

# 소셜 네트워크 서비스 데이터에서 Bi-LSTM 기반 약물 부작용 게시물 탐지 모델 연구

이충천<sup>1</sup>, 이승희<sup>2</sup>, 송미화<sup>3</sup>, 이수현<sup>1,2</sup>

<sup>1</sup> 건양대학교 정보의학교실

<sup>2</sup> 건양대학병원 헬스케어데이터사이언스센터

<sup>3</sup> 세명대학교 스마트 IT 학부

dlcndcjs91@naver.com, shleedynamics@kyuh.ac.kr, mhsong@semyung.ac.kr, shleemedi@kyuh.ac.kr

## A Study on Bi-LSTM-Based Drug Side Effects Post Detection Model in Social Network Service Data

Chung-Chun Lee<sup>1</sup>, Seunghee Lee Ph.D<sup>2</sup>, Mi-Hwa Song Ph.D<sup>3</sup>, Suehyun Lee Ph.D<sup>1</sup>

<sup>1</sup>School of Biomedical Informatics, Konyang University

<sup>2</sup>Healthcare Data Science Center, Konyang University Hospital

<sup>3</sup>School of Smart IT, Semyung University

### 요 약

본 연구에서는 소셜 네트워크 서비스(Social Network Service, SNS) 데이터로부터 약물 부작용 게시글을 추출하기 위한 순환 신경망(Recurrent Neural Network, RNN) 기반 분류 모델을 제안한다. 먼저, 처방 빈도가 높으며 게시글을 많이 확보할 수 있는 케토프로펜 약물에 대하여 국내 최대 소셜 네트워크 플랫폼인 네이버 블로그와 카페의 게시글(2005년~2020년)을 확보하고 최종 3,828건을 분석하였다. 결과적으로 케토프로펜에 대한 3종(약물, 부작용, 불용어)의 렉시콘을 정의하였으며 이를 기반으로 Bi-LSTM 분류모델 기준 87%의 정확도를 얻었다. 본 연구에서 제안하는 모델은 SNS 데이터가 약물 부작용 정보 획득을 위한 기존(전자처방기록, 자발적 약물 부작용 보고 시스템 등) 자료원에 대한 보완적 정보원이 되며, 개발된 Bi-LSTM 분류모델을 통해 약물 부작용 게시물 추출의 편리성을 제공할 것으로 기대된다.

### 1. 서론

전 세계적으로 고령화에 따른 약물 복용 빈도와 수량은 급격하게 증가하고 있는 추세이다. 이에 따른 약물 부작용 또한 급속하게 증가하여 환자의 안전을 위협하고 있는 실정이다. 이에 시판되고 있는 약물에 대한 새로운 부작용 또는 심각한 부작용이 있는지를 파악하여 조기에 발견하는 것이 중요한 이유로 부각되고 있다[1].

소셜 네트워크 서비스(Social Network Service, SNS)를 통해 게재 또는 등록되는 정보들은 개개인에 의한 직·간접적인 복용 후기를 포함하고 있기 때문에, 약물에 대한 이미 알려진 부작용 외에도 알려지지 않은 부작용이나 적응증 등과 관련한 정보를 포함하고 있어 유의미할 것이다[2].

SNS 등으로부터 약물 혹은 건강상태 부작용에 대한 텍스트를 수집하여 다양한 분석을 시도하고 있다. 이러한 연구는 약물과 증상 및 부작용에 대한 용어 정의 및 표준 용어로서의 매핑 작업이 큰 부분을 차지하여 분석 성능을 좌우하는 이슈가 있으며, 이를 보완하기 위해서 최근에 딥러닝을 활용하여 약물 이상 반응을 탐지하고자 하는 연구가 이루어지고 있다[3].

본 연구에서는 SNS 데이터를 수집하여 약물 부작용의 패턴 분석을 기반으로 양질의 약물 부작용 게시글을 확보하기 위한 연구 모델을 개발하고자 한다.

### 2. 연구 방법

본 연구에서는 SNS 데이터에서 순환 신경망 기반 분류 모델을 활용하여 약물 부작용 게시글을 분류하는 과정을 담은 파이프라인을 제안한다.

