**BERT를 이용한 혐오 표현 분류**

1. **프로젝트 개요**

인터넷 커뮤니티에서 다양한 종류의 게시물이 포스팅되고, 각각의 포스팅에는 다른 유저의 의견이 추가됩니다. 건전하고 생산적인 의견이나 별다른 의미 없는 의견들도 있지만, 보기만해도 눈쌀이 찌푸려지는 혐오 표현들도 의견으로 추가된다..

그래서 서비스 운영자들은 사용자들의 신고를 받아 삭제처리를 하거나 사용자 신고가 누적되면 자동으로 삭제하는 방식으로 시스템을 운영하고 있습니다. 이와 같은 방식은 혐오 표현의 게시와 삭제 사이에 상당한 시간 간격이 있을 수 밖에 없고, 다른 사용자들은 불쾌감을 느끼게 된다..

따라서, 글을 게시할 때 게시되는 내용을 검사하고, 혐오 표현을 분류해 낼 수 있다면 관리자들은 더욱 생산적인 업무를 수행할 수 있고, 게시물을 보는 사용자들은 혐오 표현을 적게 볼 수 있으므로 이를 위한 혐오 표현 분류 시스템을 구현하고자 한다.

1. **프로젝트 목적**

해당 프로젝트의 데이터는 kaggle 대회에서 구글이 주최한 혐오 표현 분류 대회(jigsaw Rate severity of Toxic comments)에서 주어진 데이터 셋을 이용할 계획이다. 최종 프로젝트를 위해 참여한 대회는 구글이 4번째로 주최한 대회이고, 각 대회마다 데이터 셋의 형태가 다른 것이 특징이다.

첫 번째 대회의 데이터셋은 특정 문장에 대해 유해성(toxic), 심각한 유해성(severe toxic), 외설(obscense), 위협(threat), 모욕(insult) 및 정체성 혐오(identity hate)를 특징(feature)로 주어져서 혐오로 나누고, 작업자가 판단하여 각각에 대해 0 또는 1의 값을 설정한 데이터 셋이 주어졌었다.

테이블이(가) 표시된 사진

자동 생성된 설명

그러나, 혐오 표현이라는 것은 개개인의 도덕적 기준에 따라 정도를 다르게 느낄 수 있기 때문에 4번째 대회에서는 이를 보완하여 두 개의 문장을 주고, 어느 문장이 더 혐오스러운지 비교하도록 하여 많은 일관성이 있는 데이터 셋을 확보하고자 하였다.

테이블이(가) 표시된 사진

자동 생성된 설명

본 프로젝트에서는 다양한 종류의 BERT(RoBERTa, AlBERT, GroNLP, dehateBERT, hatexplainBERT) 등 다양한 모델을 이용하여 학습시키고, 학습이 완료된 모델을 이용하여 validation 데이터를 검증한 뒤, Score를 비교하여 가장 적절한 모델을 찾고자 한다.

1. **프로젝트 내용**

본 프로젝트에서는 다양한 종류의 BERT(RoBERTa, AlBERT, GroNLP, dehateBERT, hatexplainBERT) 등 다양한 모델을 이용하여 학습시키고, 학습이 완료된 모델을 이용하여 validation 데이터를 검증한 뒤, Score를 비교하여 가장 적절한 모델을 찾고자 한다.

1. **주요 기술**

- 언어 : Python

- 모델 : BERT

- 필요 패키지 : 실습 가이드 참조

- 개발 환경 : 주피터 노트북(w. kaggle)

1. **주요 소스 코드 및 시연방법**

- 캐글에서 사용한 코드의 일부이고, 시연방법은 스코어로 갈음하겠습니다.

**1. import**

텍스트이(가) 표시된 사진

자동 생성된 설명

**2. Preprocessing**

**텍스트이(가) 표시된 사진

자동 생성된 설명**

**2. Training**

**텍스트이(가) 표시된 사진

자동 생성된 설명**

**3. Validate**

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

**4. Inference**

텍스트이(가) 표시된 사진

자동 생성된 설명

**5. 최종 스코어**



1. **후기**

BERT의 사이즈가 크다 보이 RoBERTa의 경우 오버플로우가 심심치 않게 발생했고, AlBERT나 hatebert의 경우 오버 플로우는 발생하지 않았지만, 시간이 적지 않게 걸렸습니다. 향후 자연어 처리를 좀 더 잘 하게 되면, 특허 관련 서비스를 제작해보려고 합니다. 시간이 좀 걸리겠지만, 꼭 만들어보고 싶네요.

좋은 강의 해주셔서 감사드리고, 두달동안 고생하셨습니다.