Bi-LSTM based detection of ADR posts using Korean SNS data

Chung-Chun Lee, Seunghee Lee, Mi-Hwa Song, Jong-Yeup Kim, Suehyun Lee

Submitted to: JMIR Medical Informatics
on: December 23, 2022

Table of Contents

Bi-LSTM based detection of ADR posts using Korean SNS data

Chung-Chun Lee[1*]; Seunghee Lee[2*]; Mi-Hwa Song[3]; Jong-Yeup Kim[1]; Suehyun Lee[1*]

[1]Department of Biomedical Informatics, College of Medicine Konyang University Daejeon KR
[2]Healthcare Data Science Center Konyang University Hospital Daejeon KR
[3]School of Smart IT Semyung University Jecheon KR
[*]these authors contributed equally

Corresponding Author:
Suehyun Lee
Department of Biomedical Informatics, College of Medicine
Konyang University
158 Gwanjeodong?ro, Seo?gu, Daejeon 35365, Republic of Korea.
Daejeon
KR

Abstract

Background: Social networking services (SNS) closely reflect the lives of individuals in modern society and generate large amounts of data. Previous studies have extracted drug information using relevant SNS data. In particular, it is important to detect adverse drug reactions (ADRs) early using drug surveillance systems. To this end, various deep learning methods have been employed, which have studied multiple languages in addition to English.

Objective: A cautionary drug that can cause ADRs in elderly patients was selected, and Korean SNS data containing this drug information were collected. Based on this information, we aimed to develop a deep learning model that classifies drug ADR posts based on a Recurrent Neural Network (RNN).

Methods: In previous studies, ketoprofen, which has a high prescription frequency and thus, was selected as the target drug. Posts were collected from portal site containing information about the drug, and NLP techniques for data written in Korean were applied. Posts containing relevant drug names and ADR word pairs were filtered through association analysis, and training data were generated through manual labeling tasks. So, a Bi-LSTM classification model was generated. The entire process was further verified using aceclofenac.

Results: Among the NSAIDs, Korean SNS posts containing Ketoprofen and Aceclofenac information were secured, and the generic name lexicon, ADR lexicon, and Korean stop word lexicon were generated. In addition, to improve the accuracy of the classification model, an embedding layer was created considering the association between the drug name and the ADR word. In the ADR post-classification test, ketoprofen and aceclofenac achieved 87% and 80% accuracy, respectively.

Conclusions: Here, we propose a process for developing a model for classifying ADR posts using SNS data. After analyzing drug name-ADR patterns, we filtered high-quality data by extracted posts, including known ADR words based on the analysis. Based on these data, we developed a model that classifies ADR posts. This confirmed that a model that can leverage social data to monitor ADRs automatically is feasible.

(JMIR Preprints 23/12/2022:45289)
DOI: https://doi.org/10.2196/preprints.45289

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Bi-LSTM based detection of ADR posts using Korean SNS data

Chung-Chun Lee[1,†], Seunghee Lee PhD[2,†], Mi-Hwa Song PhD[3], Jong-Yeup Kim MD PhD[1, 2], Suehyun Lee PhD[1, 2*]

[1]Department of Biomedical Informatics, College of Medicine, Konyang University, Daejeon, 35365, Republic of Korea
[2]Healthcare Data Science Center, Konyang University Hospital, Daejeon, 35365, Republic of Korea
[3]School of Smart IT, Semyung University, Jecheon, 27136, Republic of Korea

†These authors contributed equally.
*Corresponding author
Department of Biomedical Informatics
College of Medicine, Konyang University
158 Gwanjeodong-ro, Seo-gu, Daejeon 35365, Republic of Korea.
Phone: 82-42-600-8679; Fax: 82-42-543-8959; E-mail: shleemedi@kyuh.ac.kr

ORCID: Chung-Chun Lee (0000-0002-1670-6222), Seunghee Lee (0000-0002-2273-0915), Mi-Hwa Song (0000-0001-7047-8032), Jong-Yeup Kim (0000-0003-1230-9307), Suehyun Lee (0000-0003-0651-6481)

Abstract

Background: Social networking services (SNS) closely reflect the lives of individuals in modern society and generate large amounts of data. Previous studies have extracted drug information using relevant SNS data. In particular, it is important to detect adverse drug reactions (ADRs) early using drug surveillance systems. To this end, various deep learning methods have been employed, which have studied multiple languages in addition to English.

Objective: A cautionary drug that can cause ADRs in elderly patients was selected, and Korean SNS data containing this drug information were collected. Based on this information, we aimed to develop a deep learning model that classifies drug ADR posts based on a Recurrent Neural Network (RNN).

Methods: In previous studies, ketoprofen, which has a high prescription frequency and thus, was referred to the most in posts secured from SNS data, was selected as the target drug. Blogs and café, NAVER Q&A posts between 2005-2020 were collected from NAVER, a portal site containing information about the drug, and Natural Language Processing (NLP) techniques for data written in Korean were applied. Posts containing highly relevant drug names and ADR word pairs were filtered through association analysis, and training data were generated through manual labeling tasks. Using the training data, an embedded layer of word2vec was formed, and a Bi-LSTM classification model was generated. Then, we evaluated the AUC with other ML models. In addition, the entire process was further verified using the Non-steroidal Anti-inflammatory Drug (NSAID) aceclofenac.

