1. 발표 시작
2. 목차는 다음과 같습니다.
3. 먼저 저희 주제에 대해 간략히 소개하겠습니다. 저희의 큰 목표는 인터넷 내의 게시글을 분석해 범죄 예방 및 관리를 도와주는 것입니다. 최근 청소년들의 마약, 성범죄가 늘고 있는데 이런 청소년들의 범죄를 사전에 예방하고, 치료 및 재활 가능성이 있는 사람들에게는 도움을 주기 위해 프로젝트를 진행했습니다.
4. 마약거래나 성매매 대부분 주변에서 찾기 힘들기 때문에 접근성이 높은 인터넷을 통해 이루어집니다. 화면에 보이시는 것처럼 SNS 또는 커뮤니티에서 마약 거래를 위한 게시글을 올리고 실제로 거래까지 이루어지는 상황이 많이 발생합니다. 또한 이런 심각한 범죄뿐만 아니라 위협적이고 폭력적인 대화가 오가기도 하는데, 저희는 이런 점에 착안하여 저희의 서비스를 이용해 범죄를 사전에 예방하는 방식을 생각했습니다.
5. 간단한 예시를 통해 저희 서비스의 구현 방식을 말씀드리겠습니다. 우선 SNS, 딥웹 같은 게시물 또는 댓글을 저희 위험도 판별 모델에 입력합니다.
6. 모델은 입력한 텍스트에 대해 위험 카테고리를 분류하고 각 카테고리의 위험도를 측정합니다.
7. 그럼 이렇게 분석한 결과를 바탕으로 텍스트의 의도를 분석합니다. 예시를 한 가지 가져왔는데, 보시는 것처럼 게시글의 내용을 입력하면 판별 모델로부터 분석된 결과를 출력하고
8. 사생활 침해 및 불법 촬영에 해당하는 내용임을 알려주고 관련된 내용을 결과로 보여줍니다. 그럼 이렇게 분석된 내용에서, 두가지 솔루션을 제공합니다.
9. 먼저, 가해 의도가 높은 게시글에 대해서는 관련된 범죄에 해당하는 법률 조항을 제시함으로써 경각심을 조성하고 범죄를 저지르려는 의도를 줄일 수 있습니다.
10. 두번째로는 ‘성폭행을 당했다’ 또는 ‘어떤 사건에 연루되었다’ 같이 피해를 입은 듯한 사람에 대해서 챗봇 상담을 통해 도움을 줄 수 있습니다.
11. 이제 이런 서비스를 어떻게 구현했는지 구체적으로 소개드리겠습니다. 먼저 데이터셋입니다. 공공 데이터지도와 AIHUB를 이용해서 데이터를 수집했습니다. 성범죄
12. 마약,
13. 해킹을 포함한 사이버 범죄에 대한 데이터셋을 사용했고
14. 위험한 대화가 아닌 일상 대화 데이터도 사용했습니다.
15. 분류 모델을 만들기 위해서 수집한 데이터에 각 카테고리에 대해 라벨링을 진행했고
16. .
17. 텍스트 데이터의 불필요한 노이즈 제거를 위해
18. 전처리를 진행했습니다.
19. 하지만 데이터의 부족으로 인해서, 위험 대화에 대한 데이터의 비율이 적은 편이었습니다. 이를 해결하기 위해 관련 연구를 찾아본 결과 EDA 라는 텍스트 데이터의 증강 기법이 있었습니다. 이를 한글 데이터에 적용해 증강하는 방식이 있어 해당 방식을 활용했고,
20. 그 결과 데이터가 부족한 카테고리에 대해서 대화를 증강했습니다. 그리고 비율을 동일하게 맞춰주기 위해서 1500개씩 랜덤 샘플링을 진행해 데이터셋을 구축했습니다. 여기서 비율은 저희가 임의로 정한 것이기 때문에 추후에 검증 과정을 통해 모델의 성능이 가장 잘 나오는 비율을 선택해야 할 것입니다.
21. 저희 데이터가 한글 데이터임을 고려해 모델은 한국어 sbert 모델을 선택해, 구축한 데이터셋을 활용해 분류 모델을 만들기로 했습니다.
22. 배치사이즈는 16, 에포크는 3으로 파인튜닝을 진행했고, 보시는 것처럼 학습을 진행하며 loss가 떨어져 훈련 데이터에 대해서는 0.2 정도의 loss를 기록했습니다.
23. 구축한 모델에 대해서 평가는 두가지로 진행했습니다. 우선은 검증 데이터셋에 대해서 여러가지 평가지표를 활용해 전체적으로 90%의 정확도와 0.35의 loss로 성능이 나쁘지 않은 것을 확인할 수 있었고 데이터셋에는 없는 몇 가지 위험한 대화를 직접 만들어 입력해 결과를 평가해보기도 했습니다.  
