2024 CISC-W'24

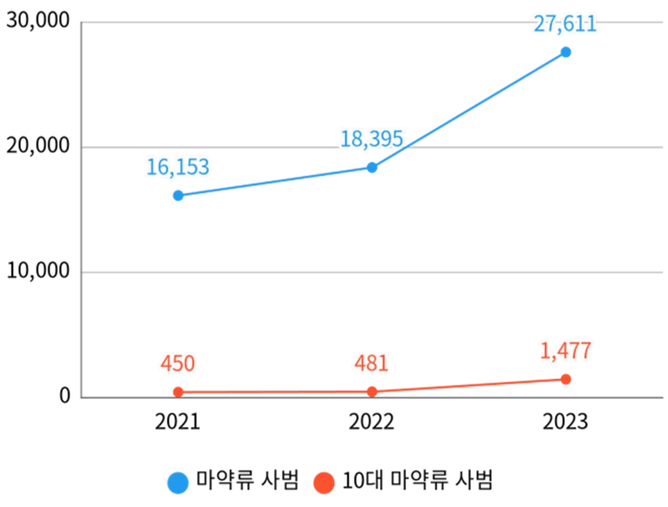
BERTweet 기반의 마약 판매 게시글 탐지 및 PCA를 통한 모델 경량화

이지용(청운대학교), 양승원(건국대학교), 윤동영(가천대학교), 이용균(인하대학교), 이은수(상명대학교)

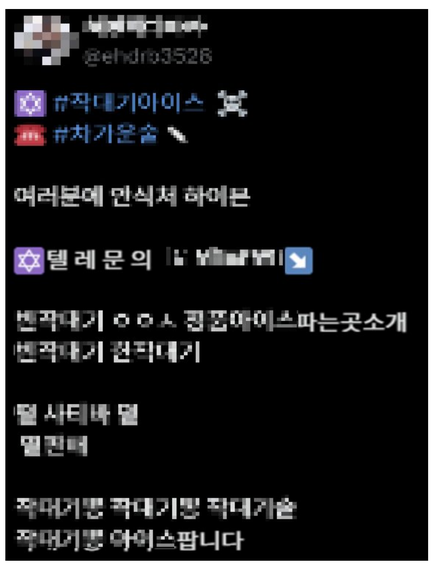
2024 한국정보보호학회 동계학술대회

I. 서론

- 대검찰청에 따르면 해를 거듭하며 마약 범죄율이 지속적으로 증가
- 마약 범죄의 대부분이 **20,30**대에서 차지하고 있고 최근 청소년 마약 범죄도 늘어나고 있는 추세
- 이러한 마약 범죄의 새로운 온상으로 **SNS**가 급부상
- 실제로 **Instagram, X(Twitter), Snapchat** 등 다양한 **SNS** 플랫폼에서 쉽게 마약 판매 게시글 확인 가능



실제 마약 판매 게시글



II. 연구방안

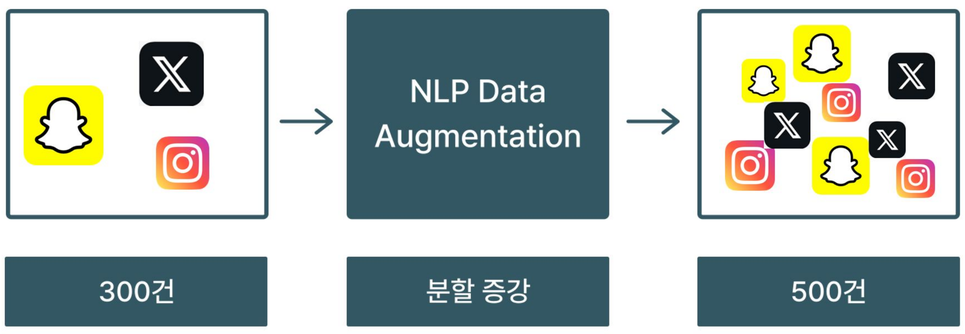
- 대중적인 **SNS**를 통해 야약 판매게시글을 지속적으로 노출시키며 경각심을 저해시키고 마약은어를 사용하여 수사망을 회피
- **Snapchat** 투명성 보고서에 따르면 이미지 기반의 게시글과 아동 성착취 게시글과 같은 플랫폼 이미지에 타격이 큰 게시글 위주로 리소스가 투입
- 이러한 이유 때문에 마약 판매 게시글에 대한 리소스 투입이 적음
- 본 연구에서는 이러한 상황을 타개할 방법으로 게시글의 독자적인 특성을 반영한 텍스트 기반의 경량화 모델을 통해 마약 판매 게시글을 탐지하며 동시에 투입되는 리소스도 줄일 방안을 제시