Results: Among the NSAIDs, Korean SNS posts containing Ketoprofen and Aceclofenac information were secured, and the generic name lexicon, ADR lexicon, and Korean stop word lexicon were generated. In addition, to improve the accuracy of the classification model, an embedding layer was created considering the association between the drug name and the ADR word. In the ADR post-classification test, ketoprofen and aceclofenac achieved 85% and 80% accuracy, respectively.

Conclusions: Here, we propose a process for developing a model for classifying ADR posts using SNS data. After analyzing drug name-ADR patterns, we filtered high-quality data by extracted posts, including known ADR words based on the analysis. Based on these data, we developed a model that classifies ADR posts. This confirmed that a model that can leverage social data to monitor ADRs automatically is feasible.

Keywords: adverse drug reaction; social network service; classification model; Korean text data

Introduction

The frequency and quantity of drugs taken worldwide due to aging are rapidly increasing. As a result, drug ADRs have also increased rapidly, threatening the safety of patients. Accordingly, the identification and early detection of new or serious ADR in drugs on the market is an increasingly important issue [1].

Data from social network services (SNS) is significant because it contains information related to, or indicating, known ADRs, in addition to unknown ADRs [2]. In Europe, the Innovative Medicines Initiative WEB-RADR (WEB-Recognizing Adverse Drug Reactions) project was carried out during 2014-2017 to address questions about the potential use of social media to monitor ADRs. Whether social media is as valuable as other data sources, such as voluntary reporting, in ADR monitoring has been studied [3].

Various analyses have been attempted by collecting text referring to ADRs, drugs, or health conditions from SNS data. These studies have played a large part in defining and mapping terminology to standard terms for drugs, symptoms, and ADRs, and as a result of this, studies have recently been conducted to detect ADRs using deep learning [4]. There has also been a study on classifying ADR posts through architectures such as HAN, FastText, and CNN based on word2vec in a dataset consisting of 11,623 posts related to drugs on ASK and Twitter [5]. To improve the performance of the model for classifying ADR posts, there are concerns regarding data balance and feature selection, which must be overcome. To address this problem, They adopted Bidirectional Encoder Presentations from Transformers (BERT) to introduce an architecture that selects keywords, such as those entered as layers before the classification model, and were recognized by humans [6].

In addition, studies have focused on mapping consumer and medical terms of posts collected from social channels [7]. Creating an ADR post-classification model based on medical terminology from databases such as Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) or MedDRA (The Medical Dictionary for Regulatory Activities) creates a clinical foundation for future applications and can be used for various drugs and ADRs.

Using Chinese social media posts as the original dataset, we proposed a semi-supervised learning framework for detecting Chinese drug terms and ADR terms [8]. RuDReC (Russian Drug Reaction Corpus) is a Russian Corpus labeled with drug terms, disease information, health terms, drug efficacy, drug adaptation, drug ADRs, and more. In addition, an experiment to automate object recognition (NER) in this corpus achieved an F1 score of approximately 74.85% [9].

Herein, we introduce a process consisting of collecting Korean posts containing drug information on social channels, selecting posts containing essential information among the collected posts, and creating and evaluating a model for classifying ADR posts. Therefore, the final purpose of this study was to create a model that effectively classifies ADR posts, written in Korean, for drugs with a high frequency of use among senile individuals.
This requires natural language preprocessing in Korean, and is verified, not only for target drugs but also for additional drugs in the process of generating a classification model for specific ADR posts.

Methods

In this study, we propose a pipeline for classifying drug ADR posts using a circular neural

network-based classification model based on SNS data.
In summary, (1) a social channel suitable for drug monitoring was selected. (2) The generic name of the target drug, ADR lexicon, was generated. (3) Text data were obtained from social channels based on the generic name lexicon for the drug for analysis. Morphological analysis was conducted using the obtained text data, and unnecessary Korean terms were added to the stop word lexicon, simultaneously. (4) The pattern of drug ADRs was analyzed by association analysis and embedding analysis of text data that had undergone preprocessing and term filtering. (5) Posts containing ADR words identified by analysis of drug ADR patterns were filtered. (6) Two researchers manually labeled the collected text data and verified their reliability. (7) We created an embedding layer that considered the correlation between generic names and ADR words. (8) We developed a recurrent neural network-based classification model using a labeled dataset. (9) To verify the generated classification model, additional drugs were selected, and the entire step was repeated. Subsequently, the model was evaluated using a matrix that evaluated the target drug and additional drugs **(Fig. 1).**

Data Set

Drug selection: A previous study [10] calculated the Beers Criteria drug prescription and the incidence of side effects in persons aged ≥65 years in Korea. The top three drugs (metoclopramide, chlorpheniramine, and ketoprofen) were selected based on prescriptions and side effects reported in the clinical environment of Konyang University Hospital [11]. Therefore, the focus of this study was on ketoprofen.

Lexicon generation

Generic name lexicon: We searched for ketoprofen based on the information provided by the Korea Pharmaceutical Information Service (KPIS). After searching for the target drug, the drug name lexicon was generated by collecting the searched drug brand names.

ADR lexicon: The ADR lexicon was generated using a list of drug ADRs based on a standardized and published database of drug ADR, WHO-ART [12], SIDER [13], and a lexicon of pre-generated consumer terms. WHO-ART represents an international classification system for drug ADR terms and is the most widely used system in ADR reporting.

Accordingly, the Korean version of WHO-ART ver.092 has been produced by the Korea Institute of Drug Safety and Risk Management (KIDSRM) [14]. SIDER is a drug ADR database that includes information on drug-ADR relationships and drug ADRs to commercially available drugs. After searching for “ketoprofen” in SIDER, an ADR lexicon was established by mapping definitions from the listed MedDRA Preferred Term and WHO-ART ver.092 Korean version.

