Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.34, NO.2, Apr. 2024 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2024.34.2.207

GPT를 활용한 개인정보 처리방침 안전성 검증 기법*

심 혜 연,[1†] 권 민 서,[2] 윤 다 영,[2] 서 지 영,[2] 이 일 구[3‡]
[1,2,3]성신여자대학교 (대학원생, 학생, 교수)

Safety Verification Techniques of Privacy Policy Using GPT*

Hye-Yeon Shim,[1†] MinSeo Kweun,[2] DaYoung Yoon,[2] JiYoung Seo,[2] Il-Gu Lee[3‡]
[1,2,3]Sungshin Women's University (Graduate student, Student, Professor)

요 약

4차 산업혁명으로 인해 빅데이터가 구축됨에 따라 개인 맞춤형 서비스가 급증했다. 이로 인해 온라인 서비스에서 수집하는 개인정보의 양이 늘어났으며, 사용자들의 개인정보 유출 및 프라이버시 침해 우려가 높아졌다. 온라인 서비스 제공자들은 이용자들의 프라이버시 침해 우려를 해소하기 위해 개인정보 처리방침을 제공하고 있으나, 개인정보 처리방침은 길이가 길고 복잡하여 이용자가 직접 위험 항목을 파악하기 어려운 문제로 인해 오남용되는 경우가 많다. 따라서 자동으로 개인정보 처리방침이 안전한지 여부를 검사할 수 있는 방법이 필요하다. 그러나 종래의 블랙리스트 및 기계학습 기반의 개인정보 처리방침 안전성 검증 기법은 확장이 어렵거나 접근성이 낮은 문제가 있다. 본 논문에서는 문제를 해결하기 위해 생성형 인공지능인 GPT-3.5 API를 이용한 개인정보 처리방침 안전성 검증 기법을 제안한다. 새로운 환경에서도 분류 작업을 수행할 수 있고, 전문 지식이 없는 일반인이 쉽게 개인정보 처리방침을 검사할 수 있다는 가능성을 보인다. 실험에서는 블랙리스트 기반 개인정보 처리방침과 GPT 기반 개인정보 처리방침이 안전한 문장과 안전하지 않은 문장의 분류를 얼마나 정확하게 하는지와 분류에 소요된 시간을 측정했다. 실험 결과에 따르면, 제안하는 기법은 종래의 블랙리스트 기반 문장 안전성 검증 기법보다 평균적으로 10.34% 높은 정확도를 보였다.

ABSTRACT

As big data was built due to the 4th Industrial Revolution, personalized services increased rapidly. As a result, the amount of personal information collected from online services has increased, and concerns about users' personal information leakage and privacy infringement have increased. Online service providers provide privacy policies to address concerns about privacy infringement of users, but privacy policies are often misused due to the long and complex problem that it is difficult for users to directly identify risk items. Therefore, there is a need for a method that can automatically check whether the privacy policy is safe. However, the safety verification technique of the conventional blacklist and machine learning-based privacy policy has a problem that is difficult to expand or has low accessibility. In this paper, to solve the problem, we propose a safety verification technique for the privacy policy using the GPT-3.5 API, which is a generative artificial intelligence. Classification work can be performed even in a new environment, and it shows the possibility that the general public without expertise can easily inspect the privacy policy. In the experiment, how accurately the blacklist-based privacy policy and the GPT-based privacy policy classify safe and unsafe sentences and the time spent on classification was measured. According to the experimental results, the proposed technique showed 10.34% higher accuracy on average than the conventional blacklist-based sentence safety verification technique.

Keywords: GPT, Generative AI, Privacy Policy, Blacklist

Received(01. 11. 2024), Modified(1st: 02. 05. 2024, 2nd: 02. 21. 2024), Accepted(02. 23. 2024)

* 본 논문은 2024년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원(P0008703, 2024년 산업혁신인재성장지원사업), 2024년도 과학기술정보통신부 및 정보통신기획평가원의 ICT혁신인재4.0 사업(IITP-2022-RS-2022-00156310)의 지원을 받아 연구되었음.

† 주저자, 220237062@sungshin.ac.kr
‡ 교신저자, iglee@sungshin.ac.kr(Corresponding author)

I. 서 론

4차 산업이 발전하면서 모바일 기기, 사물 인터넷과 같은 통신 장치의 사용률이 급격하게 증가하였고, 온라인 공간을 거쳐가는 데이터의 양이 전례 없는 속도로 커지고 있다. 온라인 서비스 제공자들은 이용자들이 생성한 데이터를 통해 빅데이터를 구축하였고, 빅데이터 분석을 통해 이전보다 질 높은 개인 맞춤형 서비스를 제공할 수 있게 되었다[1]. 이처럼 초연결 사회에서 온라인 서비스는 이용자의 QoE(Quality of Experience)를 높이는 방향으로 점차 발전하고 있으나, 개인 맞춤형 서비스를 제공하기 위한 민감한 개인정보의 활용이 늘어나게 되면서 프라이버시 침해 우려가 높아졌다[2].

온라인 서비스 제공자들은 개인정보 처리방침을 통해 사용하는 개인정보의 범위와 처리 목적 등을 알림으로써 이용자들이 서비스를 신뢰할 수 있도록 한다. 하지만 대다수의 개인정보 처리방침 문서는 길이가 길고 복잡한 형태로 구성되어 있어서, 이용자가 개인정보 처리방침이 포함하고 있는 주요 항목을 파악하기 어려운 문제가 있다. Jonathan A. Obar 등에 따르면, 74%의 사용자는 개인정보 수집·이용·제공에 관한 내용을 읽지 않고 개인정보를 제공하는 것으로 나타났다[3].

