한국융합보안학회 2025년 추계학술대회 논문집

다중 문서 포맷을 지원하는 개인정보 비식별화 방법론

백서진[1]*, 박윤지[1], 정보남[1], 함근희[1], 이병천[1]
중부대학교[1]

A Methodology for Multi-Format Document De-Identification

Seojin Baek[1]*, Yunji Park[1], Bonam Jung[1], Geunhee Ham[1], Byoungcheon Lee[1]

요약 : 본 연구는 다양한 문서 포맷(HWP, DOC, XLS, PPT, PDF 등)을 대상으로 작동하는 개인정보 비식별화 도구를 설계하였다. 각 포맷의 OLE 및 OOXML 구조를 분석하여 텍스트를 추출하고, 정규표현식·유효성 검증·KLUE-BERT 기반 NER을 결합해 개인정보를 탐지한다. 탐지된 항목은 원본 좌표를 기준으로 동일 길이의 '*'로 치환되어 문서 구조를 유지한 채 비식별화된다. 이 도구는 포맷에 구애받지 않고 일관된 개인정보 제거를 수행할 수 있다.

Key Words : De-identification, Redaction, NER (Named Entity Recognition), OLE (Compound File Binary Format), XML (Extensible Markup Language)

1. 서 론

문서에는 성명, 연락처, 주민등록번호 등 개인정보가 포함된 채로 작성·저장되는 경우가 많다. 이러한 정보가 가려지지 않은 상태로 재사용되거나 외부로 공유되면 「개인정보 보호법」 제3조 제4항 및 제17조에 따라 법적 책임이 발생할 수 있다. 이에 따라 개인정보를 비식별화하는 다양한 도구들이 등장했으나 [1], 각 도구가 지원하는 문서 포맷이 상이하다. 일부는 Docx 파일만 처리하고, 일부는 PDF나 HWP 포맷만 지원하는 등 모든 문서 형식을 일괄적으로 처리할 수 있는 통합 도구는 부재하다.

본 연구는 이러한 한계를 해결하기 위해 문서 포맷별 구조를 직접 분석하고, 여러 형식을 하나의 통합 체계 내에서 비식별화할 수 있는 범용 문서 비식별화 도구를 제안한다.

2. 본 론

2.1 지원 대상 포맷

본 연구의 비식별화 도구는 PDF를 포함해 대중적으로 사용하는 OLE(Compound File Binary Format) 와 XML(Office Open XML) 기반 문서 포맷을 모두 지원한다. OLE 계열에는 HWP, DOC, PPT, XLS를 지원하며, XML 계열에는 HWPX, DOCX, PPTX, XLSX가 포함된다.[2]

2.2 처리 흐름

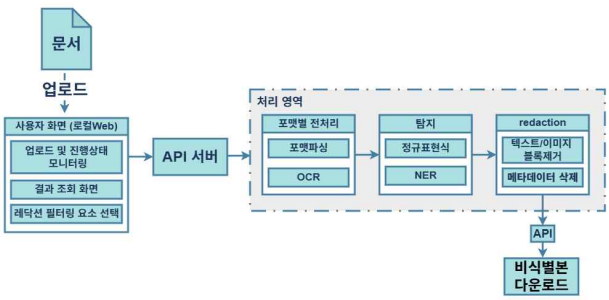

Fig. 1. Document De-identification Flow

본 도구의 전체 처리 과정은 (그림 1)과 같이 문서 입력부터 비식별화 결과 생성까지의 일련의 절차로 구성된다. 사용자가 로컬에서 동작하는 웹 인터페이스를 통해 PDF, MS Office, HWP/HWPX 등 다양한 문서를 업로드하면, API 서버가 파일 형식을 자동 판별하여 해당 포맷 전용 파서를 호출하고 비식별화 절차를 수행한다. 처리 과정은 포맷별 추출 → 정규화 → 개인정보 탐지 → 비식별화(redaction) 순으로 이루어진다.

사용자는 웹 인터페이스에서 비식별화(레닥션)된 결과를 확인할 수 있으며, 개인정보 항목별로 처리 여부를 직접 선택할 수도 있다. 기본적으로 주민등록번호, 전화번호, 이메일, 운전면허번호, 여권번호, 지역번호가 비식별화 대상으로 설정되어 있으며, 필요시 체크박스를 해제해 제외할 수 있다. 이를 통해 문서 내 어떤 정보가 비식별화되었는지 명확히 확인하고, 목적에 따라 범위를 조정할 수 있다.

2.2.1 XML 문서 추출 방식

XML 계열 파이프라인은 OOXML(DOCX, XLSX, PPTX) 과 HWPX 문서를 바탕으로[2], 컨테이너(ZIP

구조)를 열어 본문, 주석, 문서 속성 등 주요 XML 파일을 순회한다. OOXML은 document.xml, sharedStrings.xml, slides/*.xml, comments*.xml, docProps/*.xml 등을, HWPX는 Contents/*.xml을 중심으로 텍스트를 추출한다. 문서 구조 정보를 함께 보존하여, 비식별화 단계에서 정확히 치환할 수 있도록 설계했다.