Text preprocessing

Data Collection: When securing data using the Naver Open API, there is a limit to the number of posts that can be secured. To overcome this, page information on the results obtained by entering search terms on café and blog, Naver Q&A. search platforms was used. Only the body of the post was collected from the URL of the post on the page, and sensitive information, such as ID and café names, was excluded.

Morphological analysis: In the case of SNS posts, it is common to find cases in which spaces between words are not included, or morphemes are excluded for quick communication and convenience. These errors can adversely affect the analysis results; therefore, preprocessing is

necessary. The SoySpacing module was used for the spacing preprocessing, and additional spacing rules were applied and updated to the module to reduce the derived errors. The morpheme analyzer performed word tokenization using the Korean Intelligent Word Identifier (KIWI) module [15]. Words extracted through tokenization may contain many meaningless nouns that do not meet the purpose of the study; these words were selected and updated in the stop-word lexicon. The detailed process of text preprocessing using the collected data is shown in Figure 2 **(Fig. 2).**

Pattern Analysis

Association Analysis: Association analysis finds interrelationships or dependencies within data, enabling a simple and clear interpretation of results [16]. The SOC ranking of 960 patients (2013-2017) who reported spontaneous ADRs with ketoprofen from the Korea Adverse Event Reporting System (KAERS) database confirmed that the adverse reactions found in the SNS-based pattern analysis had similar patterns to the upper SOC group.

Filtering ADR Posts: Labeled text datasets are required to develop an ADR posting classification model. After preprocessing the collected data, we filtered the posts containing the generic drug words and drug ADR pairs that were significant in the pattern analysis.

Training Data Set

Labeling ADR posts: Two researchers majoring in information medicine were recruited to manually label the filtered posts. In the case of drug ADR posts, one non-drug ADR post was classified as 0, and the reliability of workers was evaluated using Kappa analysis (Cohen's or Pleiss' Kappa Coordinator).

Embedding layer created from filtered posts: We created an embedding model using an embedding technique (word2vec) for posts filtered through association analysis. The embedding model includes the association between the ADR terms ‘Ketotop’ and ‘Antipuramin’. The embedding model was applied as the input layer of the Bi-LSTM model.

Development Classification Model

Recurrent Neural Network-Based Classification Model: By applying deep learning techniques to SNS posts, the classification model creation process was established by referring to existing studies [4] to detect drug ADR posts. In the same context as time-series data with an order of events, text data is composed of a sequence for each word in a sentence. To this end, we utilized the Recurrent Neural Network (RNN) model, which utilizes the Long Short-Term Memory (LSTM) model to create a model that classifies whether drug ADR postings are recent [17]. Bi-LSTM considers the sequence of input words in both directions by adding an additional Hidden State (backward direction) to the LSTM's Hidden State (forward direction). The system used Python version 3.8 and Tensorflow, and the PC environment consisted of an Intel Core i7-8700 3.2 Ghz CPU and 16GB RAM.

The machine-learning procedure of the Bi-LSTM model is summarized as follows. (1) Pre-processing is performed on the labeled text data, where words are tokenized for each post through morpheme analysis. (2) Frequency-based integer encoding of the corresponding text data and padding steps for parallel processing were performed. (3) The input text data were embedded to fit the embedding length in the embedding layer. (4) The number of hidden units was set to 256. (5) The recurrent neural network model solving the classification problem

utilizes multi-to-one, for which the output of the model was constructed with binary classification, the activation function was set to sigmoid, the optimizer was set to Adam, and the additional parameters were learning rate = 0.001, batch_szie = 64, and epoch = 50 **(Fig. 3).**

The learning data and test data were divided by a ratio of 7:3, and the output value classification criterion (threshold) in the binary classification prediction was a default of 0.5.

Extra Drug Validation: The generation of the above classification model is the last step of the process, and the entire step is verified an extra drug. Based on the Korea Ministry of Food and Drug Safety's press release, " What medicines should senior citizens be careful of?" [18], the extra drug selected was an NSAIDs, as is Ketoprofen. Among the NSAID drugs reported in the press release, a similar amount of data to that of ketoprofen was secured for Geworin (generic name). Therefore, Geworin was determined as the extra drug.

Comparison for the other classification Method: As a validation method for Bi-LSTM machine learning techniques, in our study, classification models were generated and executed on the same dataset using GRU, CNN-Bi-LSTM, and KOGPT2.

Results

Data Acquired

From 2005 to 2020, a total of 11,693 posts were collected through crawling collectors. The ketoprofen posts that underwent the pretreatment process described in 2.3 were used for pattern analysis, resulting in the final 3,828 posts.

Table 1. Example of brand name and ADR of ketoprofen posts

Class	Context
Drug	Last year and the year before last, I used **Antiphlamine_6_drug** for insect bites and muscle pain Strangely, I don't think I can feel as strong a battle as before
ADR	When I got home, I got **blisters_153_ADR** and it was crazy. also **itch_156_ADR** I'm going to go to the hospital

Generated Lexicon

For ketoprofen, three types of lexicons (Generic name, ADR, and stop word) were generated. The drug name lexicon includes nine drug names such as Ketotop and Antiphlamine, and the ADR lexicon includes 2,925 ADR terms such as ‘rash’, ‘swelling’ and ‘stomach’. Finally, the stop word lexicon consists of 7,196 words such as 'opposite', 'everyone', and 'more'.