일부 온라인 서비스 제공자들은 개인정보 처리방침을 제대로 확인하지 않는 사용자들의 행태를 악용하여 제공하는 서비스에 불필요한 데이터를 수집하거나, 정보를 제3자에게 제공하는 등의 이용자에게 불리한 조항을 개인정보 처리방침에 포함하기도 한다. 일례로 홈플러스에서 진행한 경품 행사는 개인정보 제3자 동의를 작은 글씨로 명시함으로써 고객의 개인정보를 무단으로 사용한 사건이 있었다[4].

이용자가 온라인 서비스에서 자신의 개인정보가 오남용되는 상황을 예방하기 위해 가장 먼저 할 수 있는 조치는 서비스를 이용하기 전에 개인정보 처리방침이 안전한지를 확인하는 것이다. 그러나 긴 길이의 개인정보 처리방침을 사람이 직접 확인하는 작업은 시간적으로 매우 비효율적이며 어렵다[13]. 따라서 자동으로 개인정보 처리방침의 안전성을 검증해 줄 방법이 필요하다.

기존의 블랙리스트 기반 문장 안전성 검증 기법은 악성 조항에 주로 포함되는 어절이 개인정보 처리방침에 있는지를 파악하고, 블랙리스트 항목들과 유사성이 높은 경우를 탐지하여 차단하는 방식이다. 안전하지 않은 문장에 대한 탐지 성공률을 높이기 위해서는 모든 위험한 어절을 포함하는 블랙리스트를 구축해야 한다. 그러나 블랙리스트 기법은 탐지 기준에 따른 정확도의 차이가 상당히 크다는 한계가 있다[12]. 또한 개인정보 처리방침은 온라인 서비스를 제공하는 업체마다 표현이나 구성의 차이를 보인다. 따라서 블랙리스트를 미리 제작하더라도 달라진 문장의 표현으로 인해 안전하지 않은 문장을 탐지하고 분류하기 어려워진다는 문제가 있다.

생성형 인공지능(generative artificial intelligence)은 인간처럼 창의적인 텍스트나 이미지 등의 결과물을 생성할 수 있어서 다양한 디자인, 콘텐츠 제작 등의 산업 분야에서 주목받고 있다[5]. 생성형 인공지능은 기존의 기계학습 및 딥러닝과 다르게 사용자가 학습이나 입력 형식에 맞춰 데이터를 변환하는 작업 대신에, 자연어로 구성된 단순한 질문을 통해 결과를 도출할 수 있는 특징이 있다. OpenAI에서 개발한 딥러닝 기반 자연어 생성 모델GPT(Generative pre-trained transformer)는 생성형 인공지능 분야에서 가장 대중적인 제품으로, GPT-3을 기준으로 1,750억개의 매개변수를 통해 훈련된 자연어처리 모델을 일부 무료로 공개하고 있다[6]. 많은 양의 데이터를 통해 생성된 GPT 모델은 별도의 학습이나 fine-tuning 과정을 거치지 않더라도 기본으로 제공하는 모델만으로도 이상 탐지, 정책 위반 여부에 대한 일정 수준 이상의 결과를 도출할 수 있다. 따라서 GPT를 통해 전문 지식이 없는 일반인도 개인정보 처리방침과 같은 간단한 검사를 수행할 수 있으며, 이에 대한 가능성을 보이기 위해서 본 연구에서는 GPT 모델을 기반으로 개인정보 처리방침의 안전성을 검증하는 기법을 제안한다.

본 논문의 주요 기여점은 다음과 같다.

- GPT 기반의 개인정보 처리방침의 안전성 검증 기법을 제안하고, 종래의 대표적인 자동 검증 기법인 블랙리스트 방법과 성능 및 효율을 비교하는 프레임워크를 제안했다.
- 개인정보 처리방침 데이터셋을 수집하여 실험한 환경에서 종래 방식보다 처리 시간은 증가하지만 정확도를 개선할 수 있음을 입증했다.

본 논문의 구성은 다음과 같다. II장에서는 관련 연구를 비교·분석하고, III장에서는 종래의 기법인 블랙리스트 기반 문장 안전성 검증 기법을 설명한다. IV장에서는 GPT를 활용해 개인정보 처리방침을 검

증하는 GPT 기반 문장 안전성 검증 기법을 제안한다. V장에서는 실험에서 사용한 데이터셋의 생성 방법 및 구성에 관하여 설명하고, VI장에서 실험 결과를 분석하여 제안하는 기법의 성능을 평가하고 VII장에서 결론을 맺는다.

II. 관련 연구

본 장에서는 개인정보 관련 방침의 위험을 탐지하기 위한 주요 선행 연구의 기여점과 한계점을 분석한다. 선행 연구는 인공지능을 이용한 기법과 블랙리스트를 이용한 기법으로 구분된다. Table 1.은 관련 연구의 특징과 한계점을 분석한 내용이다.

Le Yu 등[10]은 안드로이드 애플리케이션 개인정보 처리방침에서 발생할 수 있는 문제점을 자동으로 식별하는 최초의 기법인 PPChecker를 제안한다. PPChecker는 패턴 비교와 세 가지 종류의 블랙리스트를 이용하여 개인정보 처리방침을 검사하며, 안드로이드 애플리케이션의 실제 동작과 개인정보 처리방침의 범위를 대조함으로써 실효성있는 결과를 도출할 수 있는 기여가 있다. 그러나 길고 복잡한 문장의 경우에는 블랙리스트 및 패턴 비교 방식의 한계로 인하여 검사가 불가능하며, 새로운 유형의 문제점이 발생할 경우 탐지율이 낮아진다는 문제가 있다.

