감정 분석 기반 약물 검색 시스템

김한울*O, 왕성훈*, 윤태준*, 김대영*
*순천향대학교 컴퓨터소프트웨어공학과
e-mail : hkimog13@sch.ac.kr*O, {sunghun051107, 20214004, dyoung.kim}@sch.ac.kr*

Sentiment Analysis-based Drug Search System

Hanul Kim*O, Sunghun Wang*, Tae Jun Yoon*, Dae-Young Kim*
*Dept. of Computer Software Engineering , Soonchunhyang University

요 약

약물을 복용하는 환자들은 단순한 약물 정보뿐 아니라 실제 사용 후기, 잠재적 부작용과 같은 경험적 정보를 필요로 한다. 이에 본 연구는 사용자 리뷰 데이터를 분석하여 객관적인 정보를 제공하는 감정 분석 기반 약물 검색 시스템을 설계하고 개발했다. Drug Performance Evaluation(약물 성능 평가) 데이터셋과 Drugs, Side Effects(약물 부작용) 데이터셋을 병합하여 총 33,724 개의 데이터셋을 구축하였다. 구축된 데이터셋의 리뷰 텍스트는 경량화된 DistilBERT 모델을 활용해 긍정/부정 여부를 분류하였으며, 이를 기반으로 사용자 증상에 최적화된 약물을 추천하는 웹 서비스를 구현하였다. 본 시스템은 실제 복용 경험에 기반한 신뢰도 높은 정보와 부작용 정보를 통합적으로 제공함으로써, 환자들의 안전한 약물 선택과 부작용 최소화에 기여할 것으로 기대된다.

키워드 : 감정 분석, 약물 검색 시스템, 리뷰 분석, NLP, DistilBERT

I. 서론

약물을 복용하는 환자들은 약물의 성분이나 효능 정보뿐 아니라, 실제 사용자들의 경험과 후기, 복용 시 발생할 수 있는 부작용 등 다양한 정보를 필요로 한다. 선행 연구에 따르면, 인터넷을 통해 약물 정보를 얻는 환자들의 주된 목적은 약물의 잠재적인 부작용(Side effects)을 확인하는 것으로 나타났다 [1]. 이는 환자들이 실제 복용 경험자들의 후기를 통해 다양한 정보를 얻고 싶어 함을 의미한다. 하지만 현재 인터넷상의 정보는 방대하고 비정형화되어 있어 약물에 대한 정보를 체계적으로 파악하기 어렵다는 문제가 존재한다.

따라서, 본 연구에서는 약물 리뷰 데이터에 DistilBERT [2] 모델을 활용하여 감정 분석을 적용하고, 이를 약물 부작용 데이터와 통합하여 사용자의 증상으로부터 적합한 약물을 검색할 수 있는 감정 분석 기반 약물 검색 시스템을 설계하고 개발했다. 해당 시스템은 사용자가 입력한 증상에 적합한 약물을 긍정 리뷰 순으로 추천하며, 각 약물의 상세 정보와 긍정/부정으로 분류된 사용자 리뷰를 함께 제공한다. 이를 통해 사용자들은 실제 사용자 경험을 바탕으로 객관적인 약물 정보를 획득함으로써 본인에게 적합한 약물을 선택할 수 있다.

II. 관련 연구

2.1 DistilBERT

DistilBERT는 BERT 모델에 지식 증류 기법(Knowledge Distillation)을 적용한 경량화 모델로, BERT 대비 40% 더 적은 파라미터와 60% 더 빠른 처리 속도를 가지면서도 원본 모델 성능의 97%를 유지한다. 본 연구에서는 Hugging Face에서 제공하는 사전 학습된

DistilBERT-base-uncased-finetuned-sst-2-english 모델을 활용하였다. 이 모델은 SST-2(Stanford Sentiment Treebank 2) 데이터셋으로 긍정/부정 이진 감정 분류에 특화되어 있어 별도의 학습 과정 없이 약물 리뷰 분석에 적용 가능하며, 경량화된 구조로 인해 대량의 리뷰 데이터를 효율적으로 처리할 수 있다는 장점이 있다.

2.2 활용 데이터셋

본 연구에서는 약물 리뷰 정보와 부작용 정보를 통합적으로 제공하기 위해 두 가지 공개 데이터셋을 활용하였다. Drug Performance Evaluation 데이터셋 [3]은 약물 사용자들의 리뷰를 포함한 데이터셋으로, 총 53,766개로 구성되어 있으며 약물명, 증상, 리뷰 텍스트 등 7개의 속성으로 이루어져 있다. 해당 데이터셋은 다양한 약물에 대한 사용자들의 리뷰를 담고 있어 감정 분석을 통한 약물 추천 시스템 구축에 적합하다.