ADR Pattern Analysis Results

Pattern analysis was performed using R version 4.1.0. We examined the correlation between ketoprofen and its ADR terms based on the association analysis and found the highest correlation (support:0.01, confidence:0.6) was with ‘muscle pain’ and ‘disability’, whilst correlation analysis of Antiphlamine showed the highest correlation with ‘dryness’ and ‘allergy’. Although these support values are low, this study analyzed data with a relatively low frequency, which also has an impact.

Embedding model of filtered post

When creating an embedding model for filtered posts, the parameters used in word2vec were skip-gram and 300 dimensions. Through this, it was found that there was a high degree of cosine similarity between 'ketotop' and 'itching', 'heat rash', and 'rash'. In addition, antipuramin had a high degree of cosine similarity with 'itching', 'burning', and 'swelling'.

Model classification results

Two researchers majoring in information medicine manually labeled the text data for training data generation. The labeled text consisted of 403 posts containing a significant drug name and drug ADR word pair after pattern analysis, indicating reliability with a kappa Score of 0.8.

Table 2. Manually labeled text table

Category		Ketoprofen		Total
		Antipuramin	Ketotop	
Number of data		403	173	576
Labeling	Non-ADR post	229 (56.8%)	103 (59.5%)	332 (57.6%)
	ADR post	174 (43.2%)	70 (40.5%)	244 (42.4%)

The test accuracy of the model was 85%. Table 3 presents the results of the model tests.

Table 3. Confusion matrix

		Prediction		Total
		Non-ADR post	ADR post	
Actual	Non-ADR post	93	8	101
	ADR post	18	54	72

The results showing the ROC curve of the model for each binary classification threshold point are as follows.
In terms of accuracy, the highest accuracy was approximately 85% when the threshold point was 0.5, and the lowest accuracy was approximately 58% when the threshold point was 1.0. The threshold point in the table above was 0.5, and the accuracy was 85%. In the case of recall, 54 out of 72 ADR posts were matched to ADR posts to achieve an accurate prediction of 75%.

Extra Drug Validation

For the collected data of the additional drug, Geworin, 66 (27.96%) posts referred to ADRs and 170 (72%) posts referred to non-ADRs out of the total of 236 posts. The results of the experiment with the same Bi-LSTM architecture after dividing the training and test data by 7:3. The accuracy of this test was 80%.

Table 4. Extra Drug Confusion matrix

		Prediction		Total
		Non-ADR post	ADR post	
Actual	Non-ADR post	42	5	47
	ADR post	9	15	24

The results showing the ROC curve of the model for each binary classification threshold point are as follows.
The largest accuracy was about 82% when the threshold point was 0.3, and the lowest accuracy was about 60.5% when the threshold point was 0.1. The threshold point in the table above was 0.5, and the accuracy was 80%. In the case of recall, 15 of the total 24 ADR posts were matched to make an accurate prediction of about 62% and decreased compared to the results of Ketoprofen-containing drugs (75% to 62%) **(Fig. 4).**

Comparison for the other classification Method

In classification of ADR posts, other classification model methods other than Bi-LSTM used in our study were applied and performance indicators were compared among deep learning. It was found that Bi-LSTM showed good performance overall. In particular, KOGPT2 was a transformer-based classification model and was a specialized method for Korean, but showed the lowest performance. The models's parameters were learning rate = 0.001, batch_szie = 64, and epoch = 50. Also, training data and test data were divided by a ratio of 7:3. It is same with Bi-LSTM. We experimented using Python's early stopping module. This led to the derivation of the optimal AUC value. The Comparison of roc cure is showed. **(Fig. 5)**.

Table 5. Comparison of Bi-LSTM with other classification models.

Antipuramin & Ketotop	Bi-LSTM	GRU	CNN-Bi-LSTM	KOGPT2
Precision	0.87	0.79	0.73	0.69
Recall	0.75	0.76	0.67	0.66
F1-score	0.81	0.77	0.70	0.68
AUC	0.85	0.82	0.76	0.72

Discussion

The model proposed in this study uses a Bi-LSTM-based classification model to extract drug ADR posts from SNS data and evaluate their significance. We investigated whether these processes can be effectively used for drug monitoring and proposed a classification model development process for extracting Korean ADR posts based on stepwise methods.

We comprehensively reviewed the advantages and disadvantages of using social media data for drug ADR detection [19] and investigated whether social media analysis can be integrated with voluntary reporting systems to improve ADR detection [20]. No previous studies have performed drug monitoring using KAERS, a Korean voluntary reporting system, or NAVER, a social channel.

Social media conversations also provide a wide range of information about the patient's health, such as drug use information. This may explain diversity in the data in terms of ADRs other than the known ADR of the drug [21].

When using NAVER posts, data were collected from two channels (Naver Blogs and Naver Café, Naver Q&A platform) to obtain the actual voices of consumers, not articles that contain general information. (1) For famous drug names, social network service data can be used to obtain sufficient ADR data, suggesting that they can be used in drug ADR studies or drug surveillance

studies. (2) We proposed a method to build a Korean ADR lexicon based on the mapping of the SIDER DB results with WHO-ART files (including Korean terms) for target drugs. (3) It was confirmed that the pattern of drug ADRs obtained using social network services was not different from the SOC range of self-reported ADR information obtained from KAERS. This finding adds credibility to our method. (4) Based on the results of pattern analysis, we present the process of filtering posts and generating them as artificial intelligence learning data, and propose a model that uses them to classify posts regarding ADRs of target drugs.

