텍스트 분류 성능 향상: 레이블 기반 컨텍스트 주석 생성에 관한 연구

박눈솔

한국지능정보사회진흥원

nspark@nia.or.kr

Advancing Text Classification Performance: A Study on Label-based Context Annotation Generation.

NunSol Park

NIA(National Information Society Agency).

요 약

본 논문은 텍스트 분류 성능 향상을 위한 방법론으로 레이블 기반 컨텍스트 주석 생성(Context Annotation Generation) 방안을 제안한다. 이 방법은 기존의 레이블 데이터셋(Label Dataset)을 활용하여 각 데이터 레이블에 대한 컨텍스트 주석을 자동으로 생성한다. 생성된 주석은 단순한 레이블을 넘어 텍스트 분류 결정의 근거와 관련 배경 정보를 포함하여, 분류 모델이 더 깊이 있는 학습을 할 수 있도록 지원한다. 본 논문의 연구 결과는 텍스트 분류 과제에서 도전적인 과제로 남아 있던 레이블 데이터셋 부족 문제를 효과적으로 해결하는 동시에, 데이터 레이블링(Data Labeling) 작업에 소요되는 비용과 시간을 절감할 수 있는 새로운 방향을 제시한다. 이는 다양한 도메인 환경이나 자원이 제한된 환경에서의 텍스트 분류 작업에 적용될 수 있으며, 실험에서 분류 모델 성능 개선에 대한 효과를 입증하였다.

Ⅰ. 서 론

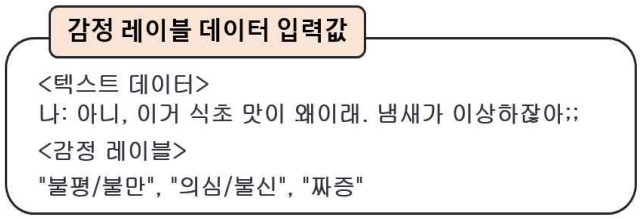
텍스트 분류(Text Classifiacation)는 자연어 처리의 핵심적인 과제로, 문서 분류, 감성 분석, 스팸 탐지 등 다양한 응용 분야에서 중요한 역할을 하고 있다[1]. 최근 딥러닝 기술의 발전으로 텍스트 분류 모델의 성능이 크게 향상되었지만, 이러한 모델들의 효과적인 학습을 위해서는 대규모의 고품질 레이블 데이터셋(Label Dataset)이 필요하다[2]. 그러나 실제 환경에서는 충분한 양의 레이블 데이터를 확보하는 것이 큰 도전 과제로 남아 있다. 데이터 레이블링(Data Labeling) 작업은 시간과 비용이 많이 소요되는 과정으로, 특히 전문 지식이 필요한 도메인(Domain)에서는 숙련된 전문가의 참여가 필수적이라 데이터셋 구축 비용을 크게 증가시킨다[3]. 또한, 레이블 데이터의 부족은 모델의 일반화 능력을 저하시킬 수 있으며, 제한된 데이터로 학습된 모델은 새로운 데이터에 대해 낮은 성능을 보일 가능성이 높다. 이러한 배경에서, 본 연구에서는 레이블 기반 컨텍스트 주석 생성(Context Annotation Generation)이라는 새로운 접근법을 제안한다. 이 방법은 기존의 레이블된 데이터셋을 최대한 활용하여, 각 레이블에 대한 컨텍스트 주석을 자동으로 생성한다. 생성된 주석은 단순한 레이블을 넘어 분류 결정의 근거와 관련 배경 정보를 포함하여, 모델이 더 깊이 있는 학습을 할 수 있도록 한다. 이러한 접근법은 데이터 레이블 구축 작업에 소요되는 비용과 시간을 절감하면서도, 모델의 성능과 일반화 능력을 향상시켜 그 효과를 입증하였다. 특히 자원이 제한된 환경이나 전문 도메인에서의 텍스트 분류 모델의 적용을 가속화할 수 있을 것으로 기대된다.

Ⅱ. 본론

본 논문에서는 초거대 언어 모델(LLM, Large Language Model)을 활용한 레이블 기반 컨텍스트 생성으로 분류 모델 성능을 향상하는 방법론을 제안한다. 최근 GPT-4(Generative Pre-trained Transformer 4)[4]와 같은 초거대 언어 모델이 다양한 자연어 처리 과제에서 뛰어난 성능을 보여주고 있으며, 탁월한 생성 능력을 갖고 있다는 점을 주목했다[5]. 이러한 초거대 언어모델을 활용하여, 본 연구를 진행하였다.