Drugs, Side Effects 데이터셋 [4]은 약물별 부작용에 대한 정보를 포함하는 데이터셋으로, 총 2,931개의 데이터로 구성되어 있으며 약물명, 부작용 등 14개의 속성으로 이루어져 있다. 이 데이터셋은 환자들이 가장 중요하게 여기는 부작용 정보를 제공하여 안전한 약물 선택을 지원한다.

III. 시스템 구현

3.1 데이터 전처리

본 연구에서는 두 데이터(D_1, D_2) 의 공통 특성인 약물명을 기준으로 병합해 통합 데이터셋(D_{new})을 구축했다. 병합 과정은 Python의 Pandas 라이브러리 [5]를 활용하여 수행하였으며, D_1을 기준으로 D_2의 레코드를 조인(join)하는 방식으로 진행했다.

$$D_{new} = \left\{ x_i \oplus y_j | x_i \in D_1, y_j \in D_2 \text{ and } x_i^1 = y_j^1 \right\} \quad (1)$$

또한, 웹 스크래핑으로 수집된 리뷰 데이터에는 HTML 엔티티가 포함된 데이터가 다수 존재했다. 이러한 HTML 데이터를 제거하기 위해 Python의 BeautifulSoup 라이브러리 [6]를 사용하여 표 1에 나타난 바와 같이 HTML 엔티티를 일반 텍스트로 정제하는 전처리를 수행했다.

표 1. 리뷰 데이터 전처리 예시

original data	refined data
"I've tried a few antidepressants over th..."	"I've tried a few antidepressants over th..."

이후, 전처리된 통합 데이터셋의 리뷰 텍스트에 대해 감정 분석을 수행했다. Hugging Face Transformers 라이브러리의 sentiment-analysis 파이프라인 [7]을 사용하여 사전 학습된 DistilBERT 모델에 각 리뷰를 입력하였으며, 모델은 각 리뷰를 긍정(POSITIVE) 또는 부정(NEGATIVE)으로 분류하고 신뢰도 점수를 함께 반환한다. 표 2는 감정 분석 결과의 예시를 보여준다.

표 2. 감정 분석 결과 예시

review	sentiment	sentiment_score
"Quick reduction of symptoms"	NEGATIVE	0.542751
"I' have been on this birth control for one cyc ...	POSITIVE	0.995449

분류된 감정 정보는 데이터셋에 새로운 속성으로 추가되어 데이터베이스에 저장되었다. 이러한 전처리 과정을 통해 약물 정보, 리뷰 데이터, 부작용 정보, 감정 분석 결과를 포함하는 총 33,724개의 통합 데이터셋을 구축했다.

3.2 시스템 아키텍처

본 시스템은 크게 서버와 클라이언트로 구성되며, 서버는 감정 분석 모듈, 데이터베이스(DB), API 서버로 이루어진다.

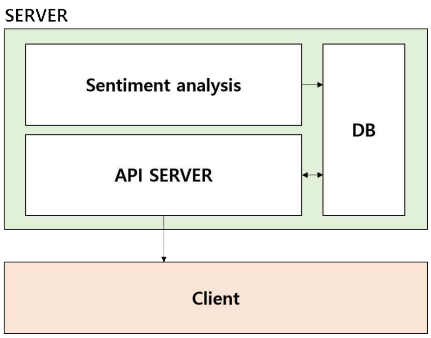

그림 1. 시스템 구조

감정 분석 모듈은 수집된 원천 데이터를 정제하고 DistilBERT 모델을 사용해 모든 리뷰의 감정을 분석한 후, 그 결과를 DB에 저장한다. DB는 약물 정보, 리뷰, 텍스트, 감정 분석 결과, 부작용 정보 등 통합 데이터셋의 모든 정보를 저장하며, API 서버의 요청에 따라 필요한 데이터를 제공한다. API 서버는 클라이언트의 요청을 처리하여 DB로부터 적절한 정보를 조회하고 응답한다. 사용자가 증상을 입력하거나 약물명을 검색하면, 해당하는 약물 목록과 상세정보, 긍정/부정 리뷰를 제공한다. 클라이언트는 웹 기반 사용자 인터페이스로 구현되었으며, 사용자의 검색 요청을 서버에 전달하고 결과를 시각적으로 표시한다.

3.3 웹 인터페이스

구현된 웹 서비스는 사용자 중심의 직관적인 인터페이스를 제공하며, 두 가지 검색 방식을 지원한다. 먼저 '증상 기반 검색'은 사용자가 자신의 증상을 입력하면, 해당 증상에 적합한 약물 목록을 제공한다. 약물 목록은 긍정 리뷰가 많은 순으로 정렬되어 표시되며, 각 약물 카드에는 약물명, 평균 평점, 긍정/부정 리뷰 개수가 함께 표시된다. 사용자가 특정 약물을 선택하면 상세 정보 페이지로 이동해 약물의 기본 정보, 관련 증상, 부작용 정보와 함께 긍정/부정으로 분류된 실제 사용자 리뷰를 확인할 수 있다.