Marco Lippi 등[7]은 잠재적인 위험을 포함하고 있는 조항을 자동으로 탐지하는 기계학습 기반의 도구인 CLAUDETTE를 제안한다. CLAUDETTE는 위험한 조항 분류에 SVM(Support Vector Machine)과 HMM(Hidden Markov Model)을 결합한 SVM-HMM을 이용하며, 데이터셋의 레이블링을 통해 잠재적인 악성 조항을 탐지할 수 있다. 그러나 한정된 데이터셋을 이용하기 때문에 이전에는 없었던 불공정 항목이 있으면 분류의 정확도가 감소할 수 있다는 문제가 있다.

Bihui Yu 등[8]은 문장이 설명하는 정책의 분야를 정확하게 결정하기 위한 생성형 인공지능 기반의 정책 분류 알고리즘을 제안한다. BERT를 사용함으로써 정책 영역을 정확하고 효율적으로 판단할 수 있다. 하지만 분류 모델을 미세조정하는 과정에서 높은 품질의 데이터셋이 요구되며, 모델 구축에 큰 비용이 든다는 한계가 있다.

Muhammad Sajidur Rahman 등[9]은 모바일 애플리케이션의 개인정보 정책이 위험한 권한과 일치하는지를 평가하기 위한 자동화된 기계학습 기반 파이프라인인 PermPress를 제안한다. PermPress는 BERT, 로지스틱 회귀, FastText 인공지능 모

Table 1. Features and limitations of related research

Related Research	Feature	Limitation
Le Yu [10]	• Proposed PPChecker, the first technique to automatically identify problems that may arise in Android application privacy policies. • Using pattern comparison and three types of blacklist to examine privacy policies.	• Uncheckable for long and complex sentences • Detection rate decreases when a new type of problem occurs
Marco Lippi [7]	• Propose CLAUDETTE, a machine learning-based tool that automatically detects provisions that contain potential risks • SVM-HMM is used and potential malicious clauses can be detected by labeling the dataset	• Using a limited dataset can reduce the accuracy of classification if there are previously unfair items
Bihui Yu [8]	• Propose a generative AI-based policy classification algorithm to accurately determine the field of policy described by the sentence • Using BERT, a generative AI, accurately and efficiently judges policy areas	• High-quality datasets are required to fine-tune classification models • Model building is expensive
Rahman [9]	• Propose PermPress, an automated machine learning-based pipeline for evaluating whether privacy policies in mobile applications are consistent with risky permissions. • Create your own application policy dataset that maps all risky privileges to privacy policies.	• Generating annotations of datasets manually, building more datasets to train models is expensive

델을 이용한다. 모델의 학습에는 모든 위험한 권한을 개인정보 정책에 매핑하는 애플리케이션 정책 데이터셋을 직접 제작하여 사용하였다. 이를 통해 모든 권한 그룹 및 모든 범주의 안드로이드 애플리케이션에서 개인정보 보호 미준수에 대한 증거를 획득할 수 있었다는 기여가 있다. 그러나 [8]와 마찬가지로 학습을 위한 데이터셋이 필요하며, 데이터셋의 주석을 수동으로 생성하였기 때문에 더 많은 데이터셋을 구축하여 모델을 학습시키는 것에 큰 비용이 든다는 한계점이 있다.

전체적으로 선행 연구에서는 개인정보 관련 방침에 대한 위험을 인공지능 기반 기법을 통해 자동으로 탐지하고 분류할 수 있다. 하지만 일부 작업은 수작업으로 수행해야 하거나, 큰 비용이 들기 때문에 서비스 이용자들과 같은 개인은 사용하기 어렵다는 한계점이 있다.

III. 블랙리스트 기반 문장 안전성 검증 기법

개인정보 처리방침을 검증하기 위한 종래의 기법 중 블랙리스트 기반의 문장 안전성 검증 기법은 인공지능을 사용하지 않고, 문장 구문 분석을 통해 안전성을 검증하는 가장 단순하면서 대표적인 방법이다[10]. 이에 따라 본 논문에서는 제안 기법을 블랙리스트 기반 안전성 검증 기법과 비교한다. 본 논문의 블랙리스트 기반 문장 안전성 검증 기법은 입력된 문장과 블랙리스트 항목을 비교하여 유사도가 임계치 이상으로 발생하는지 여부를 통해 문장의 안전성 여부를 판단하는 기법으로, [10]의 PPChecker를 기반으로 구현했다. 그러나 PPChecker는 애플리케이션을 분석하는 과정과 같이 본 연구에는 불필요한 과정이 포함되기 때문에, PPChecker의 일부 기능만을 이용해 간단한 형식으로 재현했다. 따라서 구현된 블랙리스트 기반 문장 안전성 검증 기법은 블랙리스트의 항목과 개인정보 처리방침의 비교를 통하여 안전한 문장과 안전하지 않은 문장을 구분한다. 블랙리스트 기반 문장 안전성 검증 기법의 흐름은 Fig. 1.과 같다.

