#1
안녕하세요 저희는 비속어 포함 문장 판별 AI 모델 설계 및 구현이라는 주제로 프로젝트를 진행한 아이코조의 발표를 맡은 도윤서 권주명입니다.

#2
목차는 이러합니다. 서론, 기존연구, 프로젝트 과정, 결과, 결론 순으로 발표하겠습니다.

#3,4
혹시 이런 경험이 있으신가요? 이 사진은 리그오브레전드라는 게임의 채팅화면입니다. 게임을 하다보면 이런 욕설이나 인신공격과 같이 기분나쁜 말을 듣게 되기도 합니다. 실제로 저도 오버워치라는 게임을 하다가 플레이를 못한단 이유로 욕을 먹은 경험이 있습니다. 이렇듯 우리는 게임이나 인터넷 상의 댓글에서 어렵지 않게 욕설이나 도를 지나친 비윤리적 표현들을 찾아볼 수 있습니다.

#5
그래서 저희는 인공지능을 활용하여 이러한 문제를 해결할 수 있는 프로젝트를 진행하고자 하였습니다. 인공지능을 활용하여 욕설을 필터링하는 연구는 기존에 많이 진행되었지만, 저희는 거기에서 더 나아가 욕설은 *로 마스킹 처리되어 나오고, 욕설은 아니지만 비윤리적 표현이 포함된 문장이면 경고 문구를 띄우는 모델을 제작해보고자 하였습니다.

#6
다음으로 기존연구를 소개해드리겠습니다.

#7
저희는 먼저 '딥러닝 기반 비속어 필터링 채팅 프로그램 설계 및 구현'이라는 논문을 참고하였습니다. 이 논문에서는 악성 커뮤니티 사이트에서 댓글들을 크롤링하여 데이터를 형태소 단위와 자모 단위로 전처리 한 뒤

#8
CNN 모델을 이용해 데이터를 학습시켰습니다. 이 논문에서는 총 3개의 Convolution Layer와 3개의 FC Layer를 사용한 CNN 모델을 이용했습니다.

#9
그리고 Lime 알고리즘을 사용하여 input data를 어절 단위로 나누어 욕으로 판단된 어절을 '*'로 마스킹 처리하였습니다.

#10
다음으로 bert 모델에 관한 논문을 설명해드리겠습니다. bert 모델은 트랜스포머 모델 중 하나인데 이는 모델의 인코더와 디코더에서 어텐션 기법만을 허용한 모델입니다. bert는 이러한 트랜스포머의 인코더를 쌓아 올려 양방향 인코딩 표현을 사용하는 모델입니다. bert 모델은 양방향 인코딩을 사용하기 때문에 GPT와 같은 단방향 인코딩을 사용하는 모델과 달리 문맥을

고려할 수 있어 성능면에서 우수하다고 알려져 있습니다.

#11
Bert 모델의 학습 과정은 사진과 같이 pre training과 fine tuning으로 크게 두 단계로 나뉘어집니다. bert는 위키피디아(25억 단어)와 BooksCorpus(8억 단어)와 같은 레이블(label)이 없는 텍스트 데이터로 사전 훈련되어있습니다. 이 과정이 바로 pre-training 과정입니다. 그리고 다른 작업에 대해 추가 훈련과 함께 하이퍼 파라미터를 재조정하는 과정을 fine-tuning 과정이라 합니다. 저희는 이러한 논문들을 참고하여 최종적으로 CNN 모델과 bert 모델의 한국어 버전인 kobert 모델 두가지를 선택해 모델을 제작하였습니다.

#12
다음으로 프로젝트 과정을 설명드리겠습니다.

#13
저희가 사용한 데이터는 ai hub의 텍스트 윤리검증 데이터입니다. 이 데이터는 json 형식의 한국어 텍스트로 약 45만개의 문장으로 구성되어 있고, train, val, test 세트가 각각 8:1:1로 구분되어 있습니다.

#14
train, val, test 데이터셋 각각은 모두 화자, 해당 텍스트의 비윤리성, 비윤리성의 강도, 채팅, 텍스트의 비윤리 유형, 해당 텍스트 이렇게 총 여섯 개의 칼럼으로 구성되어 있습니다.

