한글 자소분리 및 머신러닝을 이용한 불법 의약품 판매 게시물 탐지방법*

박종혁 [O1] 조소영 [2] 이종혁 [3] 임현태 [2] 정윤경 [1]
성균관대학교 인공지능학과 [1]
광운대학교 정보융합학부 [2]
성균관대학교 데이터사이언스융합학과 [3]
[1,3] {realkaya, aimecca, jhyuk2}@skku.edu, [2] {soyoung2me, 201204058}@kw.ac.kr

Detection of Illicit Drug Selling Post on an Online Community Using Phoneme Separation and Machine Learning Algorithm

Jonghyeok Park[O1] Soyoung Cho[2] Jonghyeok Lee[3] Hyuntae Lim[2] Yungyung Cheong[1]
Department of Artificial Intelligence, Sungkyunkwan University[1]
School of Information Convergence, Kwangwoon University[2]
Department of Applied Data Science, Sungkyunkwan University[3]

요 약

전문 의약품을 온라인에서 매매하는 것은 법적으로 금지되어 있다. 하지만 온라인에서 전문 의약품을 사고파는 게시물을 쉽게 찾아볼 수 있다. 이러한 불법 거래를 막기 위하여 여러가지 시도를 하고 있지만 현 시스템상 사람이 일일이 확인하기란 시간이 많이 소요되는 작업이기에 비효율적이다. 이러한 방법을 개선하기 위해 본 논문은 불법 의약품 중에서도 마약으로 주제를 한정하여 마약 매매 게시물을 수집하고 이를 머신러닝 및 딥러닝을 이용하여 판별하는 방법을 제시하고자 한다. 마약 매매업자가 SNS 상에 구현되어 있는 자동 필터링을 피하기 위해 자음 모음을 분리한 형태로 게시글을 작성한다는 점에 착안하여, 모든 단어의 자음과 모음을 분리하였다. 또한 정확도를 높이기 위하여 3 가지 모델을 사용하여 앙상블 방법을 적용하였다. 제안한 모델은 실제 마약 판매 게시물 중에 모델이 마약이라고 판별한 비율인 재현율이 자음 모음을 분리하기전보다 높게 측정되었다. 앞으로 다양한 판매 게시글을 더 많이 수집할 수 있다면 온라인상의 전문 의약품 매매 게시글을 빠르고 정확하게 판별하여 수많은 불법 거래를 막을 수 있을 것이다.

1. 서 론

최근 SNS상에서 의약품을 사고파는 게시글을 쉽게 찾아볼 수 있다. 한 매체에 따르면 온라인에서 2개월간 거래되는 의약품이 1000건을 넘는다[1]. 전문 의약품은 법적으로 온라인에서 거래가 금지되어 있지만 SNS상에서는 이러한 전문 의약품을 판매한다는 글을 발견하는 것은 어렵지 않다. 전문 의약품은 마약 같은 의약품도 있지만, 미프진(낙태), 멜라토닌(수면유도제), 삭센다(비만치료제)와 같은 의약품도 포함된다. 불법거래 의약품을 거래하기 위한 수단으로 네이버 카페, 인스타그램, 트위터 같은 SNS를 이용하는 경우가 많기 때문에, 해당 게시글이 불법의약품 판매 게시글인지 아닌지를 판별해야 할 필요성이 있다. 현재 시스템에서는 SNS 이용자들의 신고에 의해 처리되기 때문에, 대처가 늦고 사용자들이 신고를 하지 않는 경우도 많다. 따라서 본 논문은 불법 의약품 중에서도 마약 판매 SNS 게시글을 키워드를 통해 수집하고 자소 분리한 후, 머신러닝 및 딥러닝 알고리즘 앙상블 기법을 적용해 불법 판매 여부를 분류하는 방법을 제시하고자 한다.