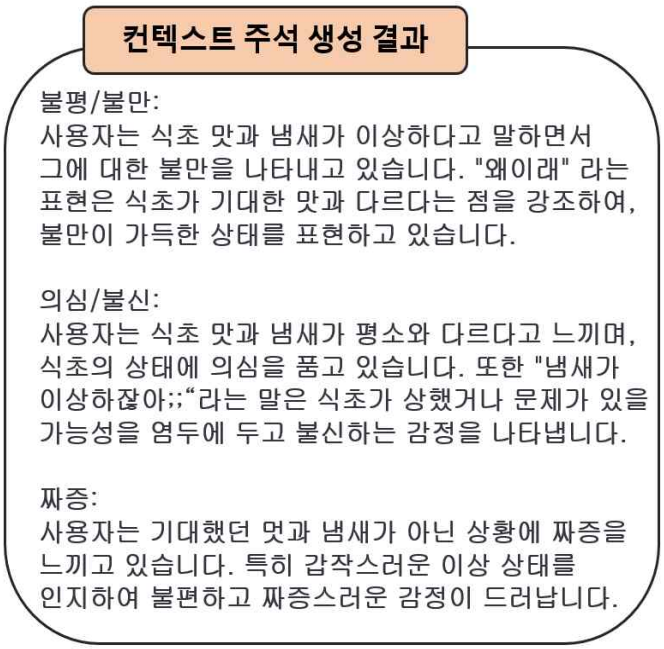
1. 감정 레이블에 대한 컨텍스트 주석 생성

컨텍스트 주석 생성(Context Annotation)은 본 연구의 핵심적인 부분이다. 컨텍스트 주석(Context Annotation)은 주어진 텍스트에 대해 왜 특정 감정 레이블이 부여 되었는지를 설명하는 맥락 정보가 반영된 주석이다. 이는 단순한 레이블 정보를 넘어서, 텍스트의 언어적 특성, 문맥적 의미, 감정 표현의 뉘앙스 등을 포함한다. 입력 데이터는 한국어 온라인 댓글 감정 데이터셋(KOTE, Korean Online That-gul Emotions)으로 다양한 플랫폼으로부터 수집된 한국어 댓글 감정 데이터셋이다[6]. 해당 데이터셋은 44개의 다양한 감정 레이블로 구성되어 있으며, 이 데이터셋에서 10,000개의 데이터를 랜덤 샘플링하여 컨텍스트 주석 생성을 위한 시드 데이터셋으로 사용하였다. 시드 데이터셋을 입력으로 하여 GPT-4 모델을 활용하여 감정 레이블에 대한 컨텍스트 주석(Context Annotation)을 생성하였다. 컨텍스트 주석을 생성하기 전 입력값 즉, 텍스트 데이터와 감정 레이블은 아래 [그림 1]과 같다.



[그림 1] 컨텍스트 주석 생성 전 입력값 예시

[그림 1]과 같은 입력값을 바탕으로 컨텍스트 주석을 생성하면 아래 이미지와 같다.

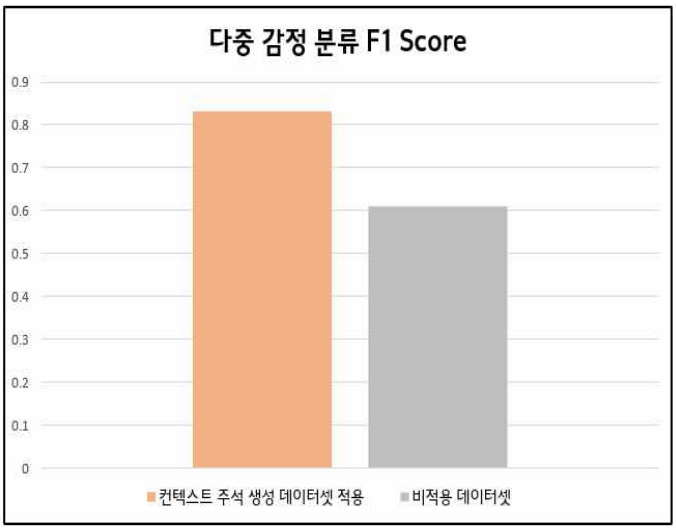


[그림 2] 컨텍스트 주석 생성 결과 예시

컨텍스트 주석은 원본 텍스트 데이터를 바탕으로 감정 레이블과 상세한 설명으로 구성된다. 각 감정 레이블에 대해 왜 해당 감정으로 분류되었는지 설명을 제공하며, 사용된 어휘, 문장 구조, 표현 방식 등을 분석하여 감정 상태를 분석한다. 또한 텍스트의 전반적인 맥락을 고려한 문맥적 해석을 바탕으로 컨텍스트 주석을 생성한다.

