발표 6분 q&a 2분

안녕하세요 5조 발표를 맡게 된 이은강 입니다. 저희 조의 주제는 텍스트 분석을 통한 사이트 범죄 위험도 분석입니다. 목차로는 이 주제를 정하게 된 배경소개, 주제 소개, 사용할 데이터 셋, 어떻게 분석을 할건지에 대한 파이프라인, 마지막으로 prototype에 대해 발표하겠습니다.

첫번째로 주제 선정 동기에 대해 말씀드리겠습니다. 뉴스기사를 보시면 다크웹이나 SNS를 통한 온라인 마약 거래가 2-30대 사이에서 증가하고 있는 것을 볼 수 있습니다. 또한, 미성년자들을 대상으로 디지털 상에서 성범죄를 일으키는 디지털 그루밍 현상도 많이 일어나고 있습니다. 예를 들어 SNS를 통해 용돈이나 아이템을 주겠다는 식으로 미성넌자들을 유혹하는 것이 디지털 그루밍의 대표적인 예시입니다.

아래 통계를 보시면 2023년동안 총 2만 6천건이 온라인에서 마약거래가 일어났습니다. 그 중 트위터에서 약 1만 4천 건 정도가 일어났습니다. 이는 다크웹 뿐만 아니라 많은 SNS상에서도 마약 거래가 일어나고 있는 것을 볼 수 있습니다. 아동 청소년을 대상으로 한 온라인 그루밍 성착취 피해도 점점 증가하고 있습니다. 2020년 이후에는 14세 미만 아동 피해자의 급증, 14-19세 미만 피해자의 수는 400건대를 돌파하였습니다. 따라서 저희는 온라인 환경에서의 마약 및 성매매 범죄율이 증가한다는 것을 문제 삼게 되었습니다.

두번째로 저희가 앞으로 분석하고 얻고자 하는 주제에 대해 말씀드리겠습니다. 앞에 통계에서도 보셨듯이, 미숙한 청소년들을 대상으로 한 마약거래와 성범죄가 많이 일어나는 것을 볼 수 있었습니다. 저희는 이를 사전 방지하여 피해자를 줄이기 위해 이 주제를 선정하게 되었습니다. 분석을 통해 얻고자 하는 것은 찾은 데이터 셋을 바탕으로 온라인 마약거래, 성범죄 관련 게시물, 채팅을 분석하여 유의한 범죄 위험도를 나타내고자 합니다. 이에 따른 활용방안으로는 온라인 범죄와 관련된 게시물이나 채팅에 노출 시 자동으로 사용자에게 경고를 주는 서비스로 활용할 수 있을 것 같습니다. 기존에 적용되는 모델은 피해영상물을 찾아낸 뒤 SNS상에 유포되는 피해 영상을 24시간 동안 실시간으로 자동 추적, 감시한다…. 하지만 저희가 만들 모델은 텍스트 마이닝 기법을 통해 마약, 성범죄 관련 게시물과 채팅을 사전에 차단하고 경고할 수 있는 차별성이 있습니다.

세번째로 사용할 데이터 셋에 대해 말씀드리겠습니다. 스마트 치안 빅데이터 플랫폼에서 얻은 마약 통계 데이터와 성범죄 관련 단어 데이터입니다. 먼저 마약 통계 데이터는 다크 웹 사이트 내 마약 관련 출현 단어를 월별로 보여줍니다. 이 데이터의 컬럼은 (1-8 말하기) 입니다. 두번쨰로 사용할 데이터는 성범죄 관련 단어 데이터입니다. 이 데이터는 다크 웹 사이트 내 성범죄 단어가 포함된 게시물을 보여줍니다. 이 데이터의 컬럼은 (1-22 말하기) 입니다.

네번째로 어떻게 분석할건지에 대한 파이프라인을 소개해드리겠습니다. 데이터 수집 및 전처리. 모델 선택, fine-tuning, 모델 평가의 순서로 진행됩니다. 먼저, 데이터 수집 및 전처리로는 스마트치안빅데이터 플랫폼에서 마약 통계 데이터와 성범죄 관련 단어 사전 데이터르 수집할 계획이고 각 데이터의 카테고리 별로 라벨링을 진행할 예정입니다. 만약 데이터의 양이 충분하지 않다고 하면 zero-shot, few-shot learning등의 기법으로 레이블 없이 비지도 학습을 진행할 것도 고려하고 있습니다. 두번째로 모델 선택으로는 범죄 용어 데이터를 확인해보면 각 단어의 파생어가 여러개가 존재하고, 일상적 단어가 은어로 사용되는 경우도 여럿 존재합니다. 예를 들면 마약은 떨이나 드럭으로 표기되고, 은어를 사용하는 경우에는 아이스나 캔디로 표기되는 경우가 있습니다. 이런 데이터를 잘 처리하기 위해 저희는 문맥 고려와 단어 관계 분석에 더 뛰어난 bert기반의 언어모델을 채택하였습니다. 그 중에서도 한국어 데이터 셋으로 사전 훈련된 클루버트, 코버트 등의 모델을 사용할 계획입니다. 추후 필요한 경우에 수집된 데이터를 가지고 추가 학습을 진행할 계획입니다. 세번째로는 이후에 저희 프로젝트가 목표로 하는 해당 게시글, 대화의 주제 분류/위험 단어 출력/ 전체적인 위험도 분석에 맞게 파인튜닝을 진행한 후, confusion matrix를 통한 평가와 ChatGPT를 이용한 직접 비교 이 두 가지 방식을 통해 모델 평가를 진행할 예정입니다.

마지막으로 프로토타입에 대해서 설명 드리겠습니다. 첫번째 프로토타입은 만든 모델을 활용하여 게시물에 대한 범죄 예측 레포트를 작성합니다. 사용자가 접근한 게시물의 텍스트를 추출하여 모델의 인풋으로 전달하고 성범죄 게시물일 확률과 마약거래 게시물일 확률을 결과로 보여줍니다. 또한. 확률을 기반으로 위험도를 측정합니다. 예를 들어 성범죄 게시물일 확률이 75%이상일 경우 위험도가 위험으로 분류됩니다. 이에 대한 위험도는 차후에 구체적으로 정할 계획입니다. 추가로 토픽분류를 통해 각각의 범죄와 관련된 토픽 목록을 보여줍니다. 두번째 프로토타입은 은어표현 추측 서비스입니다. 온라인 상에서의 마약거래에 이루어지는 은어들은 계속해서 생겨나도 변화합니다. 익숙하지 않은 은어를 사용자가 마스킹을 하면 비슷한 단어들을 목록으로 만들어 보여줍니다. 은어 표현의 Top 10 단어들을 통해 마스킹한 단어를 유추할 수 있습니다.

이렇게 5조 발표를 마무리하겠습니다. 감사합니다.