먼저, 검증하고자 하는 개인정보 처리방침을 문장 단위로 나누고 한 문장씩 블랙리스트 기반의 문장 안전성 검증 기법에 입력한다. 블랙리스트 기반의 문장 안전성 검증 기법은 입력된 개인정보 처리방침 문장을 블랙리스트에 포함된 모든 항목과 비교하고 유사도를 측정한다. 유사도를 평가하기 위해 입력된 문장을 어절 단위로 비교하여 블랙리스트 항목을 얼마나 많이 포함하고 있는지를 백분율로 나타냈다. 따라서 입력된 문장의 어절과 블랙리스트에 포함된 어절이 일치하는 경우가 많을수록 유사도가 증가한다.

블랙리스트 기반 문장 안전성 검증 기법에서는 입력된 문장에 대한 유사도가 사전에 설정된 임계치보다 높은 수치로 나타날 경우, 안전하지 않은 문장으로 분류하게 된다. 즉, 블랙리스트 항목에 대한 입력 문장의 유사도가 임계치 이상이면 위험한 문장으로 판단한다. 반대로 입력된 문장을 모든 블랙리스트의 항목과 비교했을 때의 유사도가 임계치를 넘지 않는다면 안전한 문장으로 판단한다. 이 과정을 통해 개인정보 처리방침 중 개인정보를 유출하거나 개인의 권리를 침해할 수 있는 위험 문장이 있는지 확인할 수 있다.

블랙리스트 기반 문장 안전성 검증 기법에서 사용하는 블랙리스트는 위험한 조항에 주로 나타나는 어절들의 모음으로 구성된다. 종래의 블랙리스트 기법은 1000개 이하의 항목으로 구성된 블랙리스트를 이용하기 때문에[14, 15], 본 논문의 블랙리스트에 포함되는 각 어절을 500개로 설정했다. 이에 따른 블랙리스트 생성 과정은 다음과 같다.

- 안전하지 않은 개인정보 처리방침 문장에 쓰인 악의적인 어절들을 100개 수집한다.
- GPT-3.5를 통해 의미는 유사하지만 다른 표현으로 기술된 어절을 400개 생성한다.

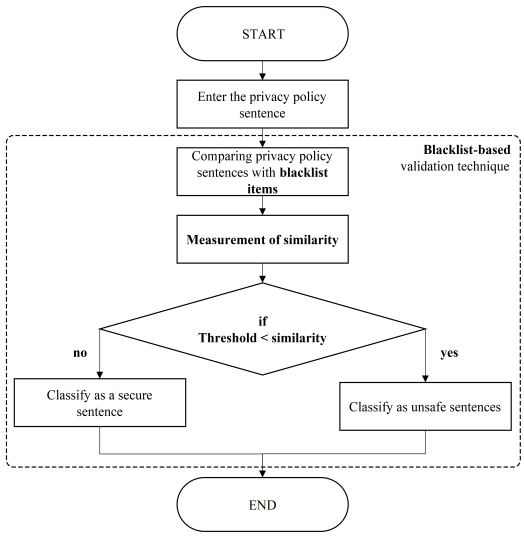

Fig. 1. Blacklist-based validation technique behavior flowchart

- 수집한 어절과 생성한 어절을 종합하여 총 500개의 어절을 통해 블랙리스트를 구성한다.

IV. GPT 기반 문장 안전성 검증 기법

본 논문에서 제안하는 GPT 기반 문장 안전성 검증 기법은 GPT를 이용하여 개인정보 처리방침이 안전한지를 판단하는 방식이다. GPT 기반 문장 안전성 검증 기법의 동작 흐름은 Fig. 2.와 같다.

먼저, 검증하고자 하는 개인정보 처리방침을 문장 단위로 나누어 제안하는 기법의 입력으로 전달한다. GPT 기반 문장 안전성 검증 기법에서는 입력된 문장을 GPT Application Programming Interface(API)에 투입하여 안전한 문장이면 0으로, 안전하지 않은 문장이면 1로 분류하도록 질의한다. 질의에 사용된 API는 GPT-3.5 버전의 text-davinci-003 모델이다. 본 논문에서는 불특정 다수가 쉽게 문장 검증을 수행할 수 있다는 가능성을 보여주기 위해 text-davinci-003 모델을 fine-tuning 과정 없이 사용하였다. GPT 질의에는 입력된 문장과 "이 문장은 안전한가요? 안전하다면 0을 출력해 주시고, 안전하지 않다면 1을 출력해 주세요."라는 질문이 포함된다. API 결과로 0이 출력되는 경우는 안전한 문장으로, 1이 출력되는 경우는 안전하지 않은 문장으로 분류한다.

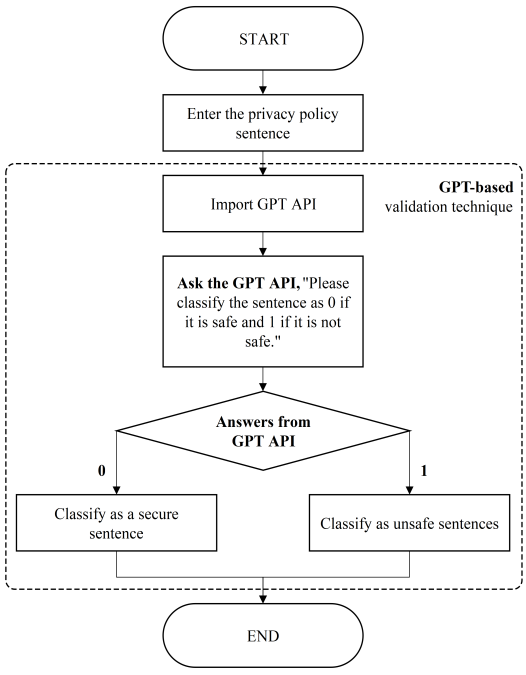

Fig. 2. GPT-based validation technique behavior flowchart

V. 데이터셋

본 논문에서 제안하는 기법 및 블랙리스트 기반 문장 안전성 검증 기법에서 공통으로 사용한 개인정보 처리방침은 GPT-3.5를 통해 생성한 문장의 집합이다. 이러한 개인정보 처리방침 데이터셋은 Fig. 3.과 같은 과정을 통해 생성된다.

