Large Language Models 의 텍스트 분류 성능 향상을 위한 기법 연구

김원철, 이웅기, 김기환
롯데이노베이트

wonchul_kim@lotte.net, unggi.lee@lotte.net, gihwan.kim@lotte.net

A Study on Techniques for Improving Text Classification Performance of Large Language Models

Kim Won Chul, Lee Ung Gi, Kim Gi Hwan
Lotte Innovate

요 약

본 논문은 대규모 언어 모델(LLM)을 활용한 텍스트 분류 태스크의 성능 향상을 위한 새로운 접근 방식을 제안한다. 기존의 LLM 기반 분류 방식이 가진 레이블 편향 및 성능 불안정성 문제를 해결하기 위해, 본 연구는 분류 태스크를 생성 문제로 변환하는 방법론을 소개한다. 이 방법은 전통적인 분류 방식보다 안정적인 성능을 제공하고, 모델의 편향을 줄이며, 여러 단계가 필요한 태스크를 하나로 통합하여 보다 효율적으로 처리할 수 있는 장점이 있다. Naver Sentiment Movie Corpus(NSMC)와 자체 구축한 웹 검색 판단 데이터셋을 사용한 실험 결과, 제안된 방법은 기존의 분류 접근법보다 우수한 성능을 달성했다. 본 연구는 LLM 을 이용한 텍스트 분류에서 생성 기반 접근 방식의 효과성을 입증하며, 이는 LLM 의 활용 범위를 확장하고 다양한 분류 문제에 대한 효과적인 해결책을 제공할 것으로 기대된다.

Ⅰ. 서 론

대규모 언어 모델(LLM)은 다양한 태스크에서 인간 수준의 성능을 보이며, 텍스트 생성, 번역, 요약 등 다양한 분야에서 활용 가능성을 보여주고 있다. 그러나, LLM 을 이용한 분류 태스크에서는 아직 몇 가지 문제점이 남아 있다. 특히, 학습 과정에서 모델이 특정 레이블에 치우치는 편향이 발생할 가능성이 존재한다는 점은 중요한 문제로 대두되고 있다.

본 연구에서는 이러한 문제를 해결하기 위해 LLM 을 이용한 직접적인 분류 접근 방식 대신 생성 문제로 변환하여 접근하는 방법론을 제안한다. 이 방법은 단순 분류 방식보다 더 안정적인 분류 성능을 제공할 뿐만 아니라, 모델의 편향을 줄이는 데 도움을 준다. 또한, 여러 태스크를 하나로 통합하여 더 효율적인 처리 과정을 가능하게 함으로써, LLM 의 활용도를 극대화할 수 있다.

제안 방법의 우수성을 입증하기 위해 두 가지 데이터셋을 사용하여 실험을 진행하였다. 첫째, 영화 리뷰 감성 분석을 위한 대표적인 한국어 데이터셋인 Naver Sentiment Movie Corpus(NSMC) [1]를 활용하였다. 둘째, 사용자 질문이 웹 검색을 요하는지 판단하기 위해 자체적으로 구축한 웹 검색 판단 데이터셋을 사용하였다. 실험 결과, 제안 방법이 기존의 일반적인 분류 방식보다 우수한 성능을 보여주었으며, 이를 통해 본 방법론이 다양한 분류 문제에 효과적으로 적용될 수 있음을 확인하였다.

Ⅱ. 관련 연구

최근 LLM 을 활용한 텍스트 분류 연구가 활발히 진행되고 있다. [2]에서 제안된 Clue And Reasoning Prompting (CARP)는 텍스트 분류를 위한 단계적 추론 전략을 도입하여 LLM 의 성능을 향상시켰다. CARP 는 입력 텍스트의 표면적 단서를 찾고, 이를 바탕으로 진단적 추론 과정을 유도하며, 미세조정된 모델로 kNN 예시 검색을 수행한다. 이 접근법은 5 개의 널리 사용되는 텍스트 분류 벤치마크 중 4 개에서 새로운 최고 성능을 달성했으며, 적은 리소스 환경과 도메인 적응 상황에서도 뛰어난 성능을 보였다. 이 연구는 LLM 을 활용한 텍스트 분류의 성능 향상을 위한 효과적인 프롬프팅 기법을 제시했다는 점에서 의의가 있다.

Ⅲ. 제안 방법

우리가 제안하는 방법론은 LLM 을 이용한 분류 태스크를 수행할 때, 직접적인 분류 접근 대신 다른 생성 문제로 변환하여 접근하는 것이다. 이 방법은 더

안정적인 분류 성능을 얻을 수 있을 뿐만 아니라, LLM 이 특정 레이블에 치우치는 문제도 해결할 수 있다. 또한, 이 접근법의 장점으로 여러 태스크를 하나로 통합할 수 있어 더 효율적인 프로세스가 가능하다는 점이 있다.

예를 들어, 웹 검색 필요성 판단과 키워드 생성이라는 두 가지 태스크가 있다면, "웹 검색이 필요하면 적절한 키워드를 생성하고, 불필요하면 '검색이 필요 없습니다'라고 출력하시오"라는 단일 지시문으로 통합할 수 있다. 이렇게 통합된 태스크는 개별적인 분류 태스크보다 더 나은 결과를 제공한다.

LLM 은 다양한 태스크를 수행할 수 있지만, 경험상 단순 분류 태스크를 학습시키면 모델이 쉽게 편향될 수 있다. 이는 학습 시간이 긴 LLM 에게 특히 치명적인 문제로 우리의 방법은 이러한 위험을 줄이면서도 높은 분류 성능과 안정적인 결과를 제공한다. 현재까지 다양한 분류 문제에 이 방법을 적용한 결과, 높은 분류 정확도와 편향되지 않은 안정적인 결과를 얻을 수 있었다.

