[결과 보고 발표 대본 – 약 10분]

발표 시간: 10분

[슬라이드 1: 발표 시작]

안녕하세요. 이번 프로젝트의 모델링 개발을 담당한 박거량입니다.

지금부터 LLM 을 활용한 보이스피싱 예방 웹 어플리케이션 프로젝트의 결과 보고를 발표하겠습니다.

[슬라이드 2: 목차]

발표는 프로젝트의 개요, 아키텍처 소개, 데이터 선정, 모델링, 프로젝트 진행 중 발생한 의문점 그리고 배포 웹사이트 시연으로 구성됩니다.

[슬라이드 3: 프로젝트 개요]

웹 어플리케이션은 총 3가지의 기능으로 구성되어 있습니다.

보이스피싱 판별 페이지에서는 사용자가 통화 음성 파일 입력 시 통화 내용이 보이스피싱일 확률을 알려줍니다.

롤플레잉 페이지에서는 사용자가 선택한 보이스피싱 상황에 대해 챗봇과 롤플레잉을 할 수 있고, 대화 종료 시에 사용자의 대응에 대한 피드백을 제공합니다.

대처방안 페이지에서는 사용자가 처한 여러 보이스피싱 피해 상황에 대해 대처방안을 제공합니다.

[슬라이드 4: 아키텍처 소개]

먼저 모델링 아키텍처에 대해 소개해드리겠습니다.

KoBERT 모델은 유저의 인풋을 받아 토크나이저를 통해 토큰화합니다. 인코더에서 각 토큰을 임베딩 처리하고 벡터로 변환합니다. Classification Head에서 인코더의 출력을 받아 이를 기반으로 Sigmoid를 통해 확률 값을 출력합니다.

KoLLaMA 모델은 유저의 인풋을 받아 이를 랭체인을 통해 전처리하고 임베딩 변환을 거쳐 RAG 기반 문서 검색을 합니다. 검색된 문서를 참고하여 QLoRA로 파인튜닝된 LLM이 응답을 생성하고 최종적으로 랭체인을 통해 가독성을 높여 사용자에게 전달합니다.

프론트엔드는 리액트를 사용하여 비동기적 통신을 했으며 반응형 UI와 코드 모듈화로 유연성을 확보했습니다.

백엔드는 기본적으로 python기반의 웹프레임 워크를 사용하여 django와 fastapi를 사용했습니다.

여기서 문제는 모델이 각각 파이썬 환경이 달랐던 건데요.

kollama는 파이썬 3.11버전으로 django 프레임워크에서 챗봇 및 유저 로그인 기능을 구현했고,

kobert는 3.8 환경에서 fastapi를 이용해 통화음성파일 분석 기능을 구현했습니다.

이러한 환경을 같이 배포하기 위해 쿠버네티스를 이용했고, 음성파일의 경우 azure blob스토리지에 저장했습니다.

[슬라이드 5: 아키텍처 소개]

전체적인 아키텍처에 대해 설명드리겠습니다. 저희 시스템은 Azure Kubernetes Service(AKS)를 기반으로 동작하며, 프론트엔드는 React로, 백엔드는 FastAPI와 Django로 구성되어 있습니다. 그리고 데이터 처리를 위해 Azure Data Lake Storage를 활용하고 있습니다.

[슬라이드 6: 데이터 소개]

모델 학습에 사용된 데이터를 소개해드리겠습니다.

KoBERT 모델은 609개의 보이스피싱 대화와 609개의 일상 대화로 구성된 데이터셋을 사용했습니다. 정규표현식 텍스트 정제, 불용어 처리 등의 전처리와 토큰화를 하여 학습에 사용했습니다.

KoLLaMA 모델은 system, assistant, user의 역할로 구성된 23개의 보이스피싱 시나리오 데이터셋을 추가학습시켰습니다. 데이터셋의 크기가 작기 때문에 epoch와 learning rate를 조절하여 학습했습니다.

KoLLaMA가 응답을 생성할 때 참고할 수 있게 RAG를 사용하였고, 롤플레잉에는 KBS 사회부가 취재 과정에서 입수한 실제 보이스피싱 시나리오를, 대응방안을 제공할 때에는 금융감독원에서 제공하는 보이스피싱 예방요령 자료를 사용했습니다.

[슬라이드 7: 모델링]

다음은 모델링에 대해 설명드리겠습니다.

KoBERT로 수행할 수 있는 task가 여러 가지이기 때문에 task에 맞추어 구현해야 합니다. 보이스피싱과 일상 대화를 분류하는 task를 수행하기 위해서 BertClassifier를 구현했습니다. 보이스피싱인지 아닌지 분류하는 이진분류이기 때문에 로스 펑션은 바이너리 크로스 엔트로피 with logits loss로 설정했습니다.

사전학습된 KoBERT 모델에 정규표현식 텍스트 정제, 불용어 처리, 토큰화 등의 전처리를 한 데이터셋을 추가학습하여 사용했습니다.