2. 컨텍스트 주석 생성 데이터 적용 분류 모델 훈련

실험에서 컨텍스트 주석 생성 데이터셋을 적용하여 분류 모델을 훈련시킨 후 성능을 비교하였다. 훈련 결과는 아래 [표 1]과 같다.



[표 1] 데이터셋에 따른 다중 감정 분류 F1 score 비교 표

좌측의 주황색 막대는 컨텍스트 주석 생성 데이터셋을 적용한 분류 모델의 F1 Score이며, 우측의 회색 막대는 컨텍스트 주석 생성 데이터셋을 적용하지 않은 분류 모델의 F1 Score이다. 분류 모델은 Polyglot[7] 모델로 두 막대 모두 같은 모델을 사용하였으며, [표 1]의 두 막대그래프를 비교했을 때, 분류 모델 성능의 큰 차이가 있음을 입증하였다. 이를 통해 레이블 기반 컨텍스트 생성 데이터셋이 감정 텍스트 분류 작업에서 모델의 성능을 크게 향상시킨다는 것을 확인할 수 있으며, 실험적으로 큰 의의가 있음을 입증하였다. 본 연구에서 진행한 실험은 44개의 다중 감정 분류 작업이기에 레이블의 개수를 줄이거나 단순한 이진 분류 작업으로 훈련시키면 F1 Score가 더 높을 것으로 예상된다.

Ⅲ. 결론

본 연구에서는 컨텍스트 주석을 생성하여 분류 모델의 성능을 향상시키는 새로운 접근법을 제시하였다. 생성된 컨텍스트 주석 데이터셋을 적용한 분류 모델은 기존 데이터셋 대비 유의미한 성능 향상을 보였으며, 44개의 다중 감정 분류 작업에서도 뛰어난 분류 성능 향상을 입증하였다. 레이블 기반 컨텍스트 주석 생성을 통해 분류 모델이 더 깊은 수준의 언어 이해와 추론 능력을 개발할 수 있음을 보여주었다. 또한 본 연구의 접근법은 높은 비용의 데이터 레이블 구축이 아닌 낮은 비용으로 고품질의 데이터셋을 구축할 수 있는 방법을 제시하여, 자원이 제한된 연구 환경에서의 활용 가능성을 보여주었다. 컨텍스트 주석을 통해 단순한 레이블 정보 이상의 풍부한 맥락 정보를 제공함으로써, 모델의 감정 인식 능력을 크게 향상시켰다. 이는 미묘한 감정 차이나 복합적인 감정 표현을 더 정확히 분류할 수 있도록 하게 한다.

본 연구에서 제시한 방법론은 텍스트 분류 작업에서의 감정 분류 작업뿐만이 아니라 다양한 분류 작업 그리고 자연어 처리 작업에 적용될 수 있는 잠재력을 보여준다. 이러한 다양한 의의를 통해, 본 연구는 자연어 처리 분야에서 레이블 기반 컨텍스트 주석 생성과 분류 모델 훈련에 대한 새로운 패러다임을 제시하고, 언어 모델의 이해 능력을 한 단계 발전시키는 데 기여하였다고 볼 수 있다.

참 고 문 헌

[1] Qian L, “A survey on Text Classification: From Traditional to Deep Learning”, arXiv, 2021.

[2] Berna A, “Semantic text classification: A survey of past and recent advances”, ScienceDirect, 2018.

[3] Laith A. “A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications”, Journal of Big Data 10, 46, 2023.

[4] Josh A, “GPT-4 Technical Report”, arXiv, 2023.

[5] Wayne X, “A Survey of Large Language Models”, arXiv, 2023.

[6] Duyoung J, “User Guide for KOTE: Korean Online Comments Emotions Dataset”, arXiv, 2022.

[7] Hyunwoong K, “A Technical Report for Polyglot-Ko: Open-Source Large Scale Korean Language Models”, arXiv, 2023.