2. 관련 연구

마약 거래 게시글뿐만 아니라 욕설과 같이 온라인상에서 발생할 수 있는 피해를 막기 위해 많은 연구가 진행되었다. Ji Ho 와 Fung 은 트윗 게시글에 포함되어 있는 욕설을 탐지하는 방법으로 CNN 기반의 3 가지 모델을 제시하였다[2]. Rout 과 그의 동료들은 기업의 수익과 연결되는 제품 리뷰 중에서도 허위 리뷰를 탐지하기 위해 준 지도학습(Semi-Supervised Learning)을 이용하여 악성 리뷰를 탐지하는 방법을 제시하였다[3]. Kulsrud 는 온라인 대화 초기 단계에서 발생할 수 있는 그루밍 성폭행을 탐지하기 위해 Logistic Regression, Ridge Regression, Naïve bayes, SVM, Neural Network 를 사용하였다. 또한 각 Bag of Word 와 TF-IDF 를 이용하여 각 임베딩에 따른 머신러닝 알고리즘의 성능을 측정하였다[4].

3. 데이터

* 이 논문은 2019 년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원(No. 2019R1A2C1006316)과 2019 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2019-0-00421, 인공지능대학원지원).

3.1. 데이터 수집

실험에 사용한 데이터는 트위터의 고급검색 기능이 적용된 웹 크롤링 패키지인 twitterscrapper 를 이용해 수집하였다. 검색 키워드는 '아이스', '물뽕', '도리도리', '크리스탈', '#아이스', '#물뽕', '#크리스탈' 이다.

검색 키워드, 사용자 아이디와 검색된 게시글을 가져올 수 있었으며 최종적으로 수집된 데이터의 양은 51,178 명의 사용자가 작성한 106,448 개의 게시글이다. 모델을 훈련시키기 위하여 각 게시글의 마약 판매를 위한 게시글 여부가 구분되어야 하기 때문에, [표 1]처럼 각 게시글이 판매 게시글인 경우 1 로, 아닌 경우엔 0 으로 라벨링(Labeling)을 진행하였다.

[표 1] Labeling 예시

집에서도 만들 수 있는 나만의 아이스크림 제조법	0
사람들이 잘 모르는데 진짜 유명해졌으면 하는…	0
ㅁㅣ프진 작ㄷㅐㄱㅣ 안전거래 ㅋㅏ톡주세요 …	1
#ㅇㅏㅇㅣ스삽니다 #얼음삽니다#ㅊㅏㄱㅏ운술…	1

게시글을 살펴보면, 마약 판매를 위한 게시글에는 '아이스',' 도리도리' 와 같은 은어가 많이 포함되어 있으며, 일반적인 단어의 형태가 아닌 'ㅇㅏㅇㅣ스',' 도ㄹㅣ도ㄹㅣ' 처럼 자음과 모음이 의도적으로 분리되어 있음을 확인할 수 있다.

3.2. 클래스 불균형 문제

라벨링 된 데이터의 분포를 살펴보면, 0 으로 라벨링 된 데이터의 수는 102,670 개이며, 1 로 라벨링 된 데이터의 수는 3,778 개로 전체 데이터의 약 3.5%에 불과하다. 이러한 데이터 불균형의 문제는 모델에 있어서 심각한 편향과 성능 저하로 이어진다.

데이터 불균형을 해소하기 위해선 여러가지 방법론이 제시되어왔다. Wei 와 Zou 는 각 단어를 랜덤하게 바꾸는 방법, 특정 단어의 동의어를 찾아서 바꾸는 방법, 두 단어의 위치를 바꾸는 방법, 마지막으로 문장 내 특정 단어를 제거하는 방법을 제시하였다[5]. 한편 Fadaee 와 그의 동료들은 NMT(Neural Machine Translation)를 이용하여 원 문장을 다른 언어로 번역하고, 번역된 문장을 다시 원 문장으로 재번역하는 방법을 이용하여 데이터의 양을 늘렸다[6]. 특정 클래스의 데이터가 많은 경우, 해당 클래스의 일부만 쓰는 undersampling 기법이 흔히 사용된다. Sharma 와 그의 동료들은 Random undersampling 을 하는 것과 데이터 샘플링을 하지 않는 것을 비교해 분류 성능을 크게 향상됨을 보였다[7]. 본 논문에서는 0 으로 라벨링 된 데이터가 압도적으로 많았기 때문에 0 으로 라벨링 된 데이터의 경우엔 1 과 비슷한 개수인 5000 개로 random undersampling 하여 사용하였다.