#15
이 중 비윤리 유형 칼럼을 보니 비윤리 문장이 아니라는 의미의 도덕/무도덕, 차별, 선정, 욕설, 폭력 이렇게 5개의 class로 나누어져 있었습니다.

#16
저희는 이러한 데이터를 목적에 맞게 활용하기 위해 데이터 전처리 과정을 거쳤습니다. 먼저 데이터를 dataframe 형식으로 사용하기 위해 엑셀에서 csv형식으로 변환하였습니다. 데이터 eda 통해 살펴본 결과 결측치는 존재하지 않아 결측치 처리는 따로 필요하지 않았습니다. 앞서 살펴본 5개의 class로 나누어진 비윤리 유형 칼럼은 라벨인코딩 후 int형으로 형변환 시켜주었습니다.

#17
저희는 욕설의 포함여부를 판별하는 이진분류 모델과 비윤리 유형을 판별하는 다중 분류 모델 두가지를 제작하였기 때문에 각각의 상황에 맞게 라벨인코딩을 진행하였습니다. 이진 분류의 경우에는 욕설만 1, 나머지는 모두 0으로 라벨 인코딩 하였고, 다중 분류의 경우에는 각 class를 0,1,2,3,4로 라벨인코딩 하였습니다.

#18
다음으로 모델링과 최적화 과정을 설명하겠습니다. cnn 이진분류 모델, cnn 다중분류 모델, kobert 이진분류 모델, kobert 다중분류 모델 순서로 설명해드리겠습니다.
먼저 model.summary를 통해 저희가 만든 CNN 모델의 구조를 확인해 보았습니다. 사진에서 볼 수 있듯이 모델에는 dense 12, dense 13 총 두 개의 dense 레이어가 생성되어 있고, 이외에도 embeding layer, convolution layer, max_pooling 레이어와 flatten 레이어가 존재합니다. 또한 마지막 레이어의 아웃풋 개수는 이진 분류이기 때문에 2개로 주어져 있는 것을 확인할 수 있었습니다.

#19
문장의 욕설 여부를 판단하는 CNN 이진 분류 모델을 학습하였습니다.
풀링 필터와 드롭아웃의 유무, learning_rate를 바꿔가며 모델을 학습한 결과 validation 정확도가 0.14 정도로 매우 낮게 측정되었습니다.

#20
따라서 저희는 activation함수를 softmax에서 sigmoid로, loss 함수를 binary crossentropy에서 mean_squared_error로 바꾸어 학습을 진행하였습니다. 그 결과 validation 정확도가 0.84로 대폭 상승하였습니다.

#21
다음으로 비윤리적 문장을 구분하는 CNN 다중 분류 모델을 학습하였습니다.
그리드 서치를 이용해 batch_size가 16, 32, 64, 128인 경우를 비교하였습니다.
그 결과 batch_size가 128인 경우 약 0.31로 가장 성능이 높음을 알 수 있었습니다.

#22
다음은 KoBERT 모델입니다.
화면 속 코드는 KoBERT 모델에 데이터를 넣기 위한 BERTDataset Class와 토큰화를 진행한 코드입니다.

#23
이 코드는 KoBERT 모델을 만들기 위한 BERTClassifier Class 코드입니다.

#24
이러한 코드를 바탕으로 욕설 여부를 판단하는 KoBERT 이진 분류 모델의 학습을 진행하였습니다.
모델의 성능에 큰 영향을 미친다고 판단한 batch_size와 learning_rate에 한해 하이퍼 파라미터 튜닝을 시도하였습니다.

#25
과대적합, 과소적합되지 않은 모델 중 test 정확도가 약 0.90으로 가장 성능이 높은

batch_size가 64, learning_rate가 5e-6인 모델을 최적의 학습환경이라고 판단하였습니다.

#26
다음으로 텍스트의 비윤리적 유형을 구분하는 KoBERT 다중 분류 모델의 학습을 진행하였습니다.
이진분류와 같은 이유로 batch_size와 learning_rate에 한하여 하이퍼 파라미터 튜닝을 시도했습니다.

#27
그 결과 과대적합, 과소적합 되지 않은 모델 중 0.68로 가장 성능이 높은 batch_size가 128, learning_rate가 5e-6인 조건을 최적의 학습환경이라 판단하였습니다.

#28
프로젝트의 결과에 대해 말씀드리겠습니다.