Despite older people having reduced use of online platforms, this may not limit the data as information shared with family members may appear on social networks. However, this study has several other limitations. First, data were only collected from Naver posts; therefore, information was limited to those users, and other platforms such as Instagram and Twitter were not considered. Second, only the known ADRs of the target drug were defined when constructing the lexicon. An extension of the unknown ADR lexicon is required to extract unknown ADRs to aid drug monitoring systems. Third, there is insufficient evidence to prove that filtered and labeled posts are clinically associated with drug ADRs. This requires a review by clinical experts during the labeling of posts. Fourth, the depth of the classification model development phase is only at the initial experimental level.

The researchers proposed a model to classify posts referring to ketoprofen-related ADRs on social media using well-known ADR information that can be further studied, including unknown ADRs. Furthermore, the model will be upgraded in the future by applying various techniques to resolve false negativity. The target drugs analyzed will also be expanded to high-demand drugs, such as tylenol and aspirin. The overall process, from the proposed collection to model development, can detect ADRs from a consumer's perspective from outside the hospital, which can be used to inform drug safety monitoring policies for a specific drug.

Conclusion

In this study, we proposed a process for developing a model for classifying drug ADR posts using social network service data. After analyzing drug ADR patterns, we filtered high-quality data by filtering posts including known drug ADR terms based on the results. Based on these data, we developed a model that classifies drug ADR posts. This confirmed that a model that can leverage social media data to automatically monitor drug ADRs is feasible.

Acknowledgments

This research was financially supported by a grant from the Korea Health Technology R&D Project of the Korea Health Industry Development Institute (KHIDI) and was funded by the Ministry of Health & Welfare, Republic of Korea (Grant No. HI19C1310).

Conflict of Interest

The authors declare no conflict of interest.

Abbreviations

ADR: Adverse Drug Reaction
API: Application Programming Interface
KAERS; Korea Adverse Event Reporting System
Kiwi: Korean Intelligent Word Identifier
KPIS: Korea Pharmaceutical Information Service
MedDRA: Medical Dictionary for Regulatory Activities

SIDER: Side Effect Resource
SNS: Social Network Service
SOC: System Organ Class
URL: Uniform Resource Locator
WHO-ART: World Health Organization Adverse Reaction Terminology
NLP: Natural Language Processing
NSAID: Non-steroidal Anti-inflammatory Drug
WEB-RADR: WEB-Recognizing Adverse Drug Reactions
BERT: Bidirectional Encoder Presentations from Transformers

References

1) Pearson TF, Pittman DG, et al. Factors associated with preventable adverse drug reactions. American Journal of Hospital Pharmacy;1994. (51):2268-2272

2) Sultana J, Cutroneo P, and Trifirò G. Clinical and economic burden of adverse drug reactions. Journal of pharmacology & pharmacotherapeutics; 2013. (4): S73

3) Caster O, Dietrich J, et al. Assessment of the utility of social media for broad-ranging statistical signal detection in pharmacovigilance: results from the WEB-RADR project. Drug safety; 2018. (41): 1355-1369.

4) Cocos A, Fiks AG, and Masino AJ, Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. Journal of the American Medical Informatics Association; 2017. (24):813-821

5) Rezaei Z, Ebrahimpour-Komleh H, et al. Adverse drug reaction detection in social media by deep learning methods. Cell Journal; 2020. (22): 319

6) Huang J-Y, Lee W-P, and Lee K-D, Predicting Adverse Drug Reactions from Social Media Posts: Data Balance, Feature Selection and Deep Learning. Healthcare; 2022. (10):618

7) Karimi S, Metke-Jimenez A, Kemp M, and Wang C, Cadec: A corpus of adverse drug event annotations. Journal of biomedical informatics; 2015. (55): 73-81.

8) Zhang M, Zhang M, et al. Automatic discovery of adverse reactions through Chinese social media. Data mining and knowledge discovery; 2019. (33): 848-870.

9) Tutubalina E, Alimova I, et al. The Russian Drug Reaction Corpus and neural models for drug reactions and effectiveness detection in user reviews. Bioinformatics; 2021. (37): 243-249.

10) Kim GJ, Lee KH, and Kim JH, South Korean geriatrics on Beers Criteria medications at risk of adverse drug events. PLoS One; 2018. (13): e0191376.

11) Shin H, Kim N, et al. Geriatrics on beers criteria medications at risk of adverse drug events using real-world data. International Journal of Medical Informatics; 2021. (154): 104542.

12) Sills JM, World health organization adverse reaction terminology dictionary. Drug Information Journal; 1989. (23):221-216

13) Kuhn M, Letunic I, Jensen LJ and Bork P, The SIDER database of drugs and side effects. Nucleic acids research; 2016. (44.D1): D1075-D1079.

14) Korea Institute of Drug Safety & Risk Management. WHO-ART 092 [Internet]. Anyang: Korea Institute of Drug Safety & Risk Management; 2015 Jul; Available from: https://www.drugsafe.or.kr/iwt/ds/ko/bbs/EgovBbs.do;jsessionid=JcVbUVsrD2anWzye6BgkPqwiw4cktqW0oiFbNuq2WQaDVhqlgoXgLidzeTEGPuTz.webint_2_servlet_engine1?bbsId=BBSMSTR_000000000024&nttId=1446&pageIndex=2&searchCnd=&searchWrd=.

15) Lee MC, Yang HM, and Kim GY, 2018. Kiwi: Korean Intelligent Word Identifier. https://github.com/bab2min/kiwi.

