2024 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB) | 979-8-3503-6426-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/BMSB62888.2024.10608314

Drug-BERT : Pre-trained Language Model Specialized for Korean Drug Crime

Jeong Min Lee[1, 2] , Suyeon Lee[3], Sungwon Byon[1], Eui-Suk Jung[1], Myung-Sun Baek[1, 2]

1: Electronics and Telecommunications Research Institute (ETRI)
Daejeon, Korea
2: University of Science and Technology (UST)
Daejeon, Korea
3: Yonsei University Department of Artificial Intelligence
Seoul, Korea
faraway@etri.re.kr, isuy.groot@yonsei.ac.kr, {swbyon, esjung, sabman}@etri.re.kr

Abstract— **We propose Drug-BERT, a specialized pre-trained language model designed for detecting drug-related content in the Korean language. Given the severity of the current drug issue in South Korea, effective responses are imperative. Focusing on the distinctive features of drug slang, this study seeks to improve the identification and classification of drug-related posts on social media platforms. Recent drug slangs are gathered and used to collect drug-related posts, and the collected data is used to train the language model. The designed pre-trained model is DRUG-BERT. The results show that fine-tuned DRUG-BERT outperforms that of the comparative models, achieving a 99.43% accuracy in classifying drug-relevant posts. Drug-BERT presents a promising solution for combatting drug-related activities, contributing to proactive measures against drug crimes in the Korean context.**

Keywords—drug slang, natural language processing, pre-trained language model, classification

I. INTRODUCTION

South Korea has long been regarded as a drug-free nation [1]. However, in recent times, drug-related crimes have emerged as a serious social issue in the country. This problem has reached such alarming proportions that it is now considered a common concern not only among law enforcement authorities but also among the general public. Particularly, with the increasing prevalence of drug transactions through social media such as X (twitter), YouTube, internet communities, the prevention and detection of the drug transactions has become more difficult.

The rise of drug transactions on social media presents new challenges, especially as the use of drug-related slang makes it increasingly difficult to discover such activities. This paper designs and implements a Drug-BERT which is a pretrained language model to detect the Korean-based drug-related contents in social media and understand the nuances of expression.

II. PRETRAING METHODOLOGY FOR DRUG-BERT MODEL

Drug-related posts encompass a broad range of content, including both posts about drug transactions and those related to drug cultivation. To capture them, it is essential to train a language model Drug-BERT optimized for the drug domain by learning text data containing drug-related terms. In drug-related posts, there is a tendency to use not only explicit terms like marijuana and methamphetamine but also slang terms (e.g. “weed(떨)” and “ice(아이스)”). Traditional language models trained on datasets composed of everyday language may fall short in detecting the specialized domain of drug crimes. Drug criminals use commonly used words to secretly express drugs. To distinguish between sentences related to drug crimes and those that are not, a specialized language model for drug slang is required to handle the ambiguity of words effectively. Therefore, Drug-BERT model is trained by drug-related terms.

III. DRUG CRIME-RELATED DATA COLLECTION AND DATA SET CONSTRUCTION

We initially recognized the need to establish drug-related slang dictionary in order to collect data related to drugs from social media platforms. After defining drug-related slang, we proceeded to crawl posts on the web based on this vocabulary. The process is outlined below.

A. Keywords Selection

The drug slang keywords in our study were extracted from the "Drug Vocabulary Dictionary" derived from the Smart Policing Big Data Platform, an entity under the National Police Agency of South Korea. We considered the keywords to be validated by domain experts for accuracy and reliability.

These consist of 678 drug-related terms, such as pruning, awakening, weed, ice, pipe, and others, which may appear to be commonly used in everyday life but are actually utilized as drug slang. Detailed explanation of the terms is presented in the Homepage of the Smart Policing Big Data Platform [2].

B. Data Collection

We choose three social media platforms, X (Twitter), Youtube and DCInside (Korean most popular anonymous internet forum), as our source of data. DCInside stands out as a widely used anonymous internet forum in South Korea with an "anything goes" characteristic, fostering an environment where diverse discussions and the sharing of information occur [3]. It includes content that might not be accessible on conventional websites, even extending to illegal information and activities such as drug trafficking or sharing illicit videos [4]. In particular, we collect comments on Tweets and YouTube videos, and titles and comments on DCInside distributed from January 2021 to January 2023. Table I describes the corrected drug-related posts according to social media platforms.

979-8-3503-6426-2/24/$31.00 ©2024 IEEE

TABLE I. DATA STATISTICS VIA SOCIAL MEDIA PLATFORM

Social Media Platform	Count
X (Twitter)	24,926
YouTube	1,395,875
DCInside	24,926
Total	1,445,727

C. Data Preprocessing

In the data preprocessing phase, duplicate data was removed to minimize redundancy and ensure accurate analysis. Additionally, unnecessary links and extraneous information were deleted to refine the dataset. It is crucial to exclude sensitive data, such as IDs, IP addresses, phone numbers, and emails, from the collected data on websites. Therefore, we removed such data before proceeding with the pretraining. The refined data were then stored in a database, ready for subsequent utilization by language models.