3.3. 전처리

게시글 데이터에는 한글뿐만 아니라 숫자, 영어, 특수문자 등이 섞여 있기에 전처리 과정을 선행하여 진행하였다.

본 논문에서는 수집한 데이터셋에 특수문자 비율이 높기 때문에 특수문자 제거를 수행하였으며, 추가로 사진자료와 링크(Link)를 제거하였다. 또한 자음 모음을 분리한 경우와 분리하지 않은 경우의 모델의 성능을 비교하기 위해 전처리한 데이터와 전처리 후 자소를 분리한 데이터를 각각 준비하였다.

4. 모델

4.1. 머신러닝 모델

우선, 기본적으로 텍스트 이상치 탐지 및 스팸 텍스트탐지 등의 분류문제에서 많이 사용되는 머신러닝 모델인 Naïve bayes 와 분류성능이 뛰어나다고 평가되는 SVM(Support Vector Machine)을 머신러닝 모델로 사용하였다.

Naïve bayes(NB)는 확률기반 모델로 베이즈 이론에 그 기초를 두고 있다. 텍스트 분류에서 Naïve bayes 는 여러 문장과 해당 문장의 클래스가 주어지면 모든 단어에 대해 각 클래스별로 단어의 출현 확률을 계산하고, 새로운 문장이 주어졌을 때, 이미 계산해둔 단어 별 출현확률을 기반으로 새로운 문장에 대해서 어느 클래스에 속할지를 계산하여 분류문제를 해결한다. 알고리즘은 비교적 간단하지만 자연어 분류문제에서 뛰어난 성능을 보여 많이 쓰는 알고리즘 중 하나이다.

SVM 은 분류, 회귀 및 특이값(outliers) 감지에 사용되는 지도학습 방법이다. SVM 은 n 차원 데이터들 사이를 구분해줄 수 있는 (n-1)차원의 초평면(Hyperplane)을 만들어 주어진 데이터에 대하여 클래스를 나누어 줄 수 있는 경계선을 만드는 과정을 진행한다. SVM 은 분류문제에 있어서 Feature 수에 영향을 받지 않기 때문에 텍스트 분류 문제에서 적합하며, 텍스트 분류는 대부분 선형으로 분리 가능하기 때문에 linear Kernel 을 사용하는 SVM 이 더 잘 작동한다. 또한 SVM 은 텍스트 데이터와 같이 매우 sparse 한 양상에 더욱 효과적이다[8].

LSTM 은 RNN 의 변형된 모델로 기존 RNN 의 문제점인 입력 벡터의 길이가 긴 경우 오차 역전파법 수행 시 뒤쪽으로 갈수록 업데이트가 잘 안되는 기울기 소실(Gradient Vanishing) 현상을 보완한 방법이다.

4.2. 앙상블 방법

앙상블(Ensemble)은 단일 모델로 성능을 내기보다 다수의 모델을 이용하여 더 나은 성능을 내는 방법이다[9]. 앙상블 방법에는 Voting, Bagging, Boosting, Stacking 등이 있으며, 본 논문에서는 Voting 중에서도 각 모델이 예측한 확률을 합산하여 결과를 내는 Soft Voting 보다 각 모델이 예측한 결과를 이용하여 투표하는 방식인 Hard Voting 을 사용하였다.

5. 실험 결과

수집한 데이터에서 0 으로 라벨링 된 데이터 중 Random undersampling 한 데이터 5,000 개와 1 로 라벨링 된 데이터 3,778 개를 사용하였다. 훈련 데이터셋과 테스트 데이터셋의 비율은 8 : 2 로 정하였다.

본 논문에서는 성능 지표로 재현율(Recall score)을 사용하였다. 재현율은 실제 정답 중에 모델이 정답이라고 판단한 비율을 의미한다. 즉, 실제 마약 판매 게시글 중 모델이 마약 판매 게시글이라고 판단한 비율을 의미한다. [표 2]는 세 모델과 앙상블의 정밀도와 재현율을 나타낸다.