이를 요약하면, (1)약물감시 대상에 적합한 소셜 채널을 선택한다. (2)분석 대상 약물에 대한 사전에 정의하여 데이터를 확보한다. (3)획득한 데이터에 대한 표준 의약품 데이터베이스에서 대상 의약품과 관련된 부작용에 대한 렉시콘(Lexicon)을 추출 및 정의하고, 텍스트 전처리를 위한 불용어 렉시콘을 확보 및 보완한다. (4)전처리 및 불용어 필터링을 거친 텍스트 데이터를 활용하여 연관성 분석, 임베딩 분석을 활용한 약물 부작용의 패턴을 분석한다. (5)약물 부작용 패턴 분석 결과 발견된 부작용 단어를 담은 게시글들을 필터링한다. (6)수집된 텍스트 데이터를 2명 이상의 연구자가 수동 라벨링을 실시하고 신뢰도를 검증한다. (7)라벨링 데이터셋을 활용하여 순환

신경망 기반 분류 모델을 개발한다. (8)분류 모델 결과를 검증하여 모델을 고도화하고 이를 활용한 게시글 추출 과정을 지속적으로 수행한다. 이를 도식화하면 그림 1과 같다.

**2.1 렉시콘(Lexicon) 정의**

**2.1.1 약물명 렉시콘:** 한국약품정보원(KPIS)에서 제공하는 정보를 기반으로 탐색한다. 타겟 약물 성분을 검색 후 검색되는 약물 이름들을 취합하여 약물 이름 렉시콘을 생성한다.

**2.1.2 약물 부작용 렉시콘:** 표준화되어 공개된 약물 부작용 관련 데이터베이스 WHO-ART[4], SIDER[5]를 기반으로 생성된 약물 부작용 리스트 및 사전 생성된 소비자 용어 사전을 이용하여 생성될 수 있다. 여기서 WHO-ART 는 의약품 부작용 용어에 관한 국제 분류 체계를 나타내는 것으로 ADR (Adverse Drug Reaction) 보고에 가장 널리 사용되고 있다. 이에 국내에서는 한국어 버전인 WHO-ART ver.092 를 한국의약품안전관리원에서 배포하고 있다. SIDER 는 Drug-ADR 관계를 제공하는 약물 부작용 데이터베이스로써 시판되는 의약품 및 그 의약품들에 대한 약물 부작용 정보를 포함한다. SIDER 에서 케토프로펜을 검색 후 나열된 MedDRA Preferred Term 과 WHO-ART ver.092 한국어 버전과의 매핑을 통해 부작용 사전을 구축한다.

**2.1.3 불용어 렉시콘:** 텍스트 기반 데이터의 불용어는 자연어 데이터 처리하기 전이나 후에 필터링 된 단어를 말한다. 상업적 고유 명사로 반복되는 용어, 약물 부작용과 상관없는 분야의 용어, 인터넷 상에서 통용되는 용어 등을 추가하여 불용어 렉시콘을 생성한다.

**2.2 데이터 수집**

네이버 오픈 API 를 사용하여 데이터를 확보하는 경우 게시글 확보 수에 제한이 있다. 이를 극복하기 위해 카페와 블로그 검색 플랫폼에서 검색어를 입력하여 얻은 결과에 대한 페이지 정보를 활용하였다.

페이지에서 게시물의 URL 로부터 게시물 본문만 수집하고, 아이디, 카페명 등 민감한 정보는 제외한다.

**2.3 데이터 전처리**

SNS 게시물의 경우 빠른 소통과 편의를 위해 단어 사이에 띄어쓰기를 하지 않거나 형태소가 파괴된 경우들을 흔히 찾아볼 수 있다. 그러나 이러한 오류들은 분석 결과에 좋지 않은 영향을 미칠 수 있기에 전처리가 필요하다. 띄어쓰기 전처리는 SoySpacing 모듈을 활용하였고, 파생되는 오류를 줄이기 위해 추가 간격 규칙을 적용하여 해당 모듈에 업데이트 하였다. 형태소 분석기는 Mecab 모듈을 이용하여 단어 토큰화를 수행하였다. 토큰화를 통해 추출된 명사에는 연구에 무의한 명사가 많이 포함되어 있을 수 있으며 이를 선별하여 불용어 렉시콘에 업데이트하였다.