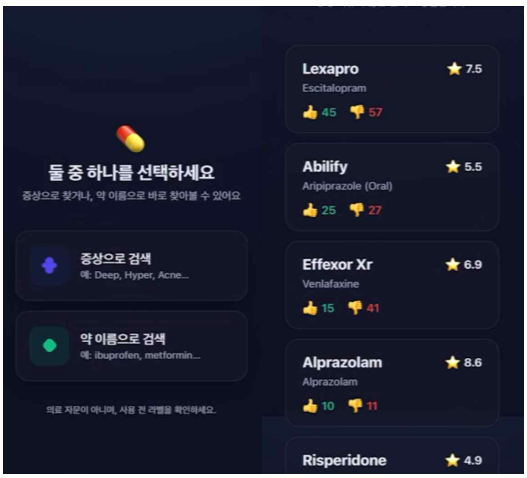

그림 2. 홈화면 및 증상 기반 검색 페이지

반면 '약물명 기반 검색'은 사용자가 특정 약물명을 직접 입력하는 방식으로, 증상 선택 단계 없이 해당 약물의 상세 정보 페이지로 즉시 이동한다. 상세 정보 페이지는 증상 기반 검색과 동일한 형식으로 구성되어 있으며, 약물의 기본 정보, 의학 정보, 감정 분석 결과에 따라 분류된 사용자 리뷰를 제공한다. 이를 통해 사용자는 자신이 찾고자 하는 약물에 대한 평가를 빠르게 확인할 수 있다.

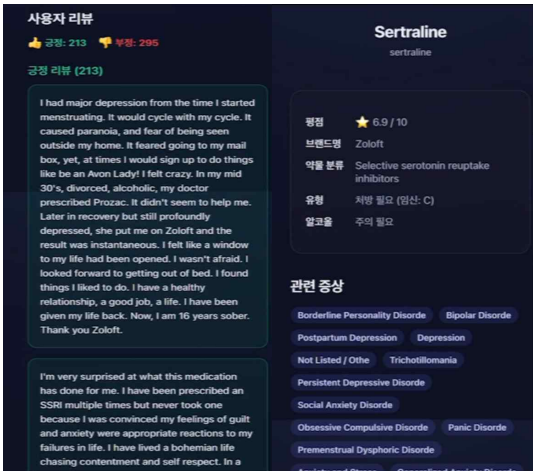

그림 3. 약물 상세 페이지

IV. 결론

본 연구에서는 약물 리뷰 데이터에 대한 감정 분석과 부작용 정보를 통합하여 사용자 중심의 지능형 약물 검색 시스템을 설계 및 구현하였다. 수집한 두 데이터셋을 공통된 약물명 기준으로 병합해 총 33,724개의 통합 데이터셋을 구축하였으며, 사전 학습된 DistilBERT 모델을 활용해 모든 리뷰에 대한 감정 분석을 수행했다. 이를 기반으로 증상 기반 검색과 약물명 기반 검색을 지원하는 웹 서비스를 개발하였으며, 긍정/부정으로 분류된 사용자 리뷰와 약물의 상세 정보를 통합적으로 제공한다.

본 시스템은 환자들이 실제 사용자 경험을 바탕으로 객관적이고 종합적인 약물 정보를 획득할 수 있도록 지원하며, 본 논문에서 사용한 데이터 분석 기법은 향후 환자들의 실제 경험을 기반으로 하여 의료 데이터 분석에 기여할 수 있다. 더 나아가, 환자 중심의 디지털 헬스케어 서비스 확장에 활용될 수 있을 것이다.

참고문헌

[1] Bergmo, Trine Strand, et al. "Internet use for obtaining medicine information: cross-sectional survey." *JMIR Formative Research*, 7(1), 2023.

[2] Sanh, Victor, et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv preprint arXiv:1910.01108, 2019).

[3] The Devastator. 2022. Drug Performance Evaluation. kaggle.

※ "본 연구는 2025 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구 결과로 수행되었음"(2021-0-01399)

https://www.kaggle.com/datasets/thedevastator/drug-performance-evaluation
[4] Jithin Varghese. 2022. Drugs, Side Effects and Medical Condition. kaggle. https://www.kaggle.com/datasets/jithinanievarghese/drugs-side-effects-and-medical-condition
[5] Pandas Development Team. (2025). https://pandas.pydata.org/
[6] Richardson, L. (2025). https://www.crummy.com/software/BeautifulSoup/
[7] Hugging Face. (2025). https://huggingface.co/docs/transformers/main/ko/main_classes/pipelines