DistilBert 모델 기반 GPT 보안 솔루션 개발

정주원
광운대학교 방산AI로봇융합학과

ABSTRACT (요약)

인공지능(AI) 기술의 발전은 자연어 처리(NLP) 및 기계 학습(Machine Learning) 분야에서 큰 혁신을 이루며, GitHub Copilot과 같은 코딩 자동화 도구의 출현을 가능하게 했다. 그러나 이러한 도구의 사용은 사용자 중요 정보 유출이라는 심각한 문제를 야기하고 있다. **본 연구는 DistilBERT 모델을 사용하여 GitHub Copilot을 이용하는 사용자의 주요 소스코드 정보를 보호하는 보안 솔루션을 개발하고 그 성능을 평가하는 것을 목적으로 한다.**

본 연구에서는 VSCode 환경에서 Copilot을 연동하여 주요 소스코드 정보를 마스킹(Masking)하는 확장 프로그램을 설계 및 개발하는 방법을 제안한다. 이 확장 프로그램은 Webpack으로 모듈화되며, DistilBERT 모델을 통합하여 사용자의 입력 정보가 실시간으로 암호화된다. 암호화된 정보는 Copilot에 안전하게 전송되며, GPT 기반의 Copilot은 이를 바탕으로 응답을 생성한다.

본 연구는 AI 기술과 정보 보호의 융합을 통해 새로운 가능성을 제시하며, 기업 및 개발자들에게 보안 중심의 솔루션을 제공함으로써 실무적 적용 가능성을 높인다.

INTRODUCTION

Copilot은 OpenAI의 GPT 모델을 이용한 인공지능 기반의 지능형 챗봇으로, AI 시대의 도래와 함께 현업 개발자들이 VSCode에 Copilot을 연동하여 프로그램 개발 속도를 단축하는 등 효율적인 프로그램 개발을 추구하고 있다. 이로 인해 **GPT를 사용한 개발이 불가피**한 상황이다. 그러나 GPT를 사용한 한국 이용자 678명의 개인정보가 유출되거나, 인공지능 학습을 목적으로 GPT에 입력된 정보가 GPT 서버로 넘어가면서 **개인정보 및 기업 기밀이 유출되는 사건이 빈번하게 발생**하고 있다. **Copilot 또한 OpenAI가 개발한 ChatGPT와 동일한 GPT 모델을 사용하고 있어 중요 정보 유출 사고에 각별한 주의가 필요**하다. 따라서 Copilot 및 GPT 사용자의 입력 정보를 보호하는 방법을 개발하는 것이 시급하다.

<보안 이슈로 인한 주요 기업의 GPT 사용 실태>

Literature Review

신영진(2021)에서 AI 데이터 처리 과정에 대해 문헌 조사 수준으로 일부 다루고 있을 뿐, 대부분의 선행 연구는 기존 법제도 조사 및 법령 해석에만 초점을 맞추고 있다. **현재까지 인공지능 이용에 있어서 개인 및 기업의 중요 정보를 사전에 보호 처리하는 기술적 방법에 대한 실질적인 논의는 이루어지지 않고 있다.**

관련 연구	연구 목적	분석 대상	분석 방법
신영진 (2021)	우리나라가 추진해야 할 개인정보보호 개선방안을 도출	우리나라의 인공지능 서비스	법제적 기준과 데이터 처리과정 기준을 중심으로 문헌 조사
이원태, 강장묵 (2016)	인공지능 기술/서비스 기반의 개인정보 보호 모델에 대한 연구	인공지능 기술/서비스	개인정보보호 침해 이슈와 개인정보보호를 보장하면서 인공지능 산업을 발전시킬 수 있는 도덕적/법제도적 모델 조사
김용대, 장원철 (2016)	인공지능의 발전과 개인정보 보호라는 두 가지 상충되는 가치를 효율적으로 증진시킬 수 있는 방법에 대한 연구	인공지능 산업	인공지능 및 빅데이터의 발전 현황 및 개인정보보호 관련 법률체계 분석

RESULTS & DISCUSSION

본 연구에서 개발한 보안 솔루션의 주요 기능은 다음과 같다.