**2.4 패턴 분석**

**2.4.1 연관성 분석**

연관성 분석은 데이터 내부에 존재하는 데이터 간의 상호관계 혹은 종속관계를 찾아내는 분석으로써, 단순하지만 명확한 결과 해석이 가능하도록 한다[6].

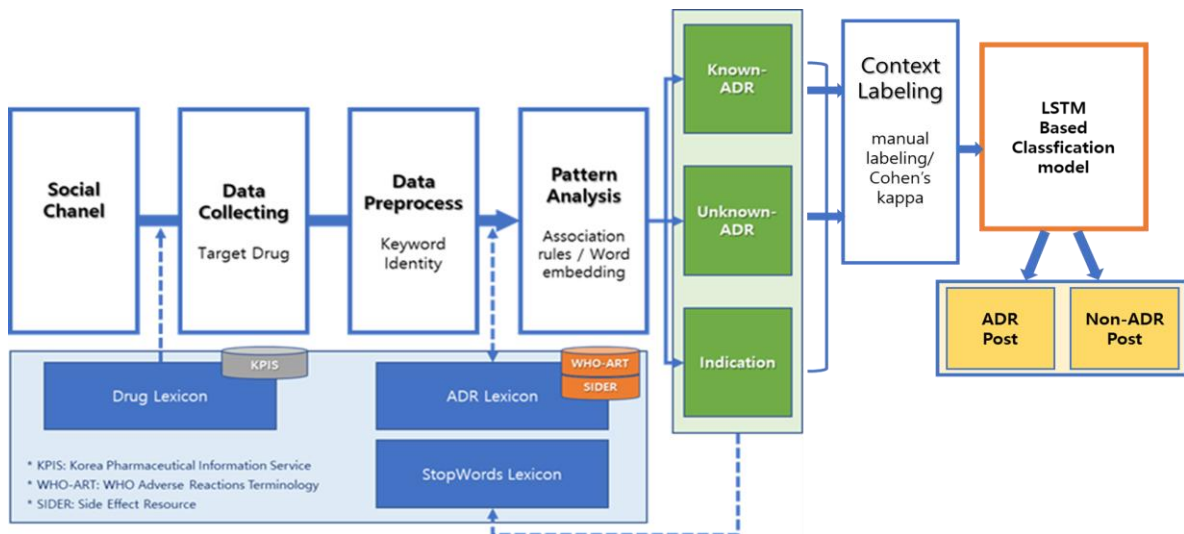
**2.5 부작용 게시글 필터링 및 라벨링**

**2.5.1 부작용 게시글 필터링**

부작용 게시글 분류 모델을 개발하기 위해서는 텍스트 데이터가 필요하다. 이를 위해 2.2 데이터 수집에서 수집된 데이터에서 전처리를 거친 후 패턴 분석에서 유의미하게 나온 약물 용어와 약물 부작용 쌍을 정리해 이를 포함하고 있는 게시글을 필터링한다.

**2.5.2 부작용 게시글 라벨링**

필터링된 게시글을 최소 2 명 이상의 수동 라벨링 작업자가 투입되어 수동 라벨링을 진행한다. 약물 부작용 게시글의 경우 1, 그렇지 않은 게시글을 0 으로 분류하며, 작업자 내 신뢰도를 Kappa 분석(Cohen's or Fleiss' Kappa Coefficient)을 사용하여 평가하여 진행한다.

