<YBIGTA 프로젝트 진행 보고서>

기수 : 23
학과 : 불어불문학과
학번 : 2020114010
이름 : 이성현

해당 보고서는 개인의 자율에 따라 작성하며, 개인 포트폴리오를 겸용하는 역할을 가집니다.

프로젝트	Ybigta 23기 신입 기수 프로젝트
주제	게임 커뮤니티 욕설 필터링 서비스
팀원	곽민규, 김현호, 정회수, 이세진(팀장), 이성현
프로젝트 의도	게임 커뮤니티의 댓글에 나오는 욕설을 필터링하는 댓글을 만듦으로써 청렴하고 정중한 커뮤니티 문화 양성에 일조할 수 있도록 기여.
역할 분배	이성현 : 데이터셋 정리, 데이터셋 전처리, 모델 후보군 제안, 모델 최적화(w. random seed), 회의 서기, 회의 간 의견 제시.
개인 역할 세부	1. 데이터셋 1-1. 선정 : 기존 데이터셋을 찾아둠. 1-2. 전처리 : malicious 데이터셋 전처리(정확한 명칭은 잘 모르겠음.) 2. 모델 2-1. 선정 : 우선 무엇을 기반으로 한 모델인지 파악하는 과정이 우선 필요했음. SVM, 사전 기반, transformer 정도의 세 가지 후보군을 제시하였고, 이전에 진행한 회의와 이번의 상의를 통해서 transformer 기반의 모델을 선정하는 것을 목표로 하였음. 이후 다양한 딥러닝 관련 라이브러리를 가지고 있는 HuggingFace에서 task에 가장 맞는 모델을 추렸고, 이에 대하여 나는 Soongsil-BERT라는 모델을 제안함. 제안된 세 가지 모델 중 해당 모델은 1순위는 아니었고, 대신 KoBERT라는 모델을 우선 사용하기로 팀원끼리 결정함. 2-2. 모델 구현 : 모델 구현은 별도로 맡은 역할은 거의 없음. 성능 최적화 과정에서 Random Seed를 설정하지 않았다는 점을 발견하여 random seed를 설정하는 버전으로 모델을 바꿨음. 2-3. 성능 최적화(Optimizing) : 스프레드시트에 모델의 성능을 최적화할 수 있도록, 주어진 hyperparameter, 그리고 이를 가독성 있게 정리함. 처음에 제안된 스프레드시트 양식은 하나의 시도에 대해서 어떠한 성과가 나왔는지를 자세히 파악할 수는 있으나 전체 성능을 파악하고자 할 때는 문제점이 있다고 판단하여 같은 파일에 새로운 시트를 만들어서 가독성을 개선함. 특히 random seed를 찾는 과정 때문에 성능만 빠르게 확인하고 넘어가야 하는 부분들이 있었는데, 가독성이 떨어질 수 있겠다는 우려를 먼저 파악하여 만든 덕분에 팀원들과 함께 모델 성능 개선을 원활히 수행할 수 있었음. 모델을 돌리는 과정에서는 함께 모델 최적화를 담당한 주 구성원 세 명(곽민규, 이세진, 이성현) 가운데 가장 좋은 성능을 보여줬음. 이는 운도 있었다고 생각함. 하지만 그와 별개로 모델에 관한 이해도, 그리고 최적점을 찾아가는 접근 방식은 다른 사람들에 비해서 좋은 편이었던 것이 좋은 성능의 모델을 찾는 데 일조하였다고 생각함. 현재 작동시키고 있는 모델은 random seed =