     위험도에 대한 측정은 어떻게 할 지 저희도 고민이 많았는데, 각 군집들의 중심과 input 데이터의 유사도를 활용해 위험도 스코어를 구하는 방법, 멀티 클래스 분류에 대한 고려도 있었는데 여러가지 문제가 있었습니다. 간단하게 각 카테고리에 대해 분류 모델이 예측하는 확률을 활용하기로 했습니다. normal이 아닌 다른 카테고리에 대한 확률의 합을 위험도로 지정했는데 그 결과 생각보다 나쁘지 않은 결과가 나와 이를 채택했습니다.
24. 다음으로 챗봇 서비스를 위해 저희는 오픈api, 랭체인 프레임 워크를 이용했습니다. 이전 모델의 분석 결과를 input으로 전달하여 게시물의 의도를 분석하고 솔루션을 제공하는 프롬프트를 작성했습니다.
25. 구현하기 위해 사용한 기술인 검색 증강 생성 기술인 rag입니다. 모르시는 분들을 위해 간략히 설명드리자면, 언어모델의 출력을 최적화하기 위해 외부 데이터를 수집해 이를 참조해서 높은 퀄리티의 output을 뽑을 수 있도록 하는 기술입니다. 저희 챗봇의 특성 상 법률에 대한 도메인 이해가 필요한데, 이 RAG 기술을 이용해 최대한 보완하도록 했습니다.
26. 자세한 과정을 설명드리겠습니다. 먼저 논문이나 공공 데이터 등을 통해 카테고리별로 기본적인 정보와 관력 법적 정보를 수집해 로드하고, 기본적인 전처리 이후 오픈Api 임베딩과 파이스를 이용해 벡터 스토어에 저장합니다. 이제 구축된 데이터를 기반으로 사용자의 질문과 유사도가 가장 높은 청크를 출력하고 이를 프롬프트에 전달해 응답을 도출합니다.
27. 이를 프롬프트에 전달해 응답을 도출하는 과정입니다.
28. 저희는 이렇게 크게 두가지 모델을 사용해서 서비스를 구축했고 다음으로 프로토타입을 간략하게 보여드리겠습니다.
29. 위험도 판별 모델의 분석 결과를 연결해 input 데이터를 분석해 위험도와 대화의도를 분석하고 관련 법 조항을 제시합니다.
30. 챗봇의 화면 구성은 다음과 같이 사이드바와 메인 화면으로 구성되어 있습니다. 사이드바에서는 사용자가 유동적으로 다른 카테고리에 대한 지원이 필요할 경우 이용할 수 있도록 카테고리를 선택할 수 있도록 되어있고, 메인화면에서 챗봇과 대화할 수 있습니다.
31. 대화하는 과정은 영상으로 먼저 보여드리겠습니다.
32. 이외에도 몇 가지 예시를 보여드리겠습니다. (설명 하면서)
33. .
34. 그럼 이 서비스가 주는 기여점, 적용 방식에 대해서 설명드리겠습니다. 우선은 웹 사이트 에 위험한 대화로 판단하면 경고를 주고 일정 위험도를 넘긴 유저에게는 저희 챗봇 또는 법률 조항 제시를 통해 사전에 범죄를 예방할 수 있습니다. 또한 피해자로 보이는 유저에게는 챗봇을 통해 도움을 주는 방식으로 사용할 수 있습니다. 이외에도 온라인 환경 안전 교육이나 유해 콘텐츠 필터링에 이용할 수 있습니다.
35. 마지막으로 저희 서비스를 구현하면서 느낀 몇 가지 한계점에 대해 설명하겠습니다. 이런 서비스를 제공하기 위해서는 결국 웹사이트나 SNS에 사용되어야 하는데, 사용자 경험에 큰 영향을 끼칠 수 있어 실현 가능성 및 적용 방식에 고민이 필요합니다. 또한 데이터셋의 부족으로 위험 카테고리가 제한되어 있고, RAG에 필요한 데이터도 마찬가지로 부족했습니다. 이는 기회가 된다면 추후에 데이터를 확보해 개선을 진행할 예정입니다. langchain에서 지원하는 프롬프트 전략인 react 전략를 사용하여 질문자의 의도에 맞게 웹검색, 특정문서반환, 등 다양한 tool들로 폭넓은 정보를 제공하고 개인 정보가 담긴 정보들은 rac을 그대로 이용해 개인 맞춤형 챗봇으로 거듭날수있도록 합니다.
36. 참고로 몇 가지 예시를 더 가져왔는데 결과가 잘 나오는 것을 확인할 수 있었습니다.
37. 이상으로 발표 마치도록 하겠습니다. 감사합니다.