BERT 기반 혐오성 텍스트 필터링 시스템 - 대학 청원 시스템을 중심으로

문태진, 배현빈, 이현수, 박상욱, 김영종

숭실대학교 소프트웨어학부

moontaijin322@gmail.com, baehynebin@gmail.com, quintuplets2000@gmail.com, cafemocha.dev@gmail.com

BERT-based Hateful Text Filtering System - Focused on University Petition System

Taejin Moon, Hynebin Bae, Hyunsu Lee, Sanguk Park, Youngjong Kim

School of Software, Soongsil University

요약

최근들어 청원 시스템은 사람들의 다양한 의견을 반영하고 대응하기 위한 중요한 수단으로 부상하고 있다. 그러나 많은 양의 청원 글들을 수작업으로 분류하는 것은 매우 시간이 많이 소요되며, 인적 오류가 발생할 수 있는 문제점이 존재한다. 이를 해결하기 위해 자연어처리(NLP) 기술을 활용한 청원 분류 시스템을 개발하는 것이 필요하다. 본 연구에서는 BERT(Bidirectional Encoder Representations from Transformers)[1]를 기반으로 한 텍스트 필터링 시스템을 제안한다. BERT 는 최근 자연어 분류 분야에서 상위 성능을 보이는 모델로, 이를 활용하여 청원 글을 분류하고 분류된 결과를 이용해 해당 글의 노출 여부를 결정한다. 본 논문에서는 BERT 모델의 이론적 배경과 구조, 그리고 미세 조정 학습 방법을 소개하고, 이를 활용하여 청원 분류 시스템을 구현하는 방법을 제시한다. 우리가 제안하는 BERT 기반의 텍스트 필터링 시스템은 청원 글 분류를 자동화하고, 이에 따른 대응 속도와 정확도를 향상시킬 것으로 기대된다. 또한, 이 시스템은 다양한 분야에서 응용 가능하며, 대용량 데이터 처리에도 적합하다. 이를 통해 대학 청원 시스템에서 혐오성 발언 등 부적절한 내용을 사전에 방지하고 학생들의 의견을 효율적으로 수집할 수 있는 기능을 제공할 수 있다는 장점을 가지고 있다.

1. 서론

최근 청원 시스템은 학생들이 학교 생활에서 직면하는 다양한 문제들을 해결하기 위한 중요한 수단으로 자리 잡고 있다. 이러한 청원 시스템은 학생들의 다양한 의견을 반영하여 대학의 정책 수립과 문제 해결에 큰 도움을 줄 수 있다. 하지만 청원 글의 수가 많아질수록 청원 글을 수작업으로 분류하는 것은 매우 어려워진다. 이는 시간과 인적 자원의 낭비를 초래할 뿐만 아니라, 오류 발생 가능성이 높아 대응 능력을 저하시키는 문제점을 안고 있다.

이러한 문제점을 해결하기 위해 자연어처리 기술을 활용하여 청원 글 분류 시스템을 개발하는 것이 필요하다. 특히 최근 BERT 모델이 자연어처리 분야에서 뛰어난 성능을 보이면서, 이를 활용한 청원 글 분류 시스템 개발이 주목받고 있다. 이러한 BERT 모델을 활용하여 대학 청원 시스템에서 청원 글을 분류하고, 혐오성 발언 등 부적절한 내용을 사전에 방지하고 학생들의 의견을 효율적으로 수집할 수 있는 청원 분류 시스템을 제안한다. 이를 통해 대학 청원 시스템의 효율성을 향상시키고, 학생들의 다양한 의견을 반영하여 대학의 발전에 기여할 수 있는 청원 시스템을 개발하는 것이 본 연구의 목적이다.

2. 시스템 구현

본 단락에서는 BERT 모델의 미세조정을 위한 데이터와 모델 구현, 전체 시스템의 구현을 설명한다.

2.1. 데이터 수집 및 전처리

본 연구에서는 청와대 국민 청원 사이트[2]에서 수집한 청원 데이터를 사용한다. 데이터는 csv 형식으로 수집한다. 수집한 데이터는 학습을 위해

각각 “hate”, “normal”, “meaningless”로 라벨링 한다. 각 라벨의 의미는 다음과 같다.