	13으로 된 모델이고, validation accuracy가 0.9522 (9/12 epoch 기준)의 성능을 보이고 있음. random seed를 정하기 이전에도 좋은 성능을 보이는 모델을 발견하였으며, 여기서 random seed를 찾아가는 과정을 거쳤음. 3. 프레젠테이션 3-1. 중간 발표 : ppt 제작은 정회수 님이 맡아주시고, 발표의 핵심 내용과 주된 흐름은 본인이 작성함. 그리고 중간 발표도 본인이 맡아서 발표를 진행함. 대본을 따로 두지 않고 발표했기 때문에 중간에 발음이 씹히기는 했지만, 필요한 내용을 전반적으로는 매끄럽게 진행함. 3-2 발표 자료 검수 : 중간 기말 발표 모두 정회수 님이 ppt 제작을 맡아주셨는데, 발표 흐름을 다루기 위해서 ppt를 검수하였다. 나 말고도 다른 분들께서도 의견을 주셨지만, 나랑 세진 님이 아무래도 가장 많은 의견을 낸 사람 두 명이다. 3-2. 최종 발표 : 전체 발표 흐름을 계획함. 4. 기타 4-1. 회의 내용 정리 및 서기 : 회의를 참고해야 하는 팀원이 있는 경우, 그리고 개인적인 포트폴리오 작성이 필요한 상황을 고려해서 작성함. 회의 내용 전반에 있어서 주고받은 대화를 최대한 기록함. 4-2. 노션 생성 및 파일 정리 : 프로젝트 진행에 필요한 세부 프로젝트를 노션에 생성하고, 관련된 파일과 내용을 git, 노션 등에 정리. 4-3. 팀원 커뮤니케이션 약속 시간 선정 조율. 개별 프로젝트 관련 의견 확인/공유/제언.
프로젝트 진행 과정	날짜 순서
	주제 선정 회의를 통한 방향성 설정
	프로젝트 진행 순서
	① 데이터셋 및 모델 후보군 방향성 설정 및 결정 ② 중간발표 ③ 주제 재설정 ④ 데이터셋 전처리 및 통합 ⑤ 모델 구현 ⑥ 모델 최적화, 인벤 데이터 크롤링 ⑦ 프론트 구현 ⑧ BERT 모델과 인벤 크롤링 모델을 프론트에 연결 ⑨ 최종 발표
회의 과정 및 내용	**1차 회의 (23.08.06)** - 수집할 데이터 및 그 데이터의 특성이 어떠한 것이 좋을 것인가? - 카카오톡 대화 내역을 활용하거나 기존 데이터를 활용하면 좋을 듯. → 기존 데이터 수집(프로젝트 기간이 부족한 점을 고려) - 어떤 종류의 모델을 구현할 것인가? - 사전 기반 vs 딥러닝 기반

	- 사전 기반 : 일상에서 쓰이지 않는 표현들도 들어갈 수 있음. - 딥러닝 기반 : 맥락을 통해서 욕설 여부를 파악할 수 있음. → 딥러닝 기반 모델로 선정. transformer를 활용해보는 것은 향후 프로젝트에도 도움이 될 것이고, 맥락이 중요한 것도 맞음. - 각자 할애 가능한 시간은? - 곽민규(DS) : 9~13시 대외 활동. 그 외에는 시간 있음. - 김현호(DE) : 평일 출근으로 기본 일과시간은 불가. 8.14부터 평일 저녁은 시간 X. - 정회수(DS) : 시간 여유 있는 편. (데이터 finetuning에 할애할 시간이 있음.) - 이세진(DS) : 석사 생활 중. 큰 프로젝트 진행 중. 중간발표 참여가 어려움. 교수님게 발표할 것이 있어서 참여가 어려움. 5일 내내 학교에 계심. - 이성현(DS) : 이번주 월~금 13~17시 불가. 이번주는 프젝에 할애할 시간이 2일 정도 있음. - 전체 워크로드를 어떻게 설정할 것인가? - 중간발표까지 목표 : 어떤 모델, 어떤 데이터로 어떤 일을 할 것인가에 대해 답해야 함. - 전체 워크로드 : 데이터셋 설정 및 전처리 → 모델 설정 → 기본 인터페이스 설정 - 다음 회의 전까지의 역할 - 어떤 데이터셋을 고를 것인가? - 고른 데이터셋의 수정이 필요하다면 어떻게 할 것인가? - 어떤 모델을 설정할 것인가?