2.2.2 OLE 포맷 문서 추출 방식

OLE 기반의 HWP, DOC, PPT, XLS 문서는 스토리지(storage) 내 스트림(stream) 단위로 파싱된다. HWP는 BodyText와 DocInfo 스트림에서 본문과 속성 정보를 추출하며, DOC는 WordDocument와 0/1Table 스트림의 피스 테이블(PlcPcd)을 이용해 조각난 텍스트를 복원한다. PPT는 TextCharsAtom 스트림에서 슬라이드별 텍스트를, XLS는 Workbook 내 SST(Shared String Table)과 CONTINUE 레코드를 해석하여 내부에 저장된 셀 데이터를 추출한다.

2.2.3 개인정보 탐지 방식

탐지 모듈은 정규표현식·유효성 검증 기반 탐지와 KLUE-BERT NER 기반 문맥 탐지를 병행하여 개인정보를 식별한다. 두 방식은 상호 보완적으로 동작하며, 형식이 명확한 데이터는 정규표현식으로, 패턴이 불규칙한 개체는 NER을 통해 문맥적으로 인식한다.

2.2.3.1 정규표현식 및 유효성 검증

정규표현식은 주민등록번호, 전화번호, 이메일, 운전면허번호 등 형식이 일정한 개인정보를 신속하게 탐지하는 데 사용된다. 탐지된 후보는 각 항목별 유효성 검증(Validation) 절차를 거쳐 오탐(False Positive)을 최소화한다. 예를 들어, 주민등록번호는 체크섬 검증, 운전면허번호는 발급년도·지역코드 검증, 이메일은 도메인 형식 확인을 수행한다. 이 과정을 통해 형식만 유사한 비개인정보 문자열이 잘못 탐지되는 문제를 방지한다.

2.2.3.2 KLUE-BERT 기반 NER

문서 본문에서 이름, 기관명, 시설·장소처럼 문맥에 의존하는 비정형 개체를 식별하고자 KLUE-BERT 계열 사전학습 모델[3] 위에 토큰 분류 헤드를 결합해 미세조정을 수행하였다. 서브워드 분절로 인한 불일치를 줄이기 위해 단어 시작 위치에만 라벨을 부여하는 BIO 정렬을 적용했고, 학습 주기마다 검증을 실시해 정밀도·재현율을 산출한 뒤 개발용 검증 기준 최고 성능 시점의 가중치를 최종 모델로 사용하였다. 모델 출력 라벨을 연속 구간으로 변환해 엔터티로 복원하고, 경계를 정제하여 원문 오프셋과 결합해 문서 전체에 일관되게 반영하였다. seqeval 기반 엔터티 단위 F1·정밀도·재현율로 수행하였으며, 검증 셋에서 F1 0.785, 정밀도 0.777, 재현율 0.792를 기록했다.

2.2.4 비식별화(Redaction) 처리

탐지된 개인정보는 정규화된 텍스트와 원본 데이터 간의 매핑 정보를 기반으로 비식별 처리된다. 텍스트 추출 단계에서 각 문서 객체의 오프셋 정보를 보존하기 때문에, 탐지된 문자열의 위치를 원본 문서 내 좌표로 역추적할 수 있다. 이후 해당 구간을 동일 길이의 '*' 문자열로 치환하여 비식별화를 수행하며[1], 문서의 레이아웃과 구조는 그대로 유지된다.

2.2.5 PDF 비식별화 처리

PDF 문서는 PyMuPDF 라이브러리를 사용하여 처리한다. 텍스트 레이어가 포함된 PDF의 경우, PyMuPDF를 통해 페이지 단위로 텍스트 블록을 추출하고, 각 블록의 위치좌표와 함께 문자열 데이터를 탐지 모듈로 전달한다.

3. 검증

본 도구의 탐지 성능을 검증하기 위해, 총 52개의 개인정보 데이터셋(하이픈 포함 25개, 미포함 27개)을 구성하여 HWP, DOCX, XLSX, PPTX, PDF 등 모든 지원 포맷에서 테스트를 수행하였다. 정규표현식 및 유효성 검증 기반 탐지 모듈을 적용한 결과, 하이픈이 포함된 항목의 경우 단 1건의 미탐을 제외하고 모든 항목이 정확히 인식되었으며, 하이픈이 없는 경우 11건의 오탐이 발생하였다. 이는 하이픈 유무로 인한 문자열 패턴 차이에서 비롯된 결과로, 본 검증은 정규표현식 기반 탐지만을 대상으로 하였으며 NER 병행 시 탐지 정확도 향상이 기대된다.

4. 결론

본 연구는 여러 문서 포맷을 통합적으로 비식별화할 수 있는 비식별화 도구를 구현하고, 정규표현식 기반 탐지 성능을 검증하였다. 실제 데이터셋 실험을 통해 포맷 간 일관된 탐지 결과를 확인하였으며, 본 도구는 기업·기관의 문서 보안 자동화에 활용될 수 있다. 향후 NER 기반 문맥 분석을 결합해 비정형 정보 처리 범위를 확장할 예정이다.

참고문헌

[1] Garfinkel, S. L., "De-Identification of Personal Information," NISTIR 8053, National Institute of Standards and Technology, 2015.

[2] Rohlmann, S., Mladenov, V., Mainka, C., Hirschberger, D., and Schwenk, J., "Every Signature is Broken: On the Insecurity of Microsoft Office's OOXML Signatures," Proceedings of the 32nd USENIX Security Symposium, 2023.

[3] 이민호, "사전학습 언어모델을 이용한 논술 구성요소 분석방법," 한국콘텐츠학회 2022 국내종합학술대회 논문집, 2022, pp. 167-168.