먼저 Table 2.의 예시와 같은 안전한 개인정보 처리방침 문장 12개를 GPT-3.5에 입력하고 반대 의미의 문장을 생성하도록 질의한다. 안전한 문장은 서비스 범위 내의 정보를 정당하게 처리를 하면서 서비스 이용자의 권리를 보호하고, 법적인 문제가 없도록 작성된 개인정보 처리방침이다. 위험한 문장은 이용자의 개인정보를 동의없이 제 3자에게 제공하는 것과 같이 서비스 이용자의 권리를 침해하거나 법적으로 문제가 있는 개인정보 처리방침 문장이다. 이때 안전한 개인정보 처리방침은 실제 온라인 서비스 기

Table 2. Examples of safe privacy policy sentences and unsafe privacy policy sentences

Safe sentence	The company immediately removes the personal information collected after the service ends.
Safe sentence generated by GPT	The company immediately removes the personal information collected after the service ends without consent.
	After the service is terminated, the company may own personal information for a certain period of time according to laws and regulations.
Unsafe sentence	The company may permanently own the personal information collected after the end of the service.
Unsafe sentence generated by GPT	The company can safely and permanently own the personal information collected after the end of the service.
	The company will immediately remove some personal information collected after the end of the service.

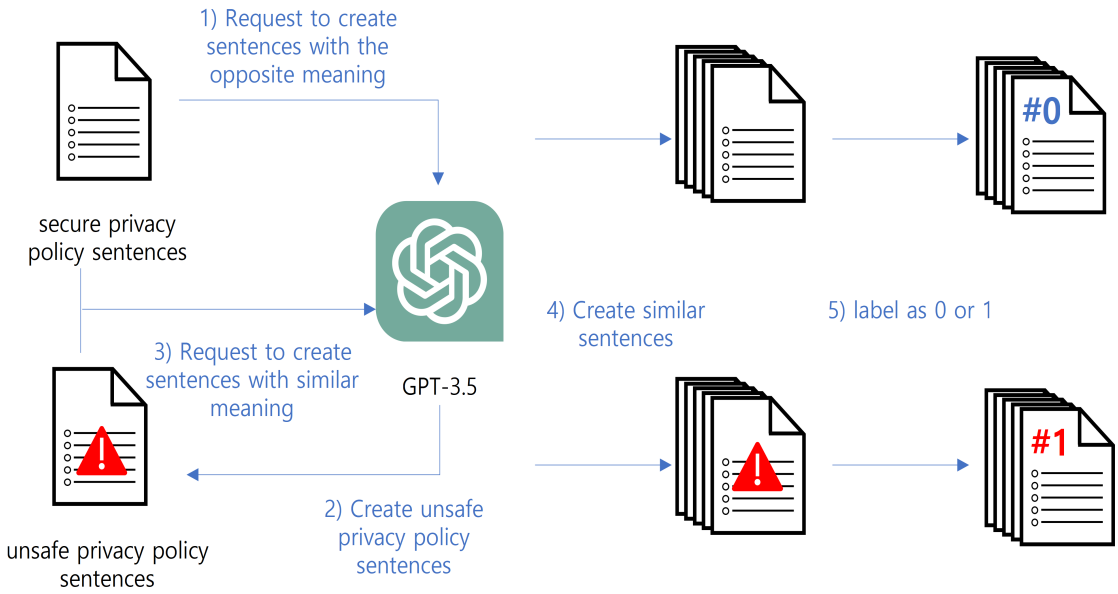

Fig. 3. Process of creating a privacy policy dataset.

업에서 제공해 주고 있는 개인정보 처리방침 문장을 참고하여 생성했다. GPT-3.5는 질의에 대한 응답으로 안전하지 않은 개인정보 처리방침 문장 12개를 결과로 도출한다. 데이터셋의 크기를 늘리기 위해 안전한 문장과 안전하지 않은 문장을 GPT-3.5에 입력하여 연관 있는 다수의 문장을 생성한다. 4,000개의 문장을 생성할 때까지 연관성이 있는 문장을 생성하는 과정이 반복된다. 이후 안전한 개인정보 처리방침 문장을 기반으로 생성된 연관 문장은 0으로 라벨링하고, 안전하지 않은 개인정보 처리방침 문장으로 생성된 문장은 1로 라벨링 한다. 안전한 문장과 안전하지 않은 문장의 예시는 Table 2.와 같다.

블랙리스트 기반 문장 안전성 검증 기법은 입력되는 문장과 블랙리스트의 어절이 일치하는 정도를 통해 안전하지 않은 문장을 구분하기 때문에 블랙리스트 항목의 개수와 입력 데이터에 따라 보이는 성능의 차이가 크다. 본 논문에서는 제안하는 기법과 블랙리스트 기반 문장 안전성 검증 기법의 차이를 비교하기 위해서 2,000개의 데이터셋은 블랙리스트 항목과 연관 있는 개인정보 처리방침 문장으로 구성한다. 나머지 2,000개의 데이터셋은 블랙리스트 항목과 연관이 적은 개인정보 처리방침 문장으로 구성한다.