16) Tan P-N, Vipin K, and Jaideep S, Selecting the right objective measure for association analysis. Information Systems; 2004. (29.4): 293-313.

17) Noh Y-D and Cho K-C. A Text Content Classification Using LSTM For Objective Category Classification. Journal of the Korea Society of Computer and Information; 2021. (26): 39-46

18) What medicines should senior citizens be careful of? [Press release]. The Korea Ministry of Food and Drug Safety; 2021 Jul 18; Available from: https://www.mfds.go.kr/brd/m_99/view.do?seq=45568&srchFr=&srchTo=&srchWord=&srchTp=&itm_seq_1=0&itm_seq_2=0&multi_itm_seq=0&company_cd=&company_nm=&page=39.

19) van Stekelenborg J, Ellenius J, et al. Recommendations for the use of social media in pharmacovigilance: lessons from IMI WEB-RADR. Drug safety; 2019. (42.12): 1393-1407.

20) Li Y, Yepes AJ, and Xiao C. Combining social media and FDA Adverse Event Reporting System to detect adverse drug reactions. Drug safety; 2020. (43.9): 893-903.

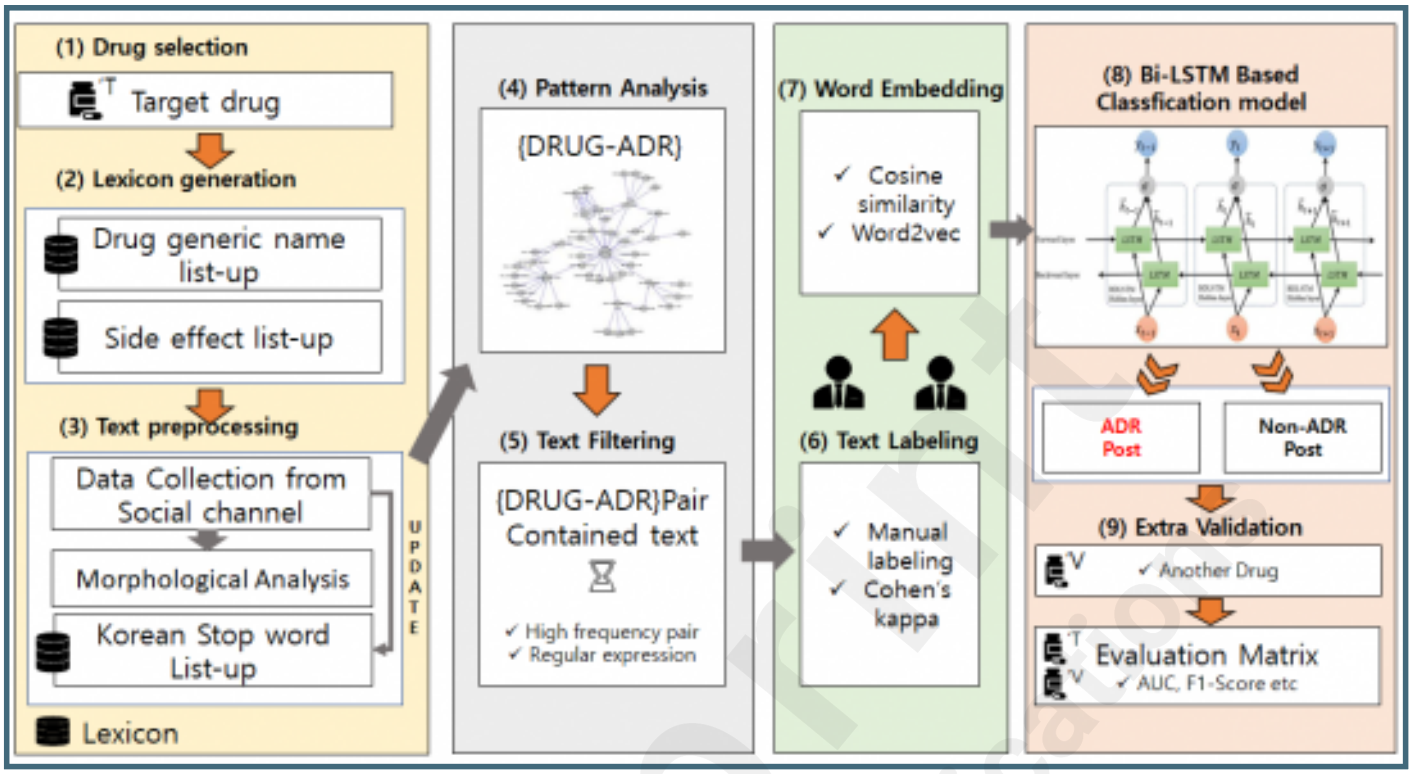
21) Kim SJ, Marsch LA, Hancock JT and Das AK, Scaling up research on drug abuse and addiction through social media big data. Journal of medical Internet research; 2017. (19.10): e6426.

