연구 회의 음성 데이터의 실시간 정보 중요도 등급화를 위한 BERT 기반 접근법

이지민[1], 최예지[2], 장항배[3*]

중앙대학교 일반대학원 융합보안학과[1,2]

중앙대학교 경영경제대학 산업보안학과[3*]

{dlwlals[1], yjji9783[2]}@cau.ac.kr, hbchang@cau.ac.kr[3*]

A BERT-based Approach for Real-time Information Importance Classification of Research Meeting Audio Data

Lee Ji Min[1], Choi Ye Ji[2], Chang Hang Bae[3*]

Department of Security Convergence, Graduate School, Chung-Ang University[1,2]

Department of Industrial Security, College of Business and Economics, Chung-Ang University[3]

요 약

본 논문은 4차 산업혁명과 데이터 기반 경제 시대에서 연구 보안을 강화하기 위해 음성 데이터를 활용하여 BERT 기반 정보 중요도 등급화 접근법을 제안한다. 연구 회의에서는 중요한 아이디어, 프로젝트 보완, 전략적 결정 및 민감한 정보가 논의되며, 이러한 정보를 신속히 식별하고 차등적으로 등급화하여 보호하는 것은 효율적인 자원 관리와 보안 조치에 필수적이다. 기존의 텍스트 기반 정보 등급화 연구와 달리, 본 연구는 AI-Hub의 주요 영역별 회의 음성인식 데이터를 활용한다. 음성 데이터를 SpeechRecognition을 사용하여 텍스트로 변환한 후, KOBERT 모델을 통해 분석하여 정보의 중요도를 분류함으로써, 문서 데이터뿐만 아니라 대화 속에서도 중요한 정보를 식별하고 보호할 수 있다. 연구 결과, 제안된 접근법으로 음성 데이터가 기밀, 비밀, 대외비, 공개의 4가지 등급으로 분류되었으며 이는 정보의 기밀성을 유지하고 자원 관리의 효율성을 높일 수 있음을 확인하였다. 향후 연구에서는 잡음, 발언이 겹치는 상황, 방언과 억양 차이를 효과적으로 처리하여 음성 인식의 정확도를 향상하고자 한다. 또한, 개선 사항들을 토대로 제안된 시스템의 성능을 평가할 것이다. 본 연구는 연구 회의에서 생성되는 중요한 정보를 보호하고, 데이터 유출 방지와 보안 사고 예방에 기여할 수 있는 실질적인 솔루션을 제공할 것으로 기대된다.

Ⅰ. 서 론

4차 산업혁명과 데이터 기반 경제 시대에서 정보의 관리와 보안은 어느 때보다 중요하다. 연구 기관과 기업은 방대한 양의 데이터를 생성하며, 이 중에는 연구 회의에서 교환되는 중요한 정보도 포함된다. 이러한 회의에서는 새로운 아이디어, 프로젝트 보완, 전략적 결정, 민감한 정보 등이 논의되며, 이는 조직의 성공에 필수적이다. 따라서 이러한 정보를 실시간으로 등급화하는 것은 효율적인 자원 관리와 강력한 보안 조치를 위해 필요하다.

특히, 연구 회의에서 다루는 정보는 기밀성이 높고 민감한 내용을 포함하는 경우가 많아 이를 적절히 보호하는 것은 중요하다. 중요도에 따라 데이터를 분류하고, 중요하지 않은 데이터에 과도한 보안 자원을 낭비하지 않도록 해야 한다. 전통적인 정보 등급화 방법은 종종 수작업으로 이루어져 비효율적이며, 데이터의 생성 속도와 동적 성격을 따라가지 못하는 경향이 있다. 이러한 한계를 극복하기 위해, 본 논문에서는 자동 음성 인식 시스템과 BERT 모델을 적용하여 실시간으로 정보 등급화를 하는 방법론을 제안한다.

Ⅱ. 관련 연구

디지털 시대의 데이터 증가와 민감한 정보의 보호가 중요해지면서, 정보의 차등적 중요도 평가는 데이터 유출 방지와 보안 사고 예방에 핵심적인 역할을 하고 있다. 이에 따라 정보 등급화에 관한 연구는 지속적으로 진행되고 있다. 그 예로 홍기완 등(2023)은 그래프 임베딩 기술을 활용하여 디스플레이 기술 문서를 효과적으로 분류하는 방법을 제안했다.[1] 이는 기술 문서의 등급화를 통해 정보 관리와 보안의 효율성을 높이는 방안을 탐구했다. 안성준과 장항배(2021)는 연구 정보의 중요도를 분류하여 보안 수준을 최적화하는 방안을 제안했다.[2] 연구 정보의 기밀성을 유지하고 자원 관리를 효율적으로 하기 위한 정보 등급화의 필요성을 강조했다. 윤현수 등(2018)은 클러스터링 기법을 활용하여 정보 자산을 효과적으로 등급화하는 방법을 제안했다.[3] 이는 정보 자산의 효율적인 관리와 보안 자원 배분의 최적화를 목표로 하였다.