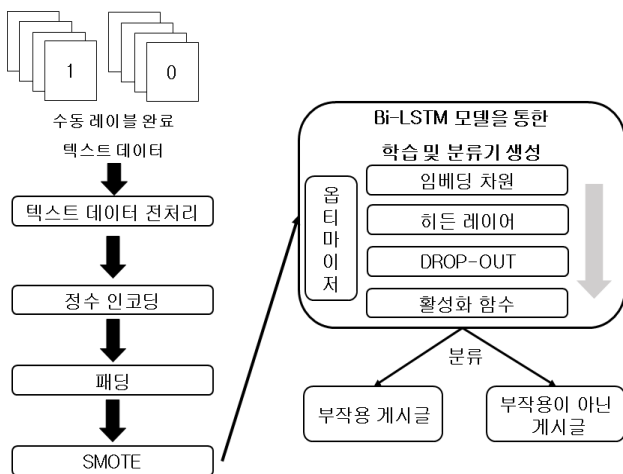


(그림 1) 연구 개요도

### 2.6 순환 신경망 기반 분류 모델

SNS 게시물에서 딥러닝 기법을 적용하여 약물 부작용을 탐지하기 위한 기존 연구[7]를 참고하여 분류 모델 생성 과정을 설정하였다. 사건의 순서를 가지고 있는 시계열 데이터와 같은 맥락으로 텍스트 데이터는 문장 내 단어마다 순서를 가지고 구성 되어있다. 이를 위해 순환신경망(Recurrent Neural Network, RNN) 모델을 활용하는데, 장기 의존성 문제(The problem of Long-Term Dependencies)를 해결하고자 LSTM(Long Short-Term Memory) 모델을 활용하여 약물 부작용 게시글 여부를 분류하는 모델을 생성한다[8]. LSTM의 Hidden State(Forward 방향)에 추가적인 Hidden State(Backward 방향)를 추가하여 양방향으로 입력 단어의 순서를 고려하는 모델이 Bi-LSTM 이다. 개발은 Python 3.8 버전과 Tensorflow 를 사용하였고, PC 환경은 Intel Core i7-8700 3.2Ghz CPU, 16GB RAM 으로 구성되었다.

Bi-LSTM 모델의 기계학습 절차를 요약하면 다음과 같다. (1)라벨링된 텍스트 데이터에 대해 전처리 작업을 수행하는데, 이때 모든 형태소 분석을 통해 각 포스트 마다 단어를 토큰화 한다. (2)해당 텍스트 데이터에 대한 빈도수 기반의 정수 인코딩과 병렬 처리를 위한 패딩 단계를 거친다. (3)입력 텍스트 데이터들은 임베딩 층에서 임베딩 길이에 맞게 임베딩 된다. (4)히든 유닛 개수를 128 개로 설정한다. (5)분류 문제를 해결하는 순환 신경망 모델은 다대일로 접근하며, 이를 위해 모델의 출력은 이진 분류로 구성하고 활성화함수는 sigmoid, 옵티마이저는 Adam 로 설정하였고, 추가적인 파라미터는 learning\_rate:0.001, batch\_size:32, epoch:50 으로 구성하였다. 이를 도식화하면 그림 2 과 같다.



(그림 2) 연구 분류 모델 구성도

부작용 게시글에 대한 불균형 문제 해결을 위하여 SMOTE(Synthetic Minority Over-sampling Technique)[9] 기법을 적용하여 부작용 게시글의 개수를 증가시켜 2:1의 비율로 tuning 하였다. 학습 데이터와 테스트 데이터는 7:3의 비율로 나누었고, 이진분류 예측에서 출력값 분류 기준(threshold)은 디폴트 0.5 이다.

### 3. 연구 결과

#### 3.1. 약물 선정

본 연구에서는 한국의 65 세 이상 노인을 대상으로 Beers Criteria 약물 처방의 발생빈도와 부작용을 계산한 연구[10]를 참조하여 건양대학교의료원의 실제 임상 환경에서 처방과 부작용을 근거로 노인 다빈도 처방 3 대 약물(메토클로프라미드, 클로르페니라민, 케토프로펜) 중에서 가장 많은 소셜 데이터를 확보할 수 있었던 케토프로펜을 선정하였다.