Figure legends

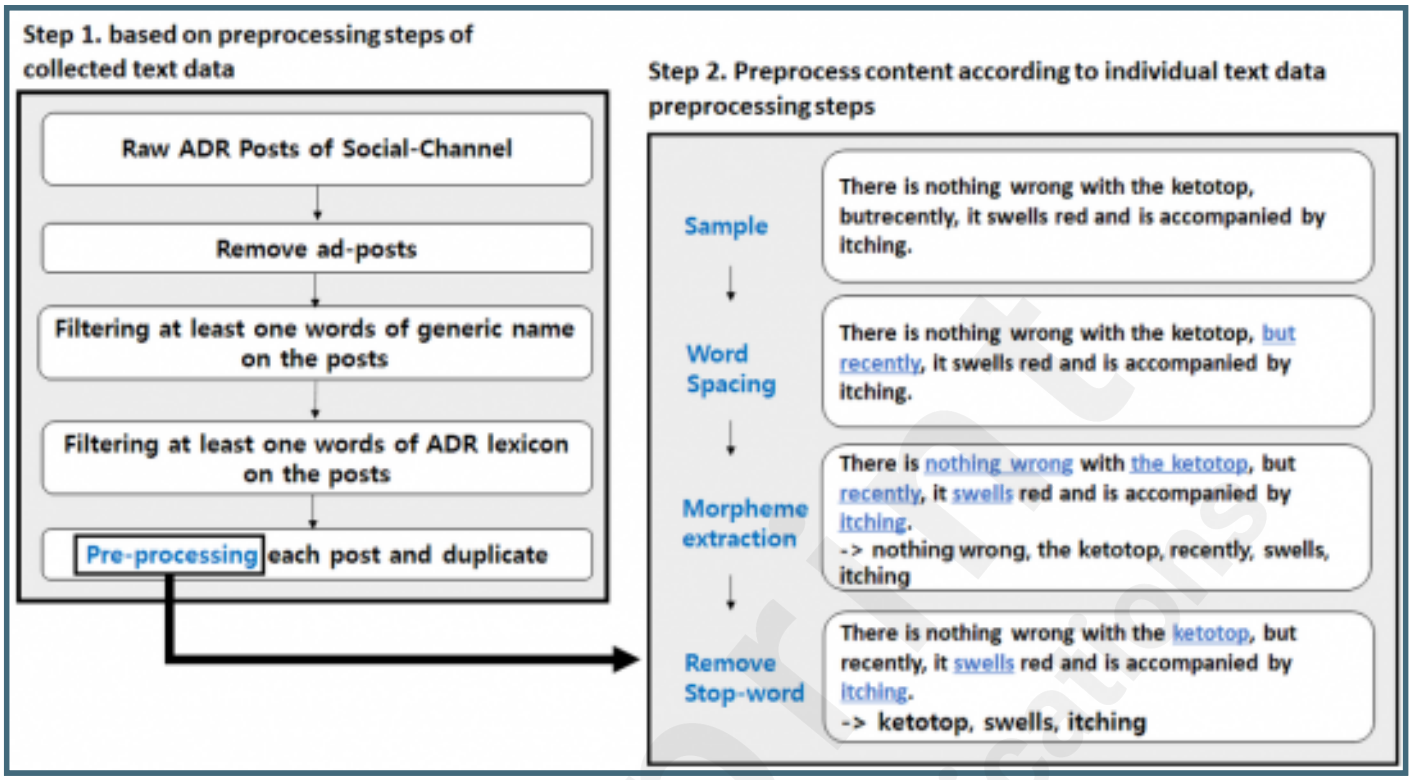
Supplementary Files

Figures

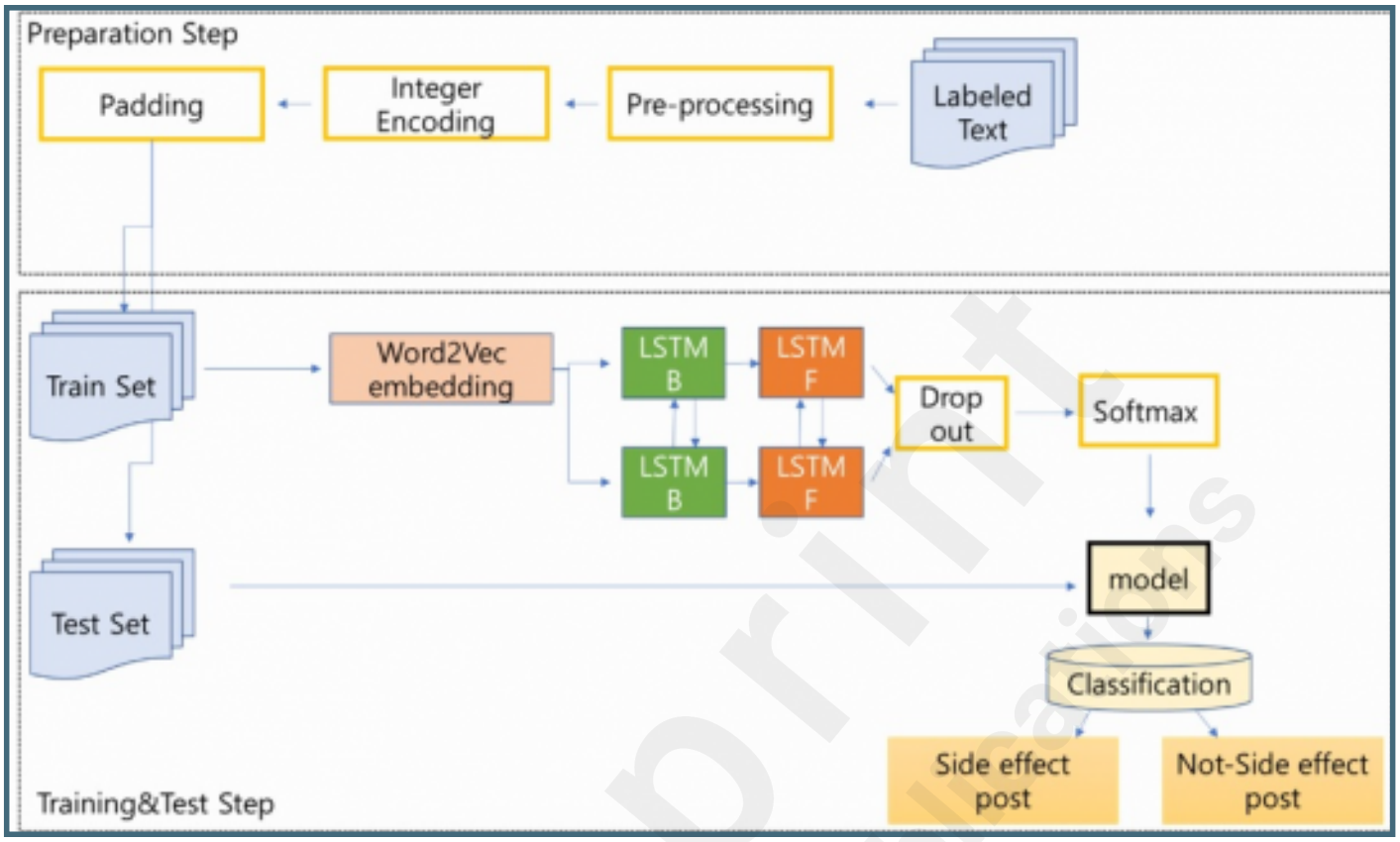
Overview of this research.