Ⅳ. 실험 데이터

본 연구는 Naver Sentiment Movie Corpus(NSMC)와 자체 제작한 웹 검색 판단 데이터셋을 실험에 사용하였다. NSMC 는 네이버 영화 리뷰를 기반으로 총 200,000 개의 문장으로 구성되어 있으며, 150,000 개는 훈련용, 50,000 개는 테스트용으로 분류되어 있다. 본 연구에서는 20 자 이상의 리뷰 문장만을 선별하고, 긍정과 부정 리뷰를 1:1 비율로 샘플링하였다. 최종적으로 학습용 500 개와 테스트용 100 개의 리뷰를 선정하여 실험에 사용하였다.

다양한 실험 결과를 얻기 위해 웹 검색 판단을 위한 자체 데이터셋을 구축하여 활용하였다. 이 데이터셋은 사용자 질의가 웹 검색이 필요한지 여부를 판별하는 분류 문제로 구성되어 있다. 두 데이터셋의 자세한 구성은 표 1 에 제시되어 있다.

(표 1) 실험 데이터 구성

Dataset	NSMC		Web Search	
	긍정	부정	필요	불필요
Train	250	250	266	110
Test	50	50	70	22

(그림 1) 웹 검색 판단 데이터 예시

```
[
  {
    "method": "Baseline",
    "instruct": "대화를 보고 검색이 필요해 보이는 상황이면 '검색 필요'라고 답변하고,
      검색이 필요 없는 상황이면 '검색 불필요'라고 답하세요.",
    "conversations": [
      {
        "from": "human",
        "value": "롯데이노베이트와 관련된 최신 기사들을 찾아주실 수 있을까요?"
      },
      {
        "from": "gpt",
        "value": "검색 필요"
      }
    ]
  },
  {
    "method": "Proposed",
    "instruct": "대화를 보고 검색이 필요해 보이는 상황이면 적절한 검색어를 생성해주고,
      검색이 필요 없는 상황이면 검색이 필요 없다고 답하세요.",
    "conversations": [
      {
        "from": "human",
        "value": "롯데이노베이트와 관련된 최신 기사들을 찾아주실 수 있을까요?"
      },
      {
        "from": "gpt",
        "value": "롯데이노베이트 뉴스"
      }
    ]
  }
]
```

Ⅴ. 실험

본 실험에서는 Llama-3 8B 모델[3]을 기본 모델로 사용하였으며, 모든 학습 파라미터와 실험 환경을 동일하게 설정하였다. 학습에 사용된 지시문은 제안 방법에 따라 분류 태스크를 키워드나 검색어 생성 태스크로 접근하였다. 각 데이터셋에 사용된 지시문은 표 2 와 3 에 제시되어 있다.

(표 2) NSMC 지시문

Method	Instruction
Baseline	다음 내용이 긍정적인 내용이면 '긍정', 부정적인 내용이면 '부정'이라고 답하세요.
Proposed	다음 내용이 긍정적인 내용이면 키워드를 생성하고, 부정적인 내용이면 '긍정적인 내용이 없습니다.'라고 답하세요.

(표 3) 웹 검색 판단 데이터 지시문

Method	Instruction
Baseline	대화를 보고 검색이 필요해 보이는 상황이면 '검색 필요'라고 답변하고, 검색이 필요 없는 상황이면 '검색 불필요'라고 답하세요.
Proposed	대화를 보고 검색이 필요해 보이는 상황이면 적절한 검색어를 생성해주고, 검색이 필요 없는 상황이면 검색이 필요 없다고 답하세요.

두 실험 데이터셋에서 제안 방법이 Baseline 보다 높은 분류 성능을 보였다. NSMC 데이터셋에서 제안 방법은 F1 0.852, 정확도 0.850, ROC-AUC 0.850 을 달성했고, 검색 판단 데이터셋에서는 F1 0.955, 정확도 0.935, ROC-AUC 0.9289 로 최고 성능을 기록했다. 이는 제안 방식의 효과성을 입증한다.

(표 4) 실험 결과

	NSMC			Web Search		
Method	F1	Acc	ROC	F1	Acc	ROC
Baseline	83.8	83.0	0.83	94.6	91.3	0.82
Proposed	**85.2**	**85.0**	**0.85**	**95.5**	**93.5**	**0.93**

Ⅵ. 결 론

본 연구는 LLM 을 이용한 분류 태스크를 생성 문제로 변환하여 수행함으로써 더 높은 성능과 안정성을 얻을 수 있음을 입증했다. 이 접근 방식은 LLM 의 레이블 편향 문제를 해결하고, 여러 단계가 필요한 태스크를 하나로 통합하여 보다 효율적으로 처리할 수 있다. NSMC 와 웹 검색 판단 데이터셋을 통해 제안 방법의 우수성을 확인했으며, 모든 지표에서 기존 분류 방식을 능가했다. 이는 본 방법론이 다양한 분류 문제에 효과적임을 의미하며, LLM 의 활용 범위를 확장하고 다양한 분류 문제에 대한 효과적인 해결책을 제공할 것으로 기대된다.

참 고 문 헌

[1] Github, https://github.com/e9t/nsmc

[2] Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T. and Wang, G., “Text classification via large language models”, arXiv preprint arXiv:2305.08377, 2023.

[3] Meta LLaMA Team. Introducing Meta Llama 3: The most capable openly available LLM to date. https://ai.meta.com/blog/meta-llama-3/, 2024.