	3차 회의 (23.08.09) - 모델 선정 - 후보군 : KcELECTRA-BERT(v3), Soongsil-BERT, KcBERT - → KcELECTRA-BERT를 우선 선정. 일반적으로 좋은 성능을 보이고, 폭넓게 쓰이며, 적은 데이터로도 학습 성능이 높기 때문. → 이번 task에 대해서도 좋은 성능을 보일 가능성이 있음. # 시간적인 여유가 있다면 Soongsil BERT도 같이 돌려서 성능 검증하기 (모델 성능 비교 작업). - 데이터의 양/질 개선 방안 - 현재 데이터의 단점은 게임이 아닌 커뮤니티나 뉴스 기사의 댓글을 크롤링한 형태. → 학습에 필요한 양이 부족할 수 있고, 데이터의 퀄리티가 부족할 수 있음. - 인벤 데이터 수합해서 모으고 싶은 특정 욕설이 포함된 문장들을 추가해서 성능 개선을 도모(ex - Tlqkf이라는 단어를 모델이 잘 판단하지 못하니까 해당 단어가 들어간 데이터를 추가하는 방식) - 중간 발표 계획 - ppt 제작은 최대한 단순하게. 발표자는 이성현. 관련 계획 및 내용은 노

	션의 '중간발표' 참고(https://www.notion.so/59552e2cd9844a86aa4b84091cc00317 https://www.notion.so/59552e2cd9844a86aa4b84091cc00317?pvs=21) - 내일 계획 - 금요일 5쯤부터 만나서 같이 데이터 전처리 작업 진행(대면으로 진행하지만, 개인 사정 고려해서 일부 비대면 진행 가능)
	4차 회의 (23.08.11) - 데이터 전처리 작업 진행 - 정회수 : selecstar - 곽민규 : curse detection data - 이성현 : korean hate speach - 데이터셋 파일 통합 - 곽민규님 또는 정회수님 담당
	5차 회의 (23.08.14) 1. 현재 모델 성능 등에 대한 문제점 확인 KoBERT로 할 수 있는 것은 필터링이 아닌 분류. BERT에서 '자지마', '느금마', '전염병' 등과 같은 단어들이 과적합 문제 등으로 인해서 올바르게 판별되지 않는 문제 발견. (앞 내용 제대로 못 들음.) 프론트로 구현하는 것이 어떻겠는지? => 욕설의 필터링이 아니라 욕설 감지로 하는 것은 어떠한지? (할거면 labeling을 일일이 해줘야 하는데, 많은 데이터에 대해서 적절히 해결하기는 힘듦.) Q. 데이터 크롤링으로 할 수 있는 것은 무엇인지? 욕설이라고 판단되면 댓글 전체를 블라인드 처리해서 욕설인지 감지. -> '해당 댓글이 욕설 데이터입니다. 블라인드 처리하겠습니까?' 라고 묻는 식으로 프론트 구현 1. 필터링을 어떻게 할 것인가? <현재는 불가한 것으로. 분류 모델부터 확실히 만든 다음에 따지는 것으로 판단하자! 안되는 이유는 게임회사들에서도 거대한 데이터베이스를 활용해서 필터링을 하고 있고, 이를 우리가 가지고 있지도, 직접 만들 수도 없기 때문에 불가한 것으로 판단.> 2. 현재의 방향성 (웹 페이지 구현과 연관지어서) 인터넷 커뮤니티나 댓글 데이터를 실시간/정기적으로 크롤링해서 그 댓글/글에 욕설이 있는지 없는지 판단하여 필터링 작업을 거침(네이버 댓글의 클린봇을 생각하면 됨). 이를 활용해서 하나의 댓글 전체를 필터링하는 방안. 프론트로 코드를 실행하는 것보다 파이썬으로 짠 코드를 구현하는 것이 더 좋음. 화면에서는 인벤을 여러 개로 나누는 편이 좋을 것 같음. 그리고 목록을 만들어서 원하는 인벤을 웹사이트 사용자가 직접 선정할 수 있게 하는 방안도 고

