BERTweet 기반의 마약 판매 게시글 탐지 및 PCA를 통한 모델 경량화

이지용† 양승원* 윤동영** 이용균*** 이은수****

† 청운대학교 (학부생) *건국대학교 (학부생) **가천대학교 (학부생) ***인하대학교 (학부생) ****상명대학교 (학부생)

Detection of drug sales posts based on BERTweet and Dimension reduction via PCA

Ji-Yong Lee† Seung-Won Yang* Dong-Yeong Youn** Yong-Gyun Lee*** Eun-Su Lee****

† Chungwoon University(Undergraduate) *Konkuk University(Undergraduate) **Gachon University(Undergraduate) ***Inha University(Undergraduate) ****Sangmyung University(Undergraduate)

요 약

오늘날 X, Instagram, Snapchat 등 SNS 플랫폼에서의 마약 판매 및 유통이 공공연하게 일어난다. 특히 마약 판매의 경우 심각한 사회 문제로써 범죄 확산의 위험이 있음에도 불구하고 특정 플랫폼에서 탐지 및 차단에 맹점이 있음을 인지하였고, 그 원인이 이미지 인식과 모델의 성능에 자원이 치중되어 텍스트 기반의 범죄 게시글 탐지 미흡의 결과로 이어졌다고 판단하였다. 따라서, 본 논문에서는 BERTweet을 기반으로 한 경량화된 마약 판매 게시글 탐지 시스템을 제안한다.

I. 서론

대검찰청 2023 마약류 범죄 백서에 따르면 해를 거듭하며 단속 현황이 증가하고 있다. 연령별 마약류 사범은 2030이 대부분을 차지하며 심각한 중독성과 함께 낮은 처벌 수위로 최근 5년간 마약류 사범의 평균 재범률은 35%를 차지하고 있다. [1]

현재 마약 범죄의 새로운 온상으로 SNS가 급부상하고 있다. 손쉽게 Telegram, Snapchat, Instagram 등의 대중적인 SNS를 통해 쉽게 마약 판매 게시글을 찾아볼 수 있으며 지속적으로 마약 판매 글을 노출시켜 경각심을 저해하고 익명성을 활용해 수사망을 회피한다. 본 연구에서는 이러한 상황을 타개할 방법으로 게시글의 독자적인 특성을 반영한 텍스트 기반의 경량화 모델을 통해 마약 범죄의 확산을 방지할 수 있는 탐지 시스템을 제안하였다. 본 연구를 통해 이미지 기반의 유해 게시글 탐지의 허점을 보완할 수 있으며 신속한 대응이 가능할 것으로 기대한다.

II. EDA 및 모델 설계

2.1 데이터셋

데이터 셋은 마약 판매 게시글인 text와 그에 따른 label 두 피처로 구성되어 있으며 X, Instagram, Snapchat, Steam Community 등 직접 여러 플랫폼을 대상으로 300건의 마약 판매 게시글을 수집하였다. 또한, 공익 게시글인 마약 관련 뉴스 기사를 300건, 일상 게시글 300건의 데이터를 확보하였다. [2][3] 마약 판매 데이터를 분석한 결과 고유의 패턴이 있음을 인지하였다. 이에 따라 마약 판매 데이터를 500건으로 분할 증강하여 대략 1,100건의 데이터 셋을 구성하였다. 범죄의 범주를 마약으로 한정한 점, 과도한 데이터 수집은 데이터의 중복성이 있음을 감안하여 적절한 규모로 구축하였다.

2.2 BERTweet

자연어 데이터를 정형화하기 위해 Transformer 기반의 자연어 처리 모델인 BERT를 사용할 것이며 특히 트윗이 사전학습 된 BERTweet 모델을 본 연구에서 선정하였다. [4][5][6]

선정한 이유는 다음과 같다. BERT의 경우 고려했던 RoBERT, KoBERT, DistilBERT 등 다양한 파인튜닝 모델이 있었지만, 공용어인 영어 트윗이 학습되어 다양한 국가의 범죄 게시글을 대상으로 탐지할 수 있다는 점, 사이버 금융 범죄, 테러 등 다중 카테고리의 범죄까지의 시스템 확장 가능하다는 점 그리고 마약 판매 게시글의 경우 정확한 문법이나 어순을 반영하지 않은 점을 근거로 BERTweet이 적합하다고 판단하였다. 자연어라는 비정형 데이터에서 임베딩 벡터를 추출해 정형화한 후 다중 분류를 진행한다.