IV. DRUG-BERT IMPLEMENTATION

A. Pretraining

We designed a pretrained language model based on the Masked Language Modeling (MLM) for the drug-specific domain. For the MLM a masking probability of 15% is applied during training [5]. The model randomly selected positions for masking, excluding special tokens such as [CLS] and [SEP], which respectively represent the beginning and end of each sentence. Subsequently, positions for replacement were identified, with an 80% probability of using a masked token and a 10% probability of employing a random token. To facilitate corpus training, the Bert Word Piece Tokenizer, with a vocabulary size of 30,000, was employed. Table II shows the hyperparameters for the designed model.

TABLE II. THE HYPERPARAMETERS USED FOR PRE-TRAINING

Hyperparameter	Value
Number of Hidden Layers	12
Hidden Size	512
Number of Attemtion Heads	8
Dropout	0.1
Max Sequence Length	512
Max Steps	20K
Hidden Act	gelu

B. Fine-Tuning for Classification

We fine-tuned the language model using drug-relevant and irrelevant posts for the single-labeled classification. This process led to the creation of a Drug-BERT classifier tailored for classifying drug-related posts. The dataset utilized for fine-tuning were randomly sampled approximately 10% of the data to ensure diverse and representative coverage of drug-related posts. The data statistics and the example is illustrated in Table Ⅲ and Table Ⅳ.

TABLE III. DATA STATISTICS FOR FINE-TUNING

	Train	Valid	Test	Total
Irrelevant	41,755	13,919	13,918	69,592
Relevant	41,756	13,918	13,919	69,593
Total	83,511	27,837	27,837	139,185

TABLE IV. EXAMPLE OF RELEVANT AND IRRELEVANT POSTS

Irrelevant	오픈 1 주년이라고 이틀동안ㅋㅋ**아이스** 카페모카 천원에 사왔삼 낼 또 가야지
	Some cafe celebrated their one-year anniversary, so they sold an **ice**d cafe mocha for just one dollar over the next two days. Gotta go again tomorrow, lol!
Relevant	북한산 최상급 퀄리티 **아이스** 팔아요. 안녕하세요, 아이스 딜러 얼음 왕자입니다. 보증금 예치 완료입니다. 정확하고 안전한 드랍 약속 드립니다.
	Offering top-quality ice from Mt. Bukhansan. Hello, I'm the **Ice** Dealer, the Ice Prince. The deposit has been completed. I guarantee accurate and secure drop-offs.

V. EXPERIMENT

We evaluated the performance of Drug-BERT through a task focused on classifying drug-related posts. The classification accuracy of Drug-BERT is compared with that of a general BERT-based models trained on a universal dataset, including BERT-base [5], KLUE-BERT [6], and KoBERT [7].

The results demonstrate that DRUG-BERT classification model achieves an accuracy of 99.43%, outperforming that of known pre-trained language models, as illustrated in Table Ⅴ. While computational efficiency between Drug-BERT and general pretrained models is similar, Drug-BERT demonstrates superior performance in detecting drug-relevant posts for real-world deployment considerations.

TABLE V. COMPARISON OF CLASSIFICATION ACCRUACY

Model	Pre-fine-tuning	Post-fine-tuning
BERT-base	0.5000	0.8969
KLUE-BERT	0.4872	0.9743
KoBERT	0.5010	0.9744
Drug-BERT (our model)	0.5422	0.9943

VI. CONCLUSIONS AND FUTURE WORK

In this paper, Drug-BERT model is designed to detect to drug crime-related posts in social media platforms. The Drug-Bert is pre-trained using specialized drug-crime related keywords and terms, and be refined through precise fine-tuning. Therefore, the designed Drug-BERT model can achieve the best performance among the existing BERT-based language models.

Our study has several limitations. Firstly, our data coverage is currently restricted, excluding the deep web and dark web, which may result in potential blind spots of drug-related activities [8]. Additionally, while our language model has undergone extensive training, further enhancement and stability can be achieved by incorporating more diverse datasets for continuous learning. Looking ahead, it is crucial to adapt to the dynamic nature of drug crimes by capturing emerging drug slang and expanding our vocabulary through periodic updates to the word dictionary. We will strengthen our response to illegal drug activities by incorporating datasets from a wider range of sources, including the deep web and dark web.

ACKNOWLEDGMENT

This research was supported and funded by the Korean National Police Agency. [Project Name: 112 Emergency Dispatch Decision Support System/ Project Number: PR08-03-000-21]

REFERENCES

[1] Cho, P. I. (2004). Drug control policy in Korea. Vancouver: International Centre for Criminal Law Reform and Criminal Justice Policy.

[2] Korea Smart Policing Big Data Platform, Accessed: Jan. 11, 2024, [Online] https://www.bigdata-policing.kr/product/view?product_id=PRDT_90

[3] Yang, S. (2017). Networking South Korea: Internet, nation, and new subjects. Media, Culture & Society, 39(5), 740-749.

[4] Joohee, K., & Chang, J. (2021). Nth room incident in the age of popular feminism: a big data analysis. Azalea: Journal of Korean Literature & Culture, 14(14), 261-287.

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[6] Park, S., Moon, J., Kim, S., Cho, W. I., Han, J., Park, J., ... & Cho, K. (2021). Klue: Korean language understanding evaluation. arXiv preprint arXiv:2105.09680.

[7] KoBERT https://huggingface.co/skt/kobert-base-v1

[8] Jin, Y., Jang, E., Cui, J., Chung, J. W., Lee, Y., & Shin, S. (2023). DarkBERT: A Language Model for the Dark Side of the Internet. arXiv preprint arXiv:2305.08596.