	려 중. 3. 팀 과제 : Optimization (팀원 전체) / 크롤링 (정회수님, 이성현님) / 프론트 (김현호님) 최적화 : 8.18(금) 자정까지 최적의 모델 성능 구현. 팀원 전체가 같은 모델과 데이터를 활용해서 성능을 점검. Github나 노션 등 다 같이 공유 가능한 곳에서 hyper parameter를 다르게 했을 때 나오는 성능들을 정리하는 것이 어떨지 확인. 스프레드 시트에 모델과 관련한 hyperparameter를 사전에 작성한 후, 모델의 성능을 작성해서 성능을 점검. optimizer, activation function, epoch, lr, regularizer, min_lr, total_rate, 엡실론, step, layer 및 input/output number 등을 활용해서 최적의 모델을 찾아낼 것. 크롤링 : 8.17(목) 자정까지 작업 완료 (for 프론트 작업). 인벤 사이트 데이터 가져오는 csv 파일로 만들기 + 다시 리스트로 부르는 파일 만들기. 프론트 : 웹사이트 구현 관련은 목요일까지 할 예정. 1. 다음 회의 : 목요일 14:30에 회의 참여. 회의 전까지 목표 : 모델 성능 확인 및 크롤링 작업 최대한 완성. 프론트는 1차적인 수준에서의 완성도 확인
	6차 회의 (23.08.17) - Optimizing : dense를 깊이 하면 더 성능이 좋아지지 않을까? 현재 스프레트 시트 250행의 모델(from 이성현) 이 validation_accuracy, model accuracy를 고려할 때 좋은 성능을 보였음. 이세진 : 대체로 모델들의 성능이 별로 안 좋게 나왔고, 0.6290 수준에서 정확도가 멈춘 경우가 많았음. 초기 설정은 overfitting 현상이 발생한 상태에서 validation accuracy의 성능이 비교적 낮게 (0.7~0.8) 나왔었는데, 이보다 좋은 성능의 모델을 찾기는 힘들었음. 곽민규 : colab과 local environment 활용해보니 모델 성능이 유의미하게 좋은 것은 없었음(0.6290에서 멈췄음). -> 향후 계획 : 현재 250행 모델을 기본 베이스로 두고 모델 성능 개선에 초점을 둘 계획. - Crawling : 현재 크롤링은 완성된 상태. 메서드 하나 호출하는 방식으로 실행하고 싶다면 셀 하나에 라이브러리 import 가능하도록 함. 작동 방식 : 이중for문을 활용하여 링크에 맞는 곳에 접속하여 그에 해당하는 게시글과 그에 맞는 텍스트들을 가져오도록 설정. 관련 데이터를 불러옴. query 함수를 활용하여 원하는 데이터를 확인할 수 있음. 확인에 다소 제한사항이 있었음. 게임 점유율에 비해서 인벤에서의 활동이 활성화되지 않은 경우가 있음. 이러한 경우를 고려하여 롤, 메이플, 피파4, 디아블로, 로아를 선정하여 데이터를 불러오도록 함. - 댓글을 가져오는 방식을 있는대로 많이 가져오기(not 베스트 댓글). 그러면 댓글을 가져오는 개수의 상한선을 정해서 한 게시글에서 가져오는 댓글의 수를 과도하게 많지 않도록 하는 것이 좋을 것임.