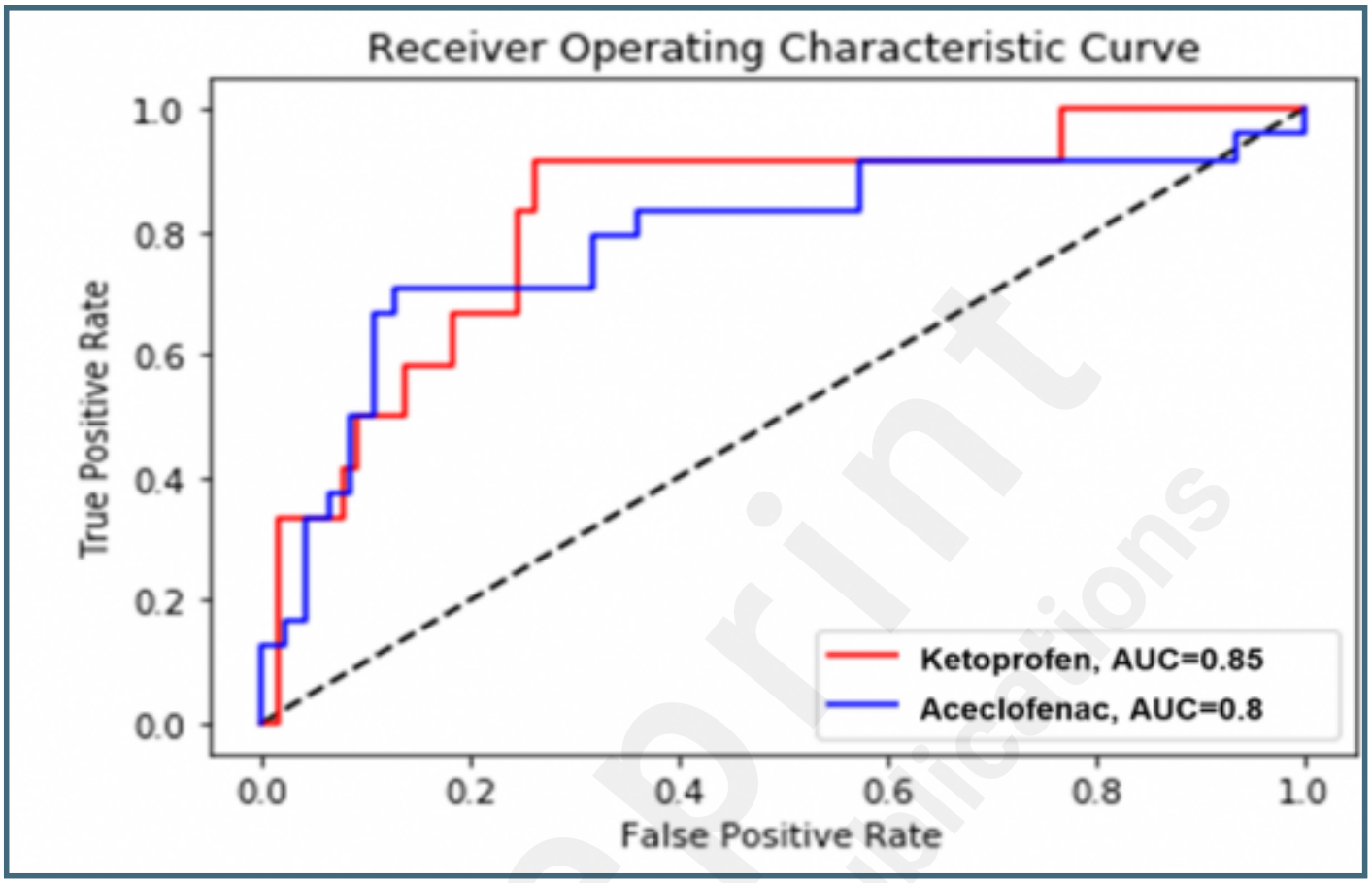
Detailed steps of text preprocessing.



The framework of the Bi-LSTM-based classification approach.



ROC analysis of two drugs to identify ADR posts.



ROC analysis of Bi-LSTM with other classification models.