	Precision	Recall	F1-Score
DNN	0.086	0.134	0.155	0.060	0.180	0.080
LSTM	0.104	0.180	0.226	0.169	0.184	0.133
Bi-LSTM	0.107	0.182	0.234	0.136	0.197	0.136
GRU	0.089	0.147	0.176	0.072	0.119	0.082
KoBERT	0.407	0.492	0.538	0.992	0.407	0.537

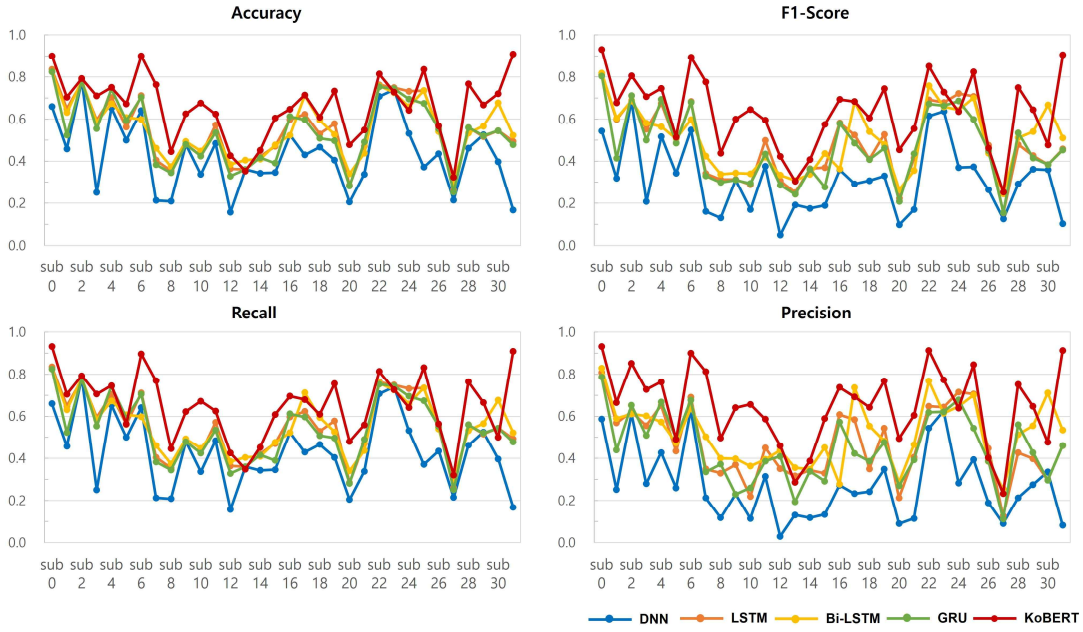


그림 9. 서브 모델 성능 비교 (DNN, LSTM, Bi-LSTM, GRU, KoBERT)
Fig. 9. Comparison of sub model performances (DNN, LSTM, Bi-LSTM, GRU, KoBERT)

VIII. 결론 및 향후 연구

이 연구에서는 인스타그램 SNS 플랫폼에서 사용자가 작성한 게시글의 문맥을 분석하여 게시글의 내용을 보완하거나 강조하는 이모티콘을 자동 추천하는 시스템을 제안하였다. 인스타그램 게시글 172,149개를 수집하여 자체 데이터셋을 구성하고 게시글에서 추출한 616여 개의 방대하고 종류도 매우 다양한 이모티콘을 대상으로 추천한 부분에 의의가 있다. 이 문제는 게시글의 문맥을 분석하여 문맥 기반으로 이모티콘을 추천하는 것이고, 이 연구에서는 이 문제를 이모티콘 카테고리를 기준으로 게시글을 분류하는 문제로 정형화하였다. 특히, 616개나 되는 대량의 이모티콘을 대상으로 추천이 가능한 원인은 클래스가 많은 텍스트 분류 문제를 해결하기 위해서 계층적 방법으로 접근하였기 때문이다. 우리의 계층적 접근 방법은 다양한 감정과 주제를 포함하며 상/하위 카테고리가 존재하는 유니코드 이모티콘의 특징을 그대로 반영한다는 점에서 의미가 크다. 314개의 이모티콘 카테고리를 계층적으로 생성하고 계층적 KoBERT 모델을 도입, 학습시켜 SNS 게시글을 32개의 메인 카테고리와 314개의 서브 카테고리로 분류하였다. 우리의 방법은 대량의 이모티콘을 대상으로 추천하는 데 적합하며 게시글의 문맥이나 내용을 반영하는 다양한 이모티콘을 추천하는 데 유용하게 활용될 수 있다.