개인정보 처리방침은 불균형 데이터와 균형 데이터로 나뉜다. 불균형 데이터셋은 안전한 문장과 안전하지 않은 문장의 비율이 균등하지 않은 데이터셋이다. 균형 데이터셋은 안전한 문장과 안전하지 않은 문장이 균형을 이루는 데이터셋을 의미한다. 즉, 데이터셋에 포함된 4,000개의 문장 중 2,000개의 문장은 안전한 개인정보 처리방침 문장이며, 그 외 2,000개 문장은 안전하지 않은 개인정보 처리방침 문장으로 구성된다. GPT-3.5로 생성된 모든 문장과 라벨이 정상적으로 설정되었는지는 4명의 검토자가 직접 검증하였다.

VI. 실험 결과 및 분석

본 장에서는 제안하는 기법과 블랙리스트 기반 문장 안전성 검증 기법에 대한 실험 환경과 실험 결과를 설명한다. 실험 결과는 정확도, F1-score, Precision, Recall 그리고 소요 시간으로 측정하였다.

6.1 실험 환경

실험은 16GB의 메모리와 11th Gen Intel(R) Core(TM) i7-1165G7 @ 2.80GHz를 장착한 단일 컴퓨터에서 수행되었다. 제안하는 GPT 기반 문장 안전성 검증 기법 및 블랙리스트 기반 문장 안전성 검증 기법은 Python 3.10 환경을 기반으로 구현된다. GPT 기반 문장 안전성 검증 기법에서 불러오는 API는 GPT-3.5의 모델인 text-davinci-

003를 사용하였다.

실험에서 사용한 성능지표는 정확도, 정밀도, 재현율 그리고 F1-score며, 해당 성능지표를 통해 GPT 기반 문장 안전성 검증 기법과 블랙리스트 기반 문장 안전성 검증 기법 각각에 대한 성능 점수를 측정하였다. 블랙리스트 기반 문장 안전성 검증 기법의 유사도 임계치는 20%부터 80%까지 20% 단위로 증가시켜 사전에 정확도를 측정하였다. 이후 가장 높은 정확도를 도출한 40%와 60% 임계치 두 가지를 선정하여 블랙리스트 기반 방식의 성능 점수를 측정했다. 블랙리스트 기반 문장 안전성 검증 기법에서 네 가지 성능 지표에 대한 점수는 40% 임계치일 때의 각 점수와 60% 임계치일 때의 점수를 평균내어 표현했다.

6.2 실험 결과

Fig. 4.는 제안 기법과 블랙리스트 기반 문장 안전성 검증 기법에서 불균형 데이터셋을 분류할 때 데이터 개수 증가에 따른 평균 정확도를 측정한 결과이다. 데이터의 증가는 블랙리스트와 유사한 데이터 2,000개에 그 외 데이터를 추가하는 방식으로 진행된다. 추가되는 데이터는 블랙리스트와 연관이 적은 데이터로 총 2,000개이고, 해당 데이터를 1,000개 단위로 증가시키며 정확도를 확인했다.

모든 경우에서 GPT 기반 문장 안전성 검증 기법은 블랙리스트 기반 문장 안전성 검증 기법보다 높은 정확도를 가지며, 최대 12.95%의 성능 차이를 보인다. 또한, 제안하는 기법 및 블랙리스트 기반 문장 안전성 검증 기법 모두에서 새로운 데이터가 추가됨에 따라 평균 정확도가 감소하는 추세를 보였다. 데이터 개수 증가에 따른 증감률은 블랙리스트에서는 약 -13%이고 제안하는 기법에서는 약 -4%로, 제안하는 기법의 낙폭이 상대적으로 적음을 확인하였다.

Fig. 5.는 제안 기법과 블랙리스트 기반 문장 안전성 검증 기법에서 불균형 데이터셋을 검증할 때 데이터 개수 증가에 따른 평균 소요 시간을 측정한 결과이다. Fig. 4.와 동일하게 데이터의 증가는 블랙리스트와 유사한 데이터 2,000개에 그 외 데이터를 추가하는 방식으로 진행된다. 추가되는 데이터는 블랙리스트와 연관이 적은 데이터로 총 2,000개이고, 해당 데이터를 1,000개 단위로 증가시키며 결과를 확인했다.

실험 결과에 따르면 데이터를 새롭게 추가하는 것과 상관없이 제안하는 기법이 블랙리스트보다 높은 소요 시간을 가졌으며, 평균 46.81%의 차이를 보였다. 이는 API를 불러오는 과정의 지연 시간 및 AI 추론 과정에서의 오버헤드로 인해 발생한 것으로 분석되었다.그러나 GPT와 같은 생성형 인공지능은 점차 발전하고 있으며, 결과 도출에 걸리는 시간이 감소하는 추세이다. 따라서 생성형 인공지능이 점차 고도화되면 본 실험에서 발생한 소요 시간 문제는 해결될 것으로 보인다.