Snapchat 투명성 보고서

Reason	Content & Account Reports	Content Enforced	% of the Total Content Enforced by Snap
선정적인 콘텐츠	4,271,116	2,266,213	42.1%
아동 성 착취	847,430	239,820	4.5%
괴롭힘	8,524,054	1,193,695	22.2%
위협 및 폭력	836,125	114,315	2.1%
자해 및 자살	188,124	32,841	0.6%
허위 정보	439,233	1,463	0.1%
사칭	440,437	14,557	0.3%
스팸	1,981,115	1,002,278	18.6%
약물	368,732	241,227	4.5%

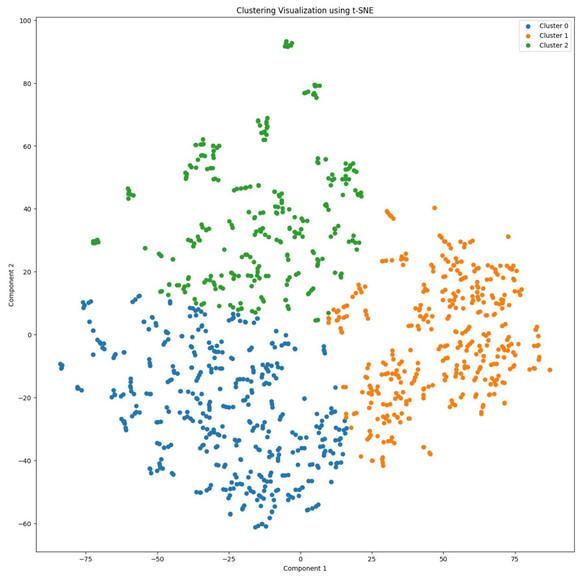
III. 데이터셋

- 데이터 셋은 마약 판매게시글 **300**건, 마약 공익 게시글과 마약관련 뉴스 기사 **300**건, 일반 게시글 **300**건을 사용
- 마약 판매 데이터를 분석한 결과 고유의 패턴이 있음을 인지하여 마약 판매 데이터를 **500**건으로 분할 증강, 총 **1100**건의 데이터셋을 구성



IV. EDA 및 모델 설계

T-SNE 시각화

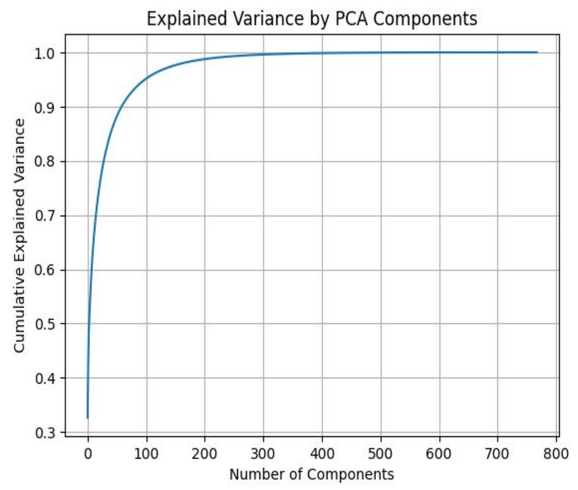


BERTweet 모델

- 자연어 데이터를 정형화하기 위해 **Transformer** 기반의 자연어 처리 모델
- **SNS**특화된 트윗 데이터를 사전 학습한 모델
- 마약 판매게시글의 경우 은어를 사용함과 동시에 정확한 문법이나 어순을 미반영