	- 동적 크롤링의 소요 시간이 긴 편이니까 정적 크롤링 방식을 더 활용한 코드가 있으면 시간소요를 줄일 수 있을 것으로 보임. Q. 크롤링 작업을 실시간으로 가져오게 하고 싶은 것인지 아니면 한 번만 가져오는 것으로 끝내고 싶은 것인지? A. 데이터 가져오는 작업의 시간이 오래 걸린다는 문제점이 있음. 현재 고려하는 해결책은 댓글을 정해진 숫자만큼 가져오면 break를 하는 방식. time.sleep(5)로 해둔 상황인데, 원하는 객체가 로드 될 때까지만 기다리는 함수가 있어서 이걸 활용할 계획. 안 되면 sleep 시간을 줄일 계획. - Front : 현재 초안이 나온 상황. 욕설 클린봇 켜기/끄기 모델이 나온 상황. 욕설 클린봇 켜기/끄기는 5개 사이트에 대해서 따로는 안 되고 동시에 되는 편. 데이터를 불러오는 것에 관한 시간의 문제가 있는데, 이는 불러오는 데이터의 양을 조절하는 방식으로 해결 가능할 것으로 예상됨(사이트 별로 30/50개씩 총 150/250). 현재 문제점 : 시간 부족. 성능 체크가 가능한 시간이 어느 정도 필요해서 Q. 주제가 좀 변경된 편인데, 나중에 발표할 때를 상상해볼 때 이 모델이 가진 잠재성이나 활용방안이 바뀔 것 같음. 어떻게 하는 것이 좋을지? A. threshold로 나오는 결과. 여기서 성능이 잘 나온다면 맥락을 고려했다는 점에서 의의가 있다고 따질 수 있을 것. 시도에 의의가 있을 것임. 당장 생각하기에는 한계가 있을 것으로 생각함. 다음 계획 : 8월 20일 (일) 한 차례 더 회의. 약속시간은 토요일에 결정. 8월 21일에 이세진님 ppt 제작. ppt 탬플릿은 추천하는 것이 있으면 팀원들이 제안해주기. 1. 모델 활용 방안에 대하여 2가지 정도 가능성을 제시해주기. 2. 각자의 파트(모델 성능 개선, 크롤링 개선, 프론트 구현)를 최대한 완성 단계로 해서 브리핑. 3. ppt 발표와 관련해서 필요하다고 생각하는 내용들 최종 정리. 4. ppt 탬플릿 제안(회의 때 또는 그 이전에) 5. 노션에 관련 진행 상황 작성.
	7차 회의 (23.08.20) 크롤링 : 크롤링 관련 변동사항. 속도가 느린 것이 현재 단점임. 정적 크롤링으로 속도를 높이고자 하였음. 어쩔 수 없이 동적 크롤링으로만 데이터를 가져올 수 있는 부분들이 있어서 동적 크롤링도 혼용하여 사용함. 불필요한 option들을 적용하면 모델의 속도를 키움. time.sleep()함수를 하용하는 대신 EC 라이브러리를 활용하여 소요 시간을 줄임(찾는 XPATH가 로드될 때까지만 대기시키는 방식). + 멀티 thred 버전도 있는데, 한 번에 여러 사이트를 병렬적으로 접근할 수 있도록 하는 방식. 다만 로컬 컴퓨터에서 실행해본 경험에서는 성능이 별로 안 좋게 나옴. version 2는 전부 1분 이상이 걸리는 단점이 있었지만, single thred로 데이터를 크롤링하는 version 3이 현재 나온 버전 중에서 가장 좋은 성능을 보였음.