2.3 EDA

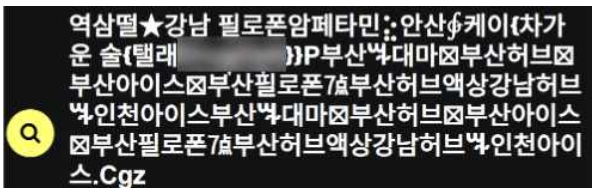

그림 1. Snapchat 마약 판매 게시글

그림 1을 통해 알 수 있듯이 마약 판매 게시글의 경우 다른 게시물과는 달리 판매자와 접촉할 수 있는 텔레그램 아이디, 각종 마약 은어 그리고 특수문자의 조합으로 구성되어 있음을 파악할 수 있다.

본래 자연어 데이터를 직관적으로 분석한 결과 고유의 패턴을 인식하였지만, 데이터를 정형화한 후에도 그 특성이 유지됨에 있어서는 확신이 필요하다. 따라서 임베딩 벡터에 차원 축소를 진행하여 비지도 학습인 k-means clustering(cluster=3)을 적용해 시각화를 진행하였다.

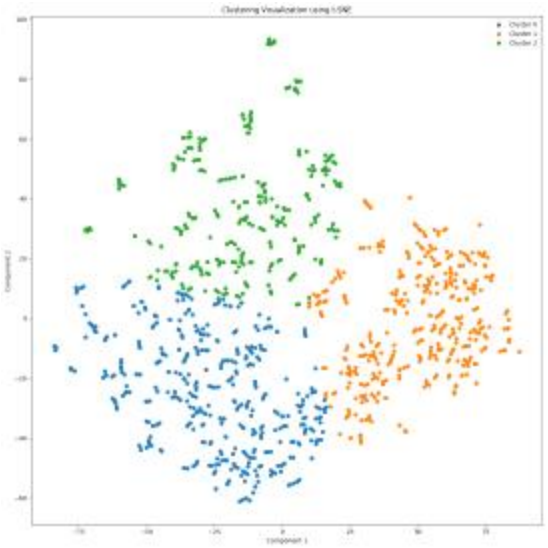

그림 2. t-sne 2차원 축소 및 k-means clustering(cluster = 3)

마약 판매 데이터의 특성을 증명하기 위한 EDA로, 상단의 cluster 2에서 마약 판매의 높은 일치율을 보인다. 이는 임베딩 벡터만으로 일정 수준 유의미한 마약 판매 탐지가 가능함을 시사한다. 다만, 나머지 cluster에서는 데이터의 혼합이 심하게 일어나 지도학습이 필요함을 알 수 있다.

2.4 모델 설계

모델은 다음의 프로세스를 따른다. 데이터셋으로부터 BERTweet 임베딩 벡터를 추출해 PCA로 차원 축소를 진행하고, RandomSearchCV로 최적의 파라미터를 선정하였다. 데이터 셋을 분할하고 다중 분류 지도학습인 XGBoost에서 모델 학습이 이루어진다. 이후 실제 Snapchat에서 수집한 마약 판매 게시글로 최종 테스트를 진행한다.

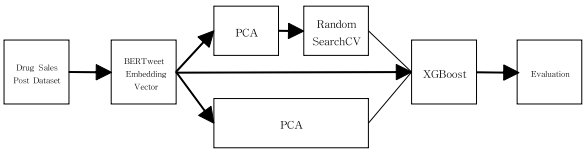

그림 3. 모델 설계 다이어그램

III. 모델 최적화 방안

3.1 모델 최적화 - PCA

BERT 모델의 기본 임베딩 벡터 차원은 768

차원으로 매우 고차원의 정형 데이터이다. 이는 연산량의 증가로 차원의 저주(Curse of Dimensionality)가 발생할 가능성이 있어 적절한 차원으로 축소하는 최적화가 필요하다. 누적 분산의 92%, 95%가 되는 지점, 즉 elbow point로 주성분을 선정 및 데이터를 축소하여 연산을 줄이는 모델 경량화를 기대할 수 있다. 해당 그림 4에서는 기존 768차원의 데이터를 72차원까지 축소가 가능함을 보인다.