모델 경량화 방안

PCA



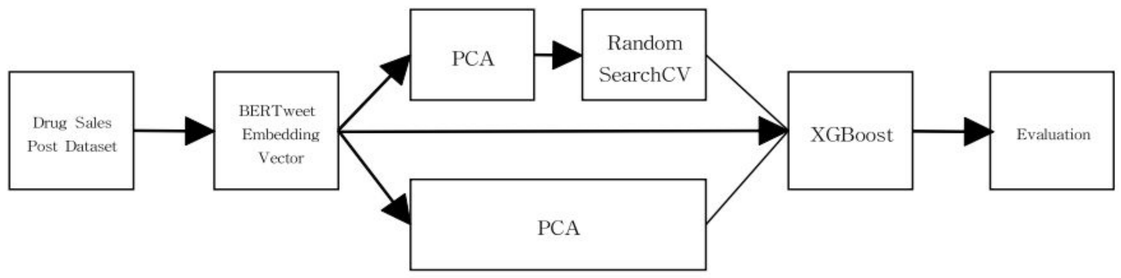
- **BERTWeet** 모델은 기본 **768**차원의 고차원 모델
- 누적 분산의 **92%, 95%**가 되는 지점, 즉 **elbow point**로 주성분을 선정 및 차원 축소

RandomSearchCV + XGBoost

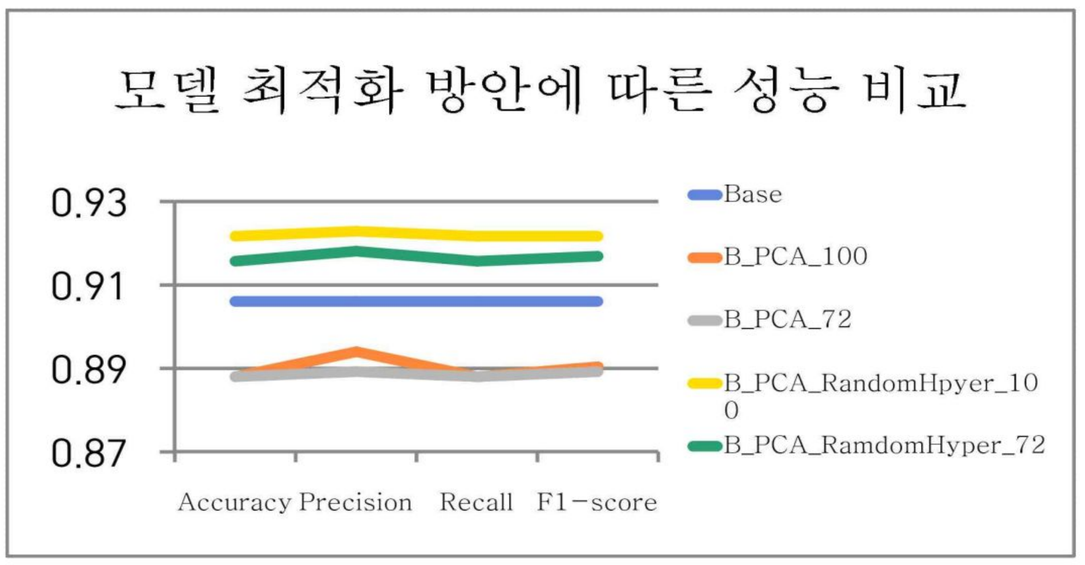
	Initial Parameters		Best Parameters
데이터 샘플링 비율	0.8	유지	0.7
트리 개수	200	100 증가	300
최소 가중치 합	1	유지	1
최대 트리 깊이	5	1 감소	4
학습률	0.1	유지	0.1
트리 분할 기준	0.1	0.05 감소	0.05
피처 샘플링 비율	0.8	0.1 증가	0.9

- 실시간 탐지에 필요한 리소스 절약을 위한 **RandomSearchCV**
- 다중분류와 성능향상을 위한 **XGBoost**

최종 모델 설계 다이어그램



V. 결론



- 모델 경량화를 진행하여도 성능 차이가 크게 떨어지지 않는 결과를 확인
- 해당 실험 결과를 통해 마약 판매 데이터 즉, 고유의 특성을 가지는 범죄 게시글 특성을 반영해 이를 정형화하여 분류가 가능
- 모델 최적화를 통해 실시간 탐지 시스템에 적용 가능한 경량 모델을 구축할 수 있음을 시사