본 연구에서 구축한 계층적 KoBERT 모델은 Top-K 정확도 0.538의 이모티콘 추천 성능을 제시하고 있다. 기존의 이모티콘 추천 연구들에서 30~100개의 소규모 이모티콘 카테고리를 추천하는 성능이 0.4을 다소 상회하는 정도인 것을 기준으로 볼 때, 우리의 시스템은 이모티콘 추천 성능을 향상시켰다고 할 수 있다. 계층적 KoBERT 모델의 성능 검증을 위하여 DNN, LSTM, Bi-LSTM, GRU의 4개 모델을 비교군으로 선정하여 성능 비교 실험을 진행하였다. 메인 모델, 서브 모델, 그리고 계층적 모델에 대해 정확도, 정밀도, 재현율, F1-Score를 각각 측정하였다. 실험 결과, 다른 비교군 모델 대비, 계층적 KoBERT 모델은 모든 성능지표에서 높은 성능을 보여주고 있다. 또한, 임의의 게시글을 입력으로 이모티콘을 추천하는 추론 실험을 진행하고 임의의 게시글에 추천된 이모티콘을 살펴본 결과, 계층적 KoBERT 모델이 LSTM 모델보다 텍스트 문맥 분석 및 파악 능력이 우수한 것을 확인할 수 있었다. 이는 BERT 모델이 양방향으로 자연어를 처리하는 언어모델이고, 특히 KoBERT 모델의 경우 한국어의 불규칙한 언어 변화의 특성을 반영하기 위해 개선된 모델로서 기존 LSTM 계열의 모델보다 텍스트 문맥이나 내용 분석에 유리하기 때문이다.

현재, 이모티콘 추천 정확도는 0.538 정도의 성능을 보이고 있는데, 이는 이모티콘 사용 형태가 개인의 연령, 취향이나 감성에 의존적이기 때문이다. 향후, 이모티콘을 사용하는 사용자들과 사용 형태를 분석하고 이러한 특성을 반영하여 이모티콘을 추천하는 새로운 방법론을 연구하고 추천 성능을 향상시키고자 한다.

감사의 글

이 논문은 2021년도 성신여자대학교 학술연구조성비 지원에 의하여 연구되었음.

참고문헌

[1] E. J. Lee, “Motivations for the Using Emoticon : Exploring the Effect of Motivations and Intimacies between Users on the Attitude and Behaviors of Using Emoticon,” *Journal of the HCI Society of Korea*, Vol. 12, No. 2, pp. 5-12, May 2017. https://doi.org/10.17210/jhsk.2017.05.12.2.5

[2] Y. I. Hong and S. K. Yim, “The Effect of Emoticon Expression Type on User Satisfaction Factors,” *The Treatise on The Plastic Media,* Vol. 25, No.1, pp. 33-41, February 2022. https://doi.org/10.35280/KOTPM.2022.25.1.4

[3] V. N. Durga Kollipara, V. N. Hemanth Kollipara, and M. Durga Prakash, “Emoji Prediction from Twitter Data Using Deep Learning Approach,” *Asian Conference on Innovation in Technology (ASIANCON)*, Pune, India, August 2021. https://doi.org/10.1109/ASIANCON51346.2021.9544680

[4] X. Zheng, G. Zhao, L. Zhu, and X. Qian, “PERD: Personalized Emoji Recommendation with Dynamic User Preference,” *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp. 1922-1926, July 2022. https://doi.org/10.1145/3477495.3531779

[5] S. Lee, E. Lee, and D. Park, “Emoticon Recommendation using Emotional Analysis,” *Proceedings of 2021 Winter Symposium of the Korean Institute of Communications and Information Sciences*, pp. 864-865, February 2021.

[6] L. Zhao and C. Zeng, “Using Neural Networks To Predict Emoji Usage from Twitter Data,” *Computer Science*, 2017.

[7] K. Matsumoto, M. Yoshida, and K. Kita, “Classification of Emoji Categories from Tweet Based on Deep Neural Networks,” *Proceedings of the 2nd International Conference on Natural Language Processing and Information Retrieval (NLPIR),* pp. 17-25, September 2018. https://doi.org/10.1145/3278293.3278306

[8] M. Pek and M. Turan, “Sentiment Analysis from Tweets using Convolutional Neural Networks,” *The Journal of Communications, Information, Electronic and Energy System (CIEES)*, Vol. 1, No. 1, June 2021. https://doi.org/10.48149/jciees.2021.1.1.2

[9] M. Bień, K. Guyard, and Y. Li, “Predicting Topic Change and Emoji Usage from Twitter Data,” *Computer Science*, December 2020.