[표 2] Precision & Recall score

구분	자소 분리 전		**자소 분리 후**	
Score	Precision	Recall	Precision	Recall
NB	0.999	0.994	0.993	0.994
SVM	0.998	0.988	0.998	0.993
LSTM	0.997	0.994	0.993	0.996
Ensemble	0.999	0.994	0.996	0.997

재현율은 Naïve bayes 를 제외한 모델에서 자소 분리 후가 자소 분리 전보다 더 높아졌음을 볼 수 있다. 즉, 자음 모음 분리를 하지 않은 경우보다 분리한 경우 재현율이 더욱 높아짐을 알 수 있다. [표 4]는 [표 3]의 문장에 대해 자음 모음이 분리되기 전과 후의 각 모델 및 앙상블한 모델의 판별 결과 예시이다. 예시 문장에 대해 자음과 모음을 분리한 것이 분리를 하지 않았을 때보다 더 정확하게 판별함을 확인할 수 있다.

[표 3] 불법 의약품 판매 문장 예시

자소 분리 전	작ㄷㅐ기 떨 ㅇㅏ이스 차가운술 자세한 내용은 쪽지로 …
자소분리 후	ㅈㅏㄱㄷㅐㄱㅣ ㄸㅓㄹ ㅇㅏㅇㅣㅅㅡ ㅊㅏㄱㅏㅇㅜㄴㅅㅜㄹ ㅈㅏㅅㅔㅎㅏㄴ ㄴㅐㅇㅛㅇㅇㅡㄴ ㅉㅗㄱㅈㅣㄹㅗ …

[표 4] 자소 분리 전 후 모델의 분류 결과 비교 예시

구분	자소 분리 전	**자소 분리 후**
실제 정답	True	
Naïve bayes	False	True
SVM	False	False
LSTM	False	True
Ensemble	False	True

6. 결론 및 향후 연구

본 논문에서는 머신러닝 및 딥러닝을 이용하여 불법 마약 판매 게시글을 판별하는 방법을 제시하였다. 불법 판매 게시글의 특성을 파악하여 자음과 모음을 분리하는 전처리를 모든 데이터에 적용하였고, 상호보완을 위해 각기 다른 모델들을 앙상블하여 정확도를 향상시켰다.

본 연구를 활용하면 게임 내 욕설탐지 및 일부 부적절한 게시글 탐지에도 적용할 수 있을 것으로 예상된다.

향후 연구 과제로는 undersampling 과 같이 전체 데이터셋 중 일부를 사용하는 경우 모집단을 대표할 수 있는 데이터를 추출하는 방법과, [꺄 = '77 ㅑ', 나 = 'ㄴr']와 같이 숫자와 문자의 조합으로 변형된 글자를 인식하는 방법에 대해 연구하고자 한다.

참 고 문 헌

[1] 정새임, "불법거래 판치는 온라인…일평균 31 건 이상", 청년의사, 2019 년 12 월 3 일

[2] Ji Ho, P., Fung, P., "One-step and two-step classification for abusive language detection on twitter.", *in proceedings of the Workshop on Abusive Language Online, Association for Computational Linguistics, Vancouver, Canada*, pp. 41-45, 2017

[3] Rout, J. K., et al., "Revisiting semi-superviesd learning for online deceptive review detection." *in proceedings of IEEE Access 5*, pp. 1319-1327, 2017

[4] Kulsrud, H. K., "Detection of cyber grooming during an online conversation" MS thesis, NTNU, 2019

[5] Wei, J. W., Zou, K., "EDA : Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks " *in proceedings of the 2019 Conference on Emprical Methods in Neural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 6382-6388, 2019

[6] Fadaee, M., Bisazza, A., Monz, C., "Data augmentation for low-resource neural machine translation." *in proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 567- 573, 2017

[7] Sharma, N., Dileep, A. D., Thenkanidiyoor, V., "Text Classification Using Hierarchical Sparse Representation Classifiers." *in proceedings of the 16th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE*, pp. 1015-1019, 2017.

[8] Joachims, T., "Text Categorization with Support Vector Machines : Learning with Many Relevant Features" *in proceedings of the ECML-98, 10th European Conference on Machine Learning*, pp. 137-142, 1998.

[9] Polikar, R., "Ensemble based systems in decision making" *in proceedings of the IEEE Circuits and Systems Magazine (Volume:6 , Issue3,Third Quarter),* pp. 21-45, 2006