Table 3.은 균형 데이터와 불균형 데이터 각각을 검증하는 경우에서 제안하는 GPT 기반 문장 안전성 검증 기법과 블랙리스트 기반 문장 안전성 검증 기법

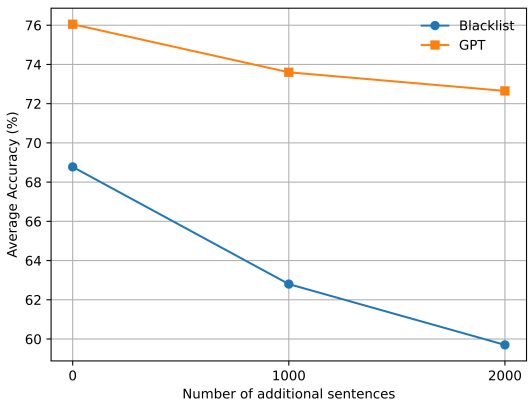

Fig. 4. Average accuracy according to the number of data added by the proposed and blacklist-based validation techniques for verifying unbalanced data

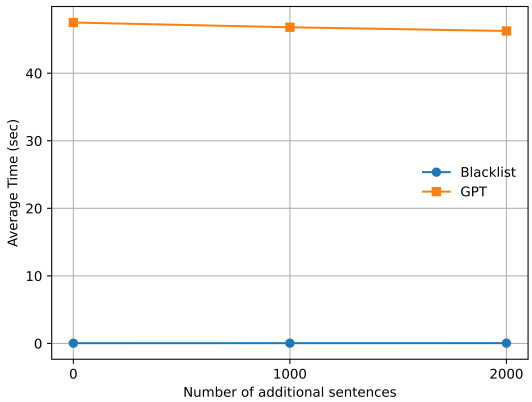

Fig. 5. Average time spent depending on the number of data added by the proposed and blacklist-based validation techniques for verifying unbalanced data

Table 3. Accuracy of proposed and blacklist-based validation technique for verifying balanced and unbalanced data

Method	Accuracy(%)		F1-score		Precision		Recall	
	BD	UD	BD	UD	BD	UD	BD	UD
Blacklist-based validation technique	68.9	68.8	0.568	0.664	0.409	0.617	0.931	0.719
GPT-based validation technique	88.7	76.1	0.873	0.688	0.779	0.527	0.994	0.989

BD: Balanced Data, UD: Unbalanced Data

의 정확도, 정밀도, 재현율 그리고 F1-score를 나타낸다. 데이터셋의 특성에 상관없이 제안하는 기법이 블랙리스트 기반 문장 안전성 검증 기법보다 대부분 높은 성능을 보였다. 제안하는 GPT 기반 문장 안전성 검증 기법은 최대 88.7%의 높은 정확도를 보였으며, 높은 재현율을 유지하였다. 특히 제안하는 기법은 균형 데이터에 대해서 블랙리스트 기반 문장 안전성 검증 기법보다 약 53.87% 높은 F1-score를 나타냄을 확인했다.

실험 결과에 따르면 거의 모든 실험 조건에서 블랙리스트 기반 문장 안전성 검증 기법보다 제안하는 기법이 높은 성능을 보였다. 그러나 개인정보 처리방침을 검증하는데 걸리는 시간이 상대적으로 제안하는 기법에서 높은 수치로 측정되었으며, 이는 본 연구의 한계점이다. GPT API는 약 1,750억 개의 파라미터로 학습되었지만[11], 실험에서 사용한 블랙리스트의 항목의 개수는 500개로 상당히 적은 수이다. 따라서 두 가지 기법에서 분류의 기준이 되는 데이터의 차이가 큰 원인으로 작용해 이와 같은 결과가 발생한 것으로 분석되었다. 따라서 두 가지 기법에서 분류의 기준이 되는 데이터의 차이가 큰 원인으로 작용해 이와 같은 결과가 발생한 것으로 분석되었다.

결론적으로 제안하는 GPT 기반 문장 안전성 검증 기법은 블랙리스트 기반 문장 안전성 검증 기법과는 달리, 새로운 데이터 검증과 같은 상황에 대한 영향을 적게 받는다, 따라서 시간은 많이 소요되더라도 다양한 상황에서 좋은 분류 성능을 가짐을 알 수 있다.

VII. 결 론

본 연구에서는 GPT를 이용해 개인정보 처리방침의 안전성 여부를 검증하는 GPT 기반 문장 안전성 검증 기법을 제안했다. 실험 결과에 따르면, 제안하는 GPT 기반 문장 안전성 검증 기법은 종래의 블랙리스트 기반 문장 안전성 검증 기법보다 평균적으로 10.34% 높은 검증 정확도를 보였다. 특히 종래의 블랙리스트 기반 문장 안전성 검증 기법은 블랙리스트에 없는 항목은 검증하지 못하지만, 제안하는 기법은 새로운 개인정보 처리방침 문장에 대해서도 상대적으로 높은 정확도로 검증할 수 있다는 이점이 있다. 또한, GPT를 이용해 개인정보 처리방침의 안전성을 검증함으로써 실생활의 다양한 분야에 대한 검증을 GPT를 통해 수행할 수 있다는 가능성을 보였다.

실험 결과에 따르면 제안하는 기법은 상대적으로 블랙리스트 기반 문장 안전성 검증 기법보다는 높은 정확도를 가졌지만, 최소 72.65%로 높지 않은 정확도를 보였다. 검증하는데 걸리는 시간의 증가율은 종래의 기법 대비 평균 46.85% 정도로 상당히 높았다. 하지만 처리 시간의 증가는 탐지 정확도를 개선하고 새로운 패턴을 자동으로 탐지할 수 있는 유연한 안전성 검증 기법을 실현하기 위해 소모되는 비용임을 알 수 있다. 본 연구의 성능 실험은 블랙리스트 기반 문장 안전성 검증 기법의 블랙리스트 항목 개수의 변화에 따른 결과를 분석하지 못했다. 또한 개인정보 처리방침의 변형에 대한 검사는 가능함을 보였지만, 실제 기업 및 기관이 제공하는 개인정보 처리방침이 안전한지에 대한 검증을 하지 못했다는 한계가 있다. 향후 연구에서는 제안하는 기법의 실효성을 입증하기 위해 다양한 종류의 공공기관 및 기업의 개인정보 처리방침을 수집한 후 안전성 검증 과정을 수행하고, 실제 기업들의 안전성을 평가하고자 한다. 또한 생성형 인공지능을 통해 개인정보 처리방침뿐만 아니라 다양한 분야 및 다양한 환경에서 높은 성능으로 검증할 수 있는 기법을 연구할 계획이다.