#### 3.2 확보된 데이터

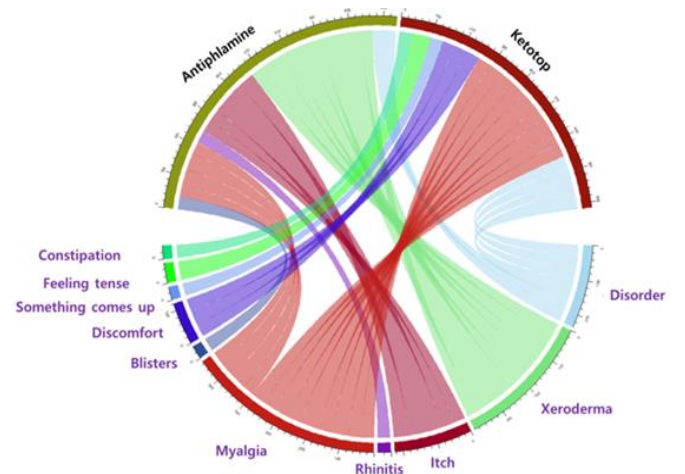
2005년부터 2020년까지 크롤링을 통해 총 11,693 개의 게시물을 수집했다. 2.4에서 설명한 전처리 과정을 거치고 패턴 분석에 사용된 케토프로펜 게시글은 최종 3,828 개의 게시글이다.

#### 3.3 생성된 렉시콘

케토프로펜에 대하여 3 중(약물명, 부작용, 불용어) 렉시콘들을 생성하였다. 약물명 사전에는 케토톱, 안티푸라민 등 9 개의 약물명이 포함되고, 부작용 사전에는 '발진', '부종', '배뎡겨' 등 2,925 개의 부작용 용어가 포함된다. 마지막으로 불용어 사전은 '반대로', '여러분', '더욱이' 등 7,196 개의 단어로 구성하였다.

#### 3.4 약물 부작용 패턴 분석 결과

패턴 분석은 R 4.1.0 버전으로 작업하였다. 우리는 연관성 분석(지지도:0.01, 신뢰도:0.6)을 바탕으로 케토프로펜과 그 부작용 용어 간 상관관계를 살펴보았으며, 그 결과를 Chord Diagram 으로 나타냈다. 케토프로펜을 함유한 대표적인 약물인 케토톱의 경우 근육통과 가장 관련성을 보였으며, 안티푸라민의 경우 건조증과 가장 높은 연관성을 보였다. 비록 낮은 지지도임에도 본 연구에서는 상대적으로 빈도가 낮은 데이터 또한 유의미성을 가지기에 이를 함께 분석하였다.



(그림 3) 케토프로펜(약물명-약물 부작용) Chord Diagram

#### 3.5 모델 분류 결과

학습 데이터 생성을 텍스트 데이터를 2 명의 연구자가 수동 라벨링 작업을 하였다. 해당 라벨링된 텍

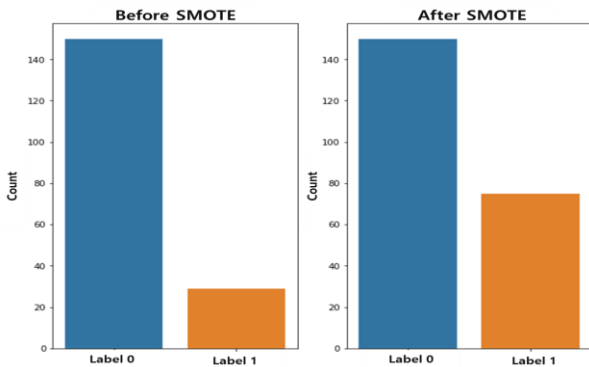
스트는 패턴분석 후 유의미한 약물 이름과 약물 부작용 단어 쌍이 포함된 총 256 개의 게시글로 구성되었고, 작업 결과 Kappa Score 0.8 의 합의를 나타냈다.

<표 1> 수동 라벨링 결과 비교표

항 목	케토프로펜		총 합
	안티푸라민	케토톱	
데이터개수	197	59	256
라벨링	0*	173 (87.8%)	214 (86.6%)
	1*	24 (12.2%)	42 (16.4%)

\*0: 부작용 게시글 아님, 1: 부작용 게시글임

학습 데이터와 테스트 데이터를 7:3 의 비율로 나누는 후 데이터 불균형 문제를 해결하고자 학습데이터 179 개(라벨 0:150, 라벨 1:29)에 대해 SMOTE 기법을 적용하여 225 개(라벨 0:150, 라벨 1:75)로 증강하였다.



(그림 4) SMOTE 적용 전후 데이터셋 변화

모델의 테스트 정확도는 87%를 기록했다. 모델 테스트의 결과는 <표 2>와 같다.