- **hate**: 혐오성 발언 내용
- **normal**: 일반적인 청원 내용
- **meaningless**: 청원과 어울리지 않는 내용

“혐오성”에 대한 분류 기준은 APEACH[3] 논문에서 사용된 것과 같이 2022 PYCON KR 의 행동 강령[4]을 따른다.

2.2. BERT 모델 Fine-Tuning

BERT 모델의 미세 조정은 Hugging Face 의 Transformers 라이브러리와 대학 커뮤니티용 기학습 언어 모델[5]을 활용하여 구현한다. 미세 조정에 필요한 하이퍼파라미터는 실험을 통해 최적값을 찾는다. 미세 조정을 마친 BERT 모델은 청원 글의 특징을 파악하고, 혐오성 발언 등 부적절한 내용을 사전에 필터링하는 데에 활용한다.

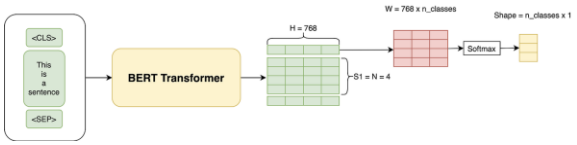

그림 1. 미세조정 된 BERT 모델 구조[6]

3.3. 청원 분류 시스템 구현

BERT 모델 미세 조정을 마친 후, 청원 글 처리 서버는 Oracle Cloud 와 SpringBoot, DB 는 MySQL, 모델 서빙 서버로는 MS Azure 을 사용하여 구현한다.

사용자가 청원 글을 입력하면 클라이언트에서 서버로 내용을 전달한다. SpringBoot 애플리케이션은 청원 글, 청원 동의에 관련된 CRUD 기능을 수행하고 MS AZURE 에 올라간 미세조정 모델을 활용하여 전달받은 글의 특징을 파악한다. 이 후, 특징을 바탕으로 분류된 내용이 부적절한 내용인지 판단하여 부적절한 내용이 포함된 글은 필터링하는 기능을 구현한다.

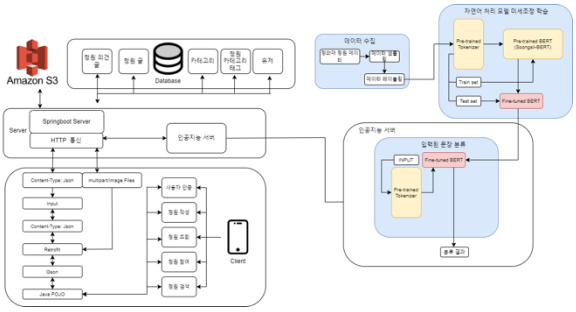

그림 2. System Architecture

4. 결론

본 연구에서는 BERT 기반의 혐오성 텍스트 필터링 시스템을 제안하였다. 제안된 시스템은 대학 청원 사이트에서 자동적으로 혐오성 발언을 필터링하여 대응 속도와 정확도를 향상시키고, 학생들이 보다 원활하게 대학과 소통할 수 있도록 도와줄 것으로 기대된다.

5. 참고문헌

[1] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[2] 청와대 청원 데이터. (2022). petition.csv. Retrieved from https://github.com/akngs/petitions (accessed March 10, 2023.). (Original Source: https://www1.president.go.kr/petitions?only=finished, The link is not valid after May 9, 2022.)

[3] Yang, Kichang, Wonjun Jang, and Won Ik Cho. "APEACH: Attacking Pejorative Expressions with Analysis on Crowd-Generated Hate Speech Evaluation Datasets." *arXiv preprint arXiv:2202.12459* (2022)

[4] PYCON Korea. Code of Conduct. Retrieved March 31. 2023. from https://2022.pycon.kr/coc

[5] jason9693/SoongsilBERT-base-beep. (2022). SoongsilBERT-base-beep. https://huggingface.co/jason9693/SoongsilBERT-base-beep

[6] James Montantes, "Softmax added BERT Architecture". https://miro.medium.com/v2/resize:fit:1400/0*zMYA0mVcM6-n_qT2. (accessed March 31, 2023.)