References

[1] M.C. Cohen, "Big data and service operations," Production and Operations Management, vol. 27, no. 9, pp. 1709-1723, Dec. 2018.

[2] P. Silva, C. Gonçalves, C. Godinho, N. Antunes and M. Curado, "Using nlp and machine learning to detect data privacy violations," In IEEE INFOCOM 2020-IEEE conference on computer communications workshops, pp. 972-977, Jul. 2020.

[3] J.A. Obar and A. Oeldorf-Hirsch, "The biggest lie on the Internet: Ignoring the privacy policies and terms of service policies of social networking services," Information, Communication & Society, vol. 23, no. 1, pp. 128-147, Jul. 2018.

[4] Gyeo-Cheol Lim, "Studying on commercial sales of personal information and conditions - Focused on the Home-Plus case of the Supreme Court," Korea Legislation Research Institute, (52), pp. 273-301, May. 2017.

[5] M. Jovanovic and M. Campbell, "Generative artificial intelligence: Trends and prospects," Computer, vol. 55, no. 10, pp. 107-112, Oct. 2022.

[6] L. Floridi and M. Chiriatti, "GPT-3: Its nature, scope, limits, and consequences," Minds and Machines, vol. 30, pp. 681-694, Nov. 2020.

[7] M. Lippi, P. Pałka, G. Contissa, F. Lagioia, H.W. Micklitz, G. Sartor and P. Torroni, "CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service," Artificial Intelligence and Law, vol. 27, pp. 117-139, Feb. 2019.

[8] B. Yu, C. Deng and L. Bu, "Policy text classification algorithm based on BERT," International Conference of Information and Communication Technology, pp. 488-491, Feb. 2022.

[9] M.S. Rahman, P. Naghavi, B. Kojusner, S. Afroz, B. Williams, S. Rampazzi and V. Bindschaedler, "Permpress: Machine learning-based pipeline to evaluate permissions in app privacy policies," IEEE Access, vol. 10, pp. 89248-89269, Aug. 2022.

[10] L. Yu, X. Luo, X. Liu and T. Zhang, "Can we trust the privacy policies of android apps?," Annual IEEE/IFIP International Conference on Dependable Systems and Networks, pp. 538-549, Jun. 2016

[11] R. Dale, "GPT-3: What's it good for?," Natural Language Engineering, vol. 27, no. 1, pp. 113-118, Jan. 2021.

[12] Chae-rim Han, Su-hyun Yun, Myeong-jin Han, Il-Gu Lee, "Machine Learning-Based Malicious URL Detection Technique," Journal of The Korea Institute of Information Security & Cryptology, 32(3), pp. 555-564, Jun. 2022.

[13] F. Xie, Y. Zhang, C. Yan, S. Li, L. Bu, K. Chen, Z. Huang and G. Bai, "Scrutinizing Privacy Policy Compliance of Virtual Personal Assistant Apps," Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering, pp. 1 - 13, Oct. 2022.

[14] M.M. Swe and N.N. My, "Fake Accounts Detection on Twitter Using Blacklist," 2018 IEEE/ACIS 17th International Conference on Computer and Information Science, pp. 562-566, Sep. 2018.

[15] R.S. Rao and A.R. Pais, "An Enhanced Blacklist Method to Detect Phishing Websites," Information Systems Security: 13th International Conference, pp. 323 - 333, Dec. 2017.

〈저자소개〉

심 혜 연 (Hye-Yeon Shim) 학생회원
2021년 8월: 성신여자대학교 융합보안공학과 졸업
2023년 2월: 성신여자대학교 미래융합기술공학과 석사 졸업
2023년 9월~현재: 성신여자대학교 미래융합기술공학과 박사
〈관심분야〉 기계학습, 생성형 인공지능, 융합보안, 정보보호

권 민 서 (Min-Seo Kweun) 학생회원
2021년 3월~현재: 성신여자대학교 융합보안공학과 학사
2023년 3월~현재: 성신여자대학교 CSE LAB 연구원
〈관심분야〉 디지털 포렌식, 융합보안, 정보보호

윤 다 영 (DaYoung Yoon) 학생회원
2021년 3월~현재: 성신여자대학교 융합보안공학과 학사
2023년 3월~현재: 성신여자대학교 CSE LAB 연구원
〈관심분야〉 융합보안, 개인정보, 정보보호

서 지 영 (JiYoung Seo) 학생회원
2021년 3월~현재: 성신여자대학교 융합보안공학과 학사
2023년 3월~현재: 성신여자대학교 CSE LAB 연구원
〈관심분야〉 정보보호, 융합보안, 악성코드

이 일 구 (Il-Gu Lee) 종신회원
2003년 2월: 서강대학교 전자공학과 졸업
2005년 2월: KAIST 정보통신대학원 석사
2016년 2월: KAIST 전산학부 박사
2005년 2월~2017년 2월: 한국전자통신연구원 5G기가통신시스템연구본부 선임연구원
2017년 3월~현재: 성신여자대학교 미래융합기술공학과/융합보안공학과 부교수
〈관심분야〉 융합보안, 미래융합기술, 정보통신