[10] J. Shobana, S. Amudha, and S. Kumar, “Emoji Anticipation and Prediction Using Deep Neural Network Model,” *2022 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS),* December 2022. https://doi.org/10.1109/ICPECTS56089.2022.10047692

[11] R. Xie, Z. Liu, R. Yan, and M. Sun, “Neural Emoji Recommendation in Dialogue Systems,” December 2016. https://doi.org/10.48550/arXiv.1612.04609

[12] G. Guibon, M. Ochs, and P. Bellot, “Emoji Recommendation in Private Instant Messages,” *Proceedings of the 33rd Annual ACM Symposium on Applied Computing (SAC),* pp. 1821-1823, April 2018. https://doi.org/10.1145/3167132.3167430

[13] J. Kim, T. Gong, B. Kim, J. Park, W. Kim, E. Huang, and S.-J. Lee, “No More One Liners: Bringing Context into Emoji Recommendations,” *ACM Transactions on Social Computing,* Vol. 3, No. 2, pp. 1-25, April 2020. https://doi.org/10.1145/3373146

[14] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann, “Using Millions of Emoji Occurrences to Learn Any-Domain Representations for Detecting Sentiment, Emotion and Sarcasm,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1615-1625, August 2017. https://doi.org/10.48550/arXiv.1708.00524

[15] T. Tomihira, A. Otsuka, A. Yamashita, and T. Satoh, “Multilingual Emoji Prediction using BERT for Sentiment Analysis,” *International Journal of Web Information System*, Vol. 16, No. 3, pp. 265-280, September 2020. https://doi.org/10.1108/IJWIS-09-2019-0042

[16] P. Zhao, J. Jia, Y. An, J. Liang, L. Xie, and J. Luo, “Analyzing and Predicting Emoji Usages in Social Media,” *WWW '18: Companion Proceedings of the The Web Conference,* pp. 327-334, April 2018. https://doi.org/10.1145/3184558.3186344

[17] G. Zhao, Z. Liu, Y. Chao, and X. Qian, “CAPER: Context-Aware Personalized Emoji Recommendation,” *IEEE Transactions on Knowledge and Data Engineering,* Vol. 33, No. 9, pp. 3160-3172, January 2020. https://doi.org/10.1109/TKDE.2020.2966971

[18] Z. Gao, A. Feng, X. Song, and X. Wu, “Target-Dependent Sentiment Classification with BERT,” *IEEE Access*, Vol. 7, pp. 154290-154299, October 2019. https://doi.org/10.1109/ACCESS.2019.2946594

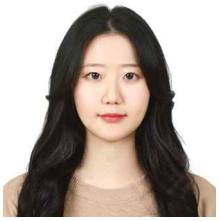
[19] Z. Ke, H. Xu, and B. Liu, “Adapting BERT for Continual Learning of a Sequence of Aspect Sentiment Classification Tasks,” *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4746-4755, June 2021. https://doi.org/10.18653/v1/2021.naacl-main.378

[20] A. S. Alammary, “BERT Models for Arabic Text Classification: A Systematic Review,” *Applied Sciences*, Vol. 12, No. 11, 5720.

https://doi.org/10.3390/app12115720
[21] X. Chen, P. Cong, and S. Lv, “A Long-Text Classification Method of Chinese News Based on BERT and CNN,” *IEEE Access*, Vol. 10, pp. 34046-34057, March 2022. https://doi.org/10.1109/ACCESS.2022.3162614
[22] Y.-J. Lee and H.-J. Choi, “Joint Learning-based KoBERT for Emotion Recognition in Korean,” in *Proceedings of the 2020 Korean Information Science Society Conference*, pp. 568-570, December 2020.
[23] C. Lee and M. Moon, “Keyword and Emotional Analysis Diary Service Using KoNLPy and KoBERT,” in *Proceedings of the 2022 Korean Society of Computer Information Conference*, Vol. 30, No. 2, pp. 501-502, July 2022.
[24] R. Pappagari, P. Zelasko, J. Villalba, Y. Carmiel, and N. Dehak, “Hierarchical Transformers for Long Document Classification,” *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 838-844, December 2019. https://doi.org/10.1109/ASRU46091.2019.9003958
[25] H. Srivastava, V. Varshney, S. Kumari, and S. Srivastava, “A Novel Hierarchical BERT Architecture for Sarcasm Detection,” iin *Proceedings of the Second Workshop on Figurative Language Processing,* pp. 93-97, July 2020. https://doi.org/10.18653/v1/2020.figlang-1.14

김지현(Jee-Hyun Kim)

2021년 : 성신여자대학교 정보시스템공학과 (공학사)
2021년~현 재: 성신여자대학교 대학원 미래융합기술공학과 석사과정
※관심분야 : 딥러닝(Deep Learning), 자연어처리(Natural Language Processing), 데이터 마이닝(Data Mining)



김예림(Ye-Rim Kim)

2020년 : 성신여자대학교 수학과 (이학사)
2020년 : 성신여자대학교 IT학부 (공학사)
2021년~현 재: 성신여자대학교 대학원 미래융합기술공학과 석사과정
※관심분야 : 딥러닝(Deep Learning), 자연어처리(Natural Language Processing), 컴퓨터 비전(Computer Vision)

변혜원(Hae-Won Byun)

1990년 : 연세대학교 전산과학과 (공학사)
1992년 : KAIST 대학원 (공학석사)
2004년 : KAIST 대학원 (공학박사-컴퓨터그래픽스)
2006년~현 재: 성신여자대학교 AI융합학부 교수
※관심분야 : 컴퓨터 그래픽스(Computer Graphics), 딥러닝(Deep Learning), 생성형 AI(Generative AI) 등