학습에 사용된 데이터셋 내의 대화가 아닌 직접 제작한 통화 녹음 파일과 금융감독원의 보이스피싱 녹음 파일을 활용하여 예측값을 비교했습니다. 오른쪽 그래프가 learning rate와 epoch에 대한 대화의 보이스피싱 확률 그래프입니다.

Learning rate 2 * 10 ^ -5, epoch 3인 경우에 보이스피싱 대화가 보이스피싱일 확률 96.27, 일상 대화가 보이스피싱일 확률 13.00으로 다른 경우에 비해 과적합이 덜하고 학습이 잘 된 것으로 판단되어서 파라미터로 선정했습니다. Max length 64, batch size 16 등의 파라미터는 gpu 메모리에 맞춰 선정했습니다.

모델을 저장할 때는 state dict 형식으로 저장하여 가중치만 KoBERT 모델에 업데이트하는 방식으로 사용했습니다.

[슬라이드 8: 모델링]

KoLLaMA 모델은 LLaMA-3-Korean-8B 모델을 사용하였는데 여기서 8B가 80억개의 파라미터를 가진 모델이라는 의미라서 모델 크기가 굉장히 큽니다. 이렇게 큰 모델을 사용하기 위해 QLoRA 방식을 사용했습니다. LLaMA-3-Korean-8B 모델을 4비트로 양자화하여 로드했고, 모델 전체 layer의 가중치를 업데이트하는 것이 아니라 LoRA 어댑터 부분의 가중치만 업데이트하였습니다. 이를 통해 GPU 메모리 사용량을 낮추고 학습 속도를 향상시켰습니다.

LangChain을 통해 프롬프트로 모델에게 보이스피싱 롤플레잉, 롤플레잉에 대한 피드백, 보이스피싱 피해 상황에 대한 대응 방법을 제공하도록 설계했고, 모델이 응답 제공 시 참고할 수 있도록 RAG를 사용했습니다.

학습에 사용되는 데이터셋의 크기가 작기 때문에 Supervised FineTuning Trainer를 사용했고 learning rate 10^-5, epoch 10으로 설정했습니다. LoRA에 대한 파라미터는 rank 4, alpha 16, dropout 0.1로 설정했습니다.

[슬라이드 9: 백엔드 채팅 아키텍처]

백엔드의 핵심기능인 채팅의 구동방식에 대해서 발표드리겠습니다.

기본적인 환경은 Django의 동기식 서버 구동방식이였지만,

비동기적이고 빠른 챗봇구현을 위해 Django의 channels 라이브러리를 설치하였고, daphne를 설치하여 비동기적 서버를 구축한 다음 redis로 채널 레이어를 설정해 django내에 consumer와의 통신이 원활할 수 있도록 구현하였습니다.

[슬라이딩 10: 프로젝트 진행 중 발생한 의문점]

프로젝트 진행 중 모델링 부분에 대해 몇 가지 의문점이 발생했습니다.

첫째로, LLM을 추가학습하더라도 추가학습하는 데이터셋의 크기가 충분히 크지 않으면 LLM의 layer 가중치를 크게 업데이트하기 힘들 것입니다. 하지만 추가학습했을 때 보이는 결과가 생각보다 차이가 커 이 부분에 대해 의문이 들었습니다.

두번째는 KoBERT 학습 시 learning rate 값을 조금만 바꾸더라도 영향을 많이 받는데 그 이유가 무엇일지에 대해 의문이 들었습니다.

마지막으로 KoLLaMA 모델이 추가학습하는 데이터셋에 영향을 덜 받게 하려고 learning rate 값을 10^-8 정도까지 많이 줄여보았습니다. 이런 경우 사전학습된 Base 모델에 가까운 모델이기 때문에 비정상적인 응답을 생성하지 않을 것이라고 생각했는데 그렇지 않았습니다.

이러한 의문점들에 대해 추가로 공부하여 해결하고자 합니다.

[슬라이딩 11: 배포 웹사이트 시연]

저희가 제작한 배포 웹사이트에 대한 시연 영상입니다.

먼저 보이스피싱 판별에 대한 영상입니다.

보이스피싱 음성 파일을 입력하게 되면 모델 분류를 통해 보이스피싱일 확률을 제공해줍니다.

일반 음성 파일을 입력했을 때는 보이스피싱이 아닌 것으로 분류해서 보이스피싱이 아닐 확률을 제공해줍니다.

다음은 롤플레잉에 대한 영상입니다.

여러 보이스피싱 시나리오 중 하나를 고르고 모델이 제공하는 응답에 맞추어 롤플레잉을 진행할 수 있습니다.

대화 종료를 입력하여 롤플레잉을 종료하게 되면 사용자의 응답에 맞춘 피드백을 제공해줍니다.

마지막으로 대처방안에 대한 영상입니다.

보이스피싱을 당한 사용자가 현재 상황을 선택하게 되면 사용자가 처한 상황에 대한 대처방안을 제공합니다.

[슬라이딩 12: Q&A]

[슬라이딩 13: 발표 종료]

이상 발표 마치겠습니다. 감사합니다.