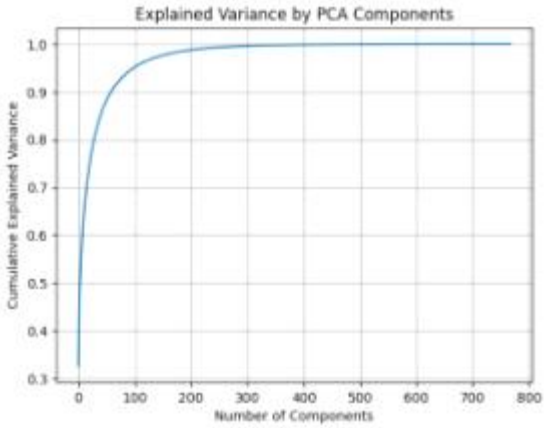

그림 4. optimal dimension - 72

3.2 모델 최적화 - RandomSearch

{'subsample': 0.7, 'n_estimators': 300, 'min_child_weight': 1, 'max_depth': 4, 'learning_rate': 0.1, 'gamma': 0.05, 'colsample_bytree': 0.9} RandomSearchCV 튜닝 후 다음의 파라미터가 선정되었고 해당 결과로 학습을 진행하였다. 해당 연구는 궁극적으로 실시간 탐지를 목적으로 하기에 실행시간에서의 이점을 위해 GridSearchCV를 배제하였다.

IV.모델 성능평가

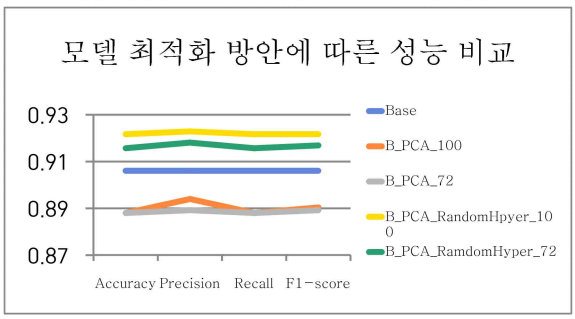

그림 5. 모델 최적화 방안에 따른 성능 비교

최적화 적용과 그렇지 않은 모델 모두 비교한 결과이다. PCA, RandomSearchCV 와 같은 모델 경량화를 진행하여도 성능 차이가 작다는 결과를 확인할 수 있다.

해당 실험 결과를 통해 마약 판매 데이터 즉, 고유의 특성을 가지는 범죄 게시글 특성을 반영해 이를 정형화하여 분류가 가능해지고 이후 모델 최적화를 통해 실시간 탐지 시스템에 적용 가능한 경량 모델을 구축할 수 있음을 시사한다.

V.결론

본 논문에서는 BERTweet, XGBoost와 PCA를 사용하여 마약 판매 데이터를 각각 공익, 일반 게시글과 함께 다중 분류 및 실시간 탐지에 적용 가능한 경량 시스템을 설계 및 개발하였고 실제 특정 플랫폼에서 그 성능이 유효함을 확인하였다.

다만, 일반 사용자의 게시글과 마약 관련 공익 게시글의 경우 분류 및 알려지지 않은 은어 또는 새로운 변조 패턴에 대한 탐지 성능의 미흡함이 확인되었다. 따라서 동일 범주 내에서의 분류 성능 향상과 시스템 확장을 후속 연구를 통해 보완할 예정이다.

[참고문헌]

[1] 대검찰청, 2023년 마약류 범죄 백서 3장, 2023.

[2] Snap Inc. Transparency Report. Snap Inc, n.d., https://values.snap.com/privacy/transparency. Accessed 2024.

[3] X Corp. DSA Transparency Report. X Corp, n.d., https://transparency.x.com/dsa-transparency-report.html. Accessed 2024.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is All You Need, Advances in Neural Information Processing Systems 30, 2017.

[5] Youngjin Jin, Eugene Jang, Jian Cui,

Jin-Woo Chung, Yongjae Lee, Seungwon Shin, DarkBERT: A Language Model for the Dark Side of the Internet, S2W Research, 2023.

[6] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171-4186, 2019.