<표 2> 모델 결과 혼동행렬

		예측		총 합
		Label 0*	Label 1*	
정답	Label 0	64	1	65
	Label 1	9	3	12

\*0: 부작용 게시글 아님, 1: 부작용 게시글임

#### 4. 결론

본 연구에서 제안하는 모델은 Bi-LSTM 기반 분류 모델을 활용하여 SNS 데이터에서 약물 부작용 게시글을 추출하고 그 유용성을 평가하였다. 다빈도 처방약물인 케토프로펜에 대해 네이버 게시글을 분석한 결과, 알려진 약물 부작용 정보를 잘 탐색할 수 있었으며 이는 SNS 데이터를 기반으로 소비자들의 약물 부작용 정보 획득에 좋은 정보원으로 활용될 수 있음을 시사한다.

본 연구에서는 몇 가지 한계점이 존재한다. 첫째, 네이버 게시글에 한정하여 약물 특화 커뮤니티에서의 정보 특성을 파악하지 못하였다. 둘째, 분류 모델 개발 단계에서 초기 실험 단계에 머물렀다. 셋째, 약물의 알려진 부작용만을 정의하였다. 넷째, 수동 라벨링

단계에서 약물 이름과 약물 부작용 단어 간의 단순 매칭을 진행하였다.

향후 연구로는 위음성(FalseNegative) 부분을 개선하고자 모델 구성 및 데이터 불균형 해소의 다양한 기법을 적용해 고도화하고자 한다. 또한, 약물의 알려지지 않은 부작용 렉시콘으로 연구 확장하고자 한다. 무엇보다 임상 전문가와의 협업을 통해 임상현장의 도메인 지식을 반영하여 임상실증 연구에 기여할 수 있기를 기대한다.

#### 감사의 글

이 논문은 정부(보건복지부)의 재원으로 보건산업진흥원(KHIDI)의 Korea Health Technology R&D 프로젝트의 지원을 받아 수행된 연구임 (grant number: HI19C1310).

#### 참고문헌

- [1] Pearson, Tonya F., et al. "Factors associated with preventable adverse drug reactions." American Journal of Health-System Pharmacy, vol. 51, no. 18, pp. 2268-2272, 1994.
- [2] Sultana, Janet, Paola Cutroneo, and Gianluca Trifirò. "Clinical and economic burden of adverse drug reactions." Journal of pharmacology & pharmacotherapeutics, vol. 4, no. Suppl1, pp. S73, 2013.
- [3] Cocos, Anne, Alexander G. Fiks, and Aaron J. Masino. "Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts." Journal of the American Medical Informatics Association, vol. 24, no. 4, pp. 813-821, 2017.
- [4] Sills, Judith M. "World health organization adverse reaction terminology dictionary." Drug Information Journal, vol. 23, no. 2, pp. 221-216, 1989.
- [5] Kuhn, Michael, et al. "A side effect resource to capture phenotypic effects of drugs." Molecular systems biology, vol. 6, no. 1, pp. 343, 2010.
- [6] Tan, Pang-Ning, Vipin Kumar, and Jaideep Srivastava. "Selecting the right objective measure for association analysis." Information Systems, vol. 29, no. 4, pp. 293-313, 2004.
- [7] Cocos, Anne, Alexander G. Fiks, and Aaron J. Masino. "Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts." Journal of the American Medical Informatics Association, vol. 24, no. 4, pp. 813-821, 2017.
- [8] Noh, Young-Dan, and Kyu-Cheol Cho. "A Text Content Classification Using LSTM For Objective Category Classification." Journal of the Korea Society of Computer and Information, vol. 26, no. 5, pp. 39-46, 2021.
- [9] Rupapara, Vaibhav, et al. "Impact of SMOTE on imbalanced text features for toxic comments classification using RVVC model." IEEE Access, vol. 9, pp. 78621-78634, 2021.
- [10] Kim, Grace Juyun, Kye Hwa Lee, and Ju Han Kim. "South Korean geriatrics on Beers Criteria medications at risk of adverse drug events." PLoS One, vol. 13, no. 3, pp. e0191376, 2018.