#29
각 모델의 성능을 비교한 결과, 이진 분류의 경우 KoBERT 모델이 0.91로 성능이 가장 좋았습니다. 다중 분류 또한 KoBERT 모델이 0.70으로 성능이 가장 높았습니다. 따라서 저희는 KoBERT 모델을 활용하여 예측을 시도하였습니다.

#30
우선 욕설과 비윤리적 텍스트를 언급하는 것에 대해 양해부탁드립니다.
KoBERT 이진 분류 모델로 직접 문장을 입력해 욕설 포함 여부를 예측한 결과,
개나가리색기, 개나라색기야는 욕설이 포함된 문장으로, 새끼를 꼬다, 새끼를 낳다는 욕설이 포함되지 않은 문장으로 구분하였습니다. 이를 통해 특정 단어를 보고 욕설이라 판단하는 것이 아닌 문맥을 고려해 예측함을 알 수 있었습니다.

#32
KoBERT 이진 분류 모델의 예측에 성공한 결과를 정리해보았습니다.
개뇬과 캐논, ㅂ1ㅅ과 방수, 및힌과 마침내 또한 잘 구분해 예측했습니다.

#33
하지만 시바견, 시발점, 슈바이처를 욕설로 예측하고, 졸라, 띠바를 욕설이 아니라 예측하는 오류가 있었습니다. 예측에 실패한 원인을 찾기 위해 예측 실패 단어가 포함된 문장으로 다시 예측을 시도하였습니다. 그 결과 내 강아지는 시바견이다, 졸라 맛있다는 예측에 성공하였으나 나머지 문장은 여전히 예측에 실패하였습니다.

#34
다음으로 KoBRET 다중 분류 모델을 이용해 문장의 비윤리적 유형을 예측해보았습니다.
test dataset의 문장을 이용해 테스트를 진행하였습니다.
너 머리가 왜그러냐, 남녀평등시대라서 그런거 없어 그런게 어딨냐의 예측에 실패하였으나 나

머지 문장은 성공적으로 예측하였습니다. 비윤리적 표현이 포함된 문장이면 아래와 같이 어떤 유형의 비윤리적 표현이 포함되어 있는지 경고 문구를 출력하도록 하였습니다.

#35
이번 프로젝트에서 LIME 알고리즘을 이용해 욕설을 '*'로 마스킹 처리하는 것을 구현하고자 했습니다.
하지만 scikit-learn의 모델은 predict_proba 함수로 예측할 수 있었으나 KoBERT 모델에서는 predict하는 함수가 존재하지 않아 새로운 함수를 선언해 예측 모델을 만들어 주어야 했습니다. 하지만 함수 선언 과정에서 차질이 생겼고, 이를 해결하려 하였으나 시간이 부족하여 LIME 알고리즘을 구현해내지 못했다는 한계가 있었습니다.

#36
이번 프로젝트의 모델의 성능을 향상시키기 위해서는 더 많은 데이터를 학습시키고, 단일 모델만 사용하는 것이 아닌 특정 모델의 한계를 극복할 수 있도록 다른 모델과 결합하는 방법을 생각해 보았습니다.
또한 설명 가능한 인공지능(XAI)를 이용하여 모델이 어느 부분에서 데이터를 잘못 판단하게 되었는지 확인 후 보강하는 과정을 거쳐야 할 것입니다.

#37
마지막으로 프로젝트의 결론에 대해 말씀드리겠습니다.

#38
이 프로젝트에서는 CNN과 KoBERT 모델을 이용한 욕설과 비윤리적 포함 문장을 판별해주는 알고리즘을 구현하였습니다. 이는 욕설 뿐만 아니라 비윤리적 표현가 포함된 문장까지 거를 수 있어 깨끗한 인터넷 환경을 만들어 나가는 데 활용될 수 있습니다. 실제 데이터를 기반으로 수업에서 배운 CNN 모델에서 나아가 KoBERT 모델까지 구현해보았다는 데에 의의를 두고 있습니다. 저희가 만든 모델에 직접 문장을 넣어 예측해보는 과정에서 인공지능에 대한 재미 또한 느낄 수 있었습니다.

#39
(3초 기다리기)

#40
역할 분담은 이와 같습니다.

#41
참고문헌도 이와 같습니다.

#42
이상으로 발표를 마치도록 하겠습니다. 감사합니다.

#43
질문 받아여~~