	– 시간 단축의 한계가 있는 이유는? XPATH로 댓글을 가져오는 방식을 채택해야 하는데, 댓글 코멘트마다 XPATH의 임의 번호들이 있음. 번호가 임의로 나오는 문제가 제대로 가져오지 못함. random sampling으로 텍스트 변환하는 방식도 사용하였는데, text 변환 과정에서 발생하는 시간 소요를 줄이고자 3개마다 바꾸는 방식을 채택함. – 기타 모델에 관한 설명 댓글 중에서 이미지 데이터는 그냥 패스, 띄어쓰기나 줄띄우기처럼 특수문자가 되는 문자들('\n, \t, 등')을 재처리(' ')함. 모델 최적화 : 모델을 h5 파일로 전환하는 작업을 거침. 파일을 github에 첨부해야 하는데 문제가 발생하는 중. 이대로 안 되면 구글드라이브를 통해 공유할 예정. Q. 모델 최종본은 언제 결정되는지? 욕설을 판별하는 기준치만 결정하면 확정하면 될 듯. drive 폴더로 따로 코드를 따로 저장해야 하는 상황. –> 내일 저녁 6시까지 모델을 주기를 권장함. – 상위폴더 –> content –> 원하는 모델 있음. 모델 20 epoch으로 돌려서 완성하면 될 듯. – 프론트 : 프론트 관련된 부분은 개인 일정 등을 고려하여 월요일 저녁에 작성하는 것이 마지막으로 할 수 있는 작업. – 최종 발표 : 각자 맡은 분야는 각자가 작성(회의 중에 완성하는 것이 목표). ppt 제작은 정회수님이 제작. 발표는 이세진님. 발표 시간은 15분 내외니까 ppt 자료를 확인해보고 그에 맞게 불필요한 내용을 제외하면 좋음. 발표전까지 준비사항 – ppt는 내일 22시까지 완성해보기. – VSCode와 연동하여 실행되는지 확인할 것.
프로젝트로 성장한 점	1. 팀 프로젝트를 전반적으로 진행해본 경험 그 자체. 내 기억이 맞자면 팀 프로젝트를 제대로 힘줘서 진행하는 것은 조직행동론 수업 이후로 처음이다. 특히 내가 해보고 싶었던 AI 분야에서의 프로젝트를 이렇게 괜찮은 완성도 속에서 보여준 것은 괄목할 성장이라고 생각한다. 당장 개강하는 2학기에도 '딥러닝의 이론과 실제'라는 수업을 통해서 팀 프로젝트를 진행할 예정이고 YBIGTA 학회를 통해서도 프로젝트를 진행할 예정이다. 팀 프로젝트를 진행할 때 더 효율적인 과정을 거칠 수 있을 것 같다. 2. DS 팀원으로서 생각해야 할 요소들의 습득.

	DS 팀원으로서 모델을 선정하고, 설계하고, task에 맞게 최적화하고, 데이터를 다루는 일까지, 다양한 업무를 다루었다. 일부 일들은 다른 팀원분(이세진 님)이 맡아주셨는데, 어쨌든 같은 팀에서 하는 일들을 구경하면서 여러 일을 생각해보는 계기가 되었다. 모델을 다루는 과정에서 내가 배경지식이 있어서 잘 해결한 부분들도 있었지만, random seed를 제대로 따지지 못한 것이 아쉬운 점이었다. 그리고 RectifiedAdam과 같은 모델을 내가 다뤄놓고서, 막상 각각의 hyperparameter가 무슨 역할을 하는지 100% 이해하지 못한 부분도 있었다. 챗지피티로 내가 hyperparameter를 공부했었지만, 이러한 지식의 빈틈은 빠르게 채워나가는 것이 좋을 것 같다. 3. DS팀의 멤버가 실제 프로젝트에서 하는 일. 프로젝트를 하니까 시간 자체는 모델 최적화에서 굉장히 긴 시간을 할애하는 것 같았다. 내가 모델을 최적화하는 데 들인 자원이 제일 많았었는데, 이 시간을 줄이는 것이 팀플레이에서 중요한 요소가 될 것 같다. 이번에는 다행히 내가 아주 좋은 성능을 내는 지점을 발견했지만, 앞으로는 또 어떻게 바뀔지 알기 힘들다. 이번에 최적점을 찾아낸 일을 기분 좋게 여길 수는 있겠지만 이게 상당한 운도 있었다는 점을 꼭 유념하면서 모델의 최적화 과정을 거쳐야 할 것 같다.