본 논문은 텍스트로 이루어진 기술 문서를 다루었던 기존 연구와 달리 데이터의 유형의 범위를 확장한 음성 데이터를 활용하여 정보의 중요도를 등급화한다는 점에서 차별성이 있다. 이는 단순한 문서 데이터 분석을 넘어 대화 속에서도 중요한 정보를 식별하는 것을 가능하게 한다.

Ⅲ. 제안하는 방법론

Ⅲ-Ⅰ. 데이터 수집 및 ASR 기반 음성 인식

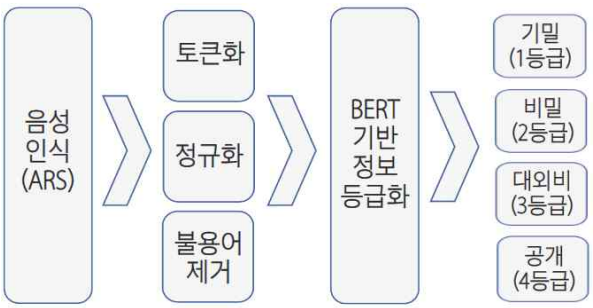
본 연구는 음성 데이터를 기반으로 정보 중요도 등급화를 위해 공공데이터포털 AI-Hub 주요 영역별 회의 음성인식 데이터셋의 일부를 사용하였다. 해당 데이터셋은 2021년 수집된 데이터로, 정치, 경제, 사회, 교육

등 10개의 카테고리에 대하여 3인 이상의 회의 녹음을 포함하고 있다. 이는 wav 파일 형식으로 수집되었으며, SpeechRecognition 음성 인식을 적용하여 텍스트 형식으로 변환하였다.

6: 최 변호사님
11: 어~ 뭐~ 어쨌든 지금 우리나라 전체적인 재판 비율 그러니까 구속되거나 재판에 회부되거나 실형 나오는 비율이 마약 사범이 거의 뭐~
(열 배에서)/(10배에서) (열)/(10) 몇 배까지 많습니다 그럴 정도로 지금 우리나라가 (()) 인데
지나친 (()) 치료 모델 (()) 도입을 해야 되겠고 두 번째가 이제 그 공급 업체하고 수요 업체 모델이 적절하게 조화를 이뤄야 되겠다 이 생각을 합니다
그리고 세 번째로는 역시 청소년이 가장 주된 포인트가 되어야 합니다 우리들의 관심을 기울여야 되는 그리고 청소년들은 단순히 마약뿐만이 아니라
온갖 중독성 있는 약물들 어~ 술 담배
그다음에 기타 환각 물질까지 포함한 그런 종합적인 대책이 좀 더 절실하게 요구된다 이렇게 말씀드릴 수 있을 것 같습니다.
6: 부장님
10: 예 그~ 그동안에 인제 그~ 어떤 의미에서는 검찰이나 뭐~ 경찰 등 그~
단속 기관 위주의 그~ 마약 통제 정책 이게 그동안의 주간을 이루어 왔는데
어~ 다행스럽게 다음주에 작년 연말에 지금 구성이 돼 있습니다.
(십사 개)/(14개) 정부 부처하고 그다음에 (오 개)/(5개) 우리 마약 퇴치 관련 민간자들이 참여 해 가지고
어~ 처음으로 지금 어~ 국무총리실 주관으로 어~ 국가마약류대책협의회가 다음주에 개최 예정으로 있습니다.
오늘 여기 그~ 같이 패널로 참석하신 분들 또 우리 @이름8 전도사님 또 시청자분 전화주신 내용을 제가 충분히
그~ 정책 수립 과정에 반영하도록 노력을 하겠습니다.
오늘 이렇게 좋은 프로그램을 마련해 주신 것에 대해서 굉장히 저희 검찰 당국에서는 감사드립니다.
6: 저희도 이렇게 나와주신 부장님과 또 토론자 여러분들께 진심으로 감사의 말씀을 드리겠습니다.
투약자와 마약 공급책에 대한 단속과 처벌을 지금까지와는 비교가 안 될 정도로 강화하는 것도 한 방법일 수 있을 것 같습니다.
또 중독자에 대한 체계적인 치료도 반드시 필요합니다.
그렇지만 건강한 사회 의식이 회복되지 않는 한 이 모든 노력은 미봉책에 불과하다는 생각을 해 봅니다.
어제 오늘 모처럼 단비가 촉촉하게 내렸습니다.
진지들과 따뜻하고 정겨운 주말 되시길 바랍니다 (사 월)/(4월) 첫째주 난상 토론 오늘 여기서 문을 닫겠습니다

(그림 1) ASR 기반 음성 인식으로 변환한 데이터

텍스트 데이터의 경우, 일반적으로 자연어 처리 과정에서 전처리 과정이 필수적으로 요구된다. 그림 2에서와 같이 전처리 과정에서는 불용어 제거, 정규화를 실시하고 토큰화를 통해 BERT 모델의 입력 데이터를 생성하였다.