1. 중요 정보 암호화 AI 모델(DistilBERT)의 구축 및 통합

: DistilBERT AI 모델을 통합하여 Copilot으로 전송되는 사용자 입력 정보 중 중요 정보를 사전에 효과적으로 식별하고 보호할 수 있다.

2. VSCode 환경에서의 Copilot 연동 및 사용자 입력 정보 보호 확장 프로그램 개발

: 사용자는 Copilot을 사용할 때 해당 확장 프로그램을 통해 중요 정보 노출을 최소화하면서, 속도 지연 없이 안전하게 작업할 수 있다.

3. Webpack으로 모듈화된 확장 프로그램

: 확장 프로그램은 Webpack으로 모듈화되어 입력 정보를 보호한다. 이를 통해 사용자 입력 정보 보안 수준을 높이고, 확장성 있는 구조를 유지한다.

이러한 보안 중심의 AI 기술 개발은 **기존 AI 기술 개발에서 해킹 및 중요 정보 보호 방안을 고려하지 않았던 점을 보완**하며, **AI 기술과 중요 정보 보호 간의 새로운 가능성을 제시**한다. 이를 통해 **기업 및 개발자들에게 보안 중심의 솔루션을 제공**할 수 있다.

또한, **해당 솔루션은 다른 AI 기반 도구와의 상호작용에서도** 적용될 수 있으며, **향후 보안 및 개인정보 보호 기술의 발전에도 기여**할 수 있다. 중요 정보 보호 기능을 갖춘 AI 모델(DistilBERT)을 사용함으로써 **사용자가 AUC 95%이상이라는 높은 정확도로 안전하게 서비스를 이용하면서도, 응답 시간 및 속도 측면에서도 지연시간 100ms이하로 성능을 유지하여 중요한 경쟁 우위를 확보**할 수 있다.

향후 연구에서는 다음과 같은 방향으로 확장될 수 있다:

1. 다양한 프로그래밍 환경 확장

: VSCode 외에도 다양한 프로그래밍 환경과 통합할 수 있는 확장 프로그램을 개발하여 다른 개발자 도구에서도 동일한 중요 정보 보호 기능을 제공할 수 있도록 연구를 확장할 수 있다.

2. 대규모 사용자 테스트

: 다양한 사용자 그룹을 대상으로 대규모 테스트를 진행하여, 실사용 환경에 맞춰 솔루션을 개선할 수 있다.

본 연구를 통해 AI 기반 도구의 중요 정보 보호 수준을 높이고, 사용자에게 더욱 안전하고 효율적인 개발 환경을 제공할 수 있을 것이다.

표 1. 모델 성능 결과

성능 분류	목표 성능 계산
정확도 성능	$Target\,AUC = \int (TPR(FPR))dFPR > 95\%$
속도 성능	$Target\,d_{latency} = d_{proc} + d_{queue} + d_{trans} + d_{prop} < 100ms$

REFERENCES

- 신영진, (2021), 우리나라의 인공지능(AI)서비스를 위한 개인정보보호 개선방안, 중소기업융합학회, 20-33
- 이원태, 강장묵, (2016), 인공지능 기술/서비스 기반의 개인정보 보호 모델에 대한 연구, 한국인터넷방송통신학회, 1-6
- 김용대, 장원철, (2016), 인공지능산업 육성을 위한 개인정보보호 규제 발전 방향, 161-176
- 중앙일보, (2023), '챗GPT, 한국인 687명 개인정보 유출...국내법 적용해 첫 제재'
- BBC, (2023), 'ChatGPT 오류 ... 다른 사용자 대화 기록 유출'
- 조선경제, (2023), '챗GPT에 묻다가 기밀 샌다" 기업마다 정보보안 골머리[NOW]'