(그림 2) 음성 데이터 기반 정보 중요도 등급화 프레임워크

Ⅲ-Ⅱ. BERT 기반 분류 모델 적용

본 연구에서는 BERT 기반 모델 중 한국어에 특화된 KOBERT 모델을 사용하여 텍스트로 변환된 연구 회의 음성 데이터를 분석한다. 정보 등급은 기업 보유 정보를 대상으로 정보 창출 및 유지 비용, 산출 정보 수준, 정보 활용도, 내부 활용 효과, 외부 유출 위험의 5가지 요소를 평가하여 기밀(1등급), 비밀(2등급), 대외비(3등급), 공개(4등급)의 네 가지로 분류한다.[4, 5] 모델의 학습은 전문가가 각 정보 항목에 대해 평가한 점수를 바탕으로 기밀, 비밀, 대외비, 공개의 네 가지 등급으로 라벨링 한 데이터를 사용하였으며, 학습된 KOBERT 모델을 사용하여 정보 중요도 등급화를 적용한 결과 예시는 표 1과 같다. 본 연구에서는 연구 회의 음성 데이터로부터 중요한 정보를 신속히 식별하여 이를 기밀, 비밀, 대외비, 공개로 등급화할 수 있음을 확인했다.

(표 1) KOBERT를 적용한 정보 중요도 등급화 결과 예시

등급	연구 회의 핵심 내용 요약
기밀(1)	정치적 이슈와 정책 결정 과정에 관한 논의
공개(4)	경제 전문가들의 경제 동향 의견 회의
대외비(3)	사회적 이슈와 복지 정책에 관한 논의

Ⅳ. 결론

본 논문에서는 연구 보안을 강화하기 위해 주요 영역별 회의 음성인식 데이터를 활용하여 음성 데이터를 텍스트로 변환한 후, KOBERT 모델을 사용하여 실시간으로 정보의 중요도를 분류하는 접근법을 제안했다. 연구 결과, 음성으로부터 중요한 정보를 신속히 식별하고, 이를 기밀, 비밀, 대외비, 공개의 네 가지 등급으로 등급화할 수 있음을 확인했다.

제안한 방법론은 연구 회의에서 논의되는 기밀성이 높은 정보를 식별하고 적절한 보안 조치를 적용함으로써 중요한 연구 정보의 유출을 방지할 수 있다. 또한, 중요도에 따라 데이터를 분류하여 보안 자원과 관리 자원을 효율적으로 배분함으로써 자원의 활용도를 극대화할 수 있다. 실시간으로 정보를 등급화하고 중요도를 평가하여 변화하는 상황 속에서도 신속하게 대응할 수 있는 중요한 기여를 할 것으로 기대된다.

향후 연구에서는 회의 환경에서 발생할 수 있는 잡음, 여러 사람의 발언이 겹치는 상황, 방언과 억양 차이 등을 효과적으로 처리하여 음성 인식의 정확도를 향상하고자 한다. 또한, 이러한 개선 사항들을 통해 제안된 시스템의 성능을 평가해보고자 한다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 정보통신방송혁신인재양성(융합보안핵심인재양성)사업의 연구 결과로 수행되었음 (IITP-2024-RS-2023-00266605)

참 고 문 헌

[1] 홍기완, 한유나, 윤원석. "그래프 임베딩 기반 디스플레이 기술문서 등급화 기술 연구." 한국산업보안연구, 13, pp. 1-25, 2023.

[2] 안성준, 장항배. "연구보안을 위한 연구정보 등급화 방안 연구." 한국정보처리학회 학술대회논문집, 28, 2, pp. 273-275, 2021.

[3] 윤현수, 김용현, 김동화, 신동규, 신동일. "클러스터링을 이용한 정보자산의 등급화 방안." 한국통신학회 학술대회논문집, pp. 368-369, 2018.

[4] Na, Onechul, et al. "The rating model of corporate information for economic security activities." Security Journal, 32, pp. 435-456, 2019.

[5] 특허청, “우리 기업의 영업비밀 등급분류 가이드(2021년 개정판)”, 2021.