온라인 상의 마약 거래 분석 및 예측을 통한 모니터링 시스템 구현

성균관대학교 글로벌융합학부 데이터사이언스융합전공

양지인 김유진 선해라 우다연

Background

마약의 주 거래처가 온라인 공간에서 이루어지면서, 누구나 어디서든 쉽게 마약 거래를 할 수 있게 되었다. 인터넷의 익명성과 비대면이라는 특성으로 인해 접근성이 높아진 것에 비해 수사와 검거는 더 어려워진 추세이다.

따라서, 본 프로젝트에서는 젊은 층을 대상으로 마약 거래가 주로 이루어지는 트위터의 최근 3년간의 트윗을 수집하여 다음과 같은 분석 목표를 정하였다. 우선, **SNS를 이용한 마약 거래의 최근 경향성을 분석**해보고자 한다. 또한, **마약 거래 게시글을 판별하는 모델을 구축**하는 것을 목표로 한다. 최종적으로 트위터에 업로드되는 마약 거래 게시글의 내용과 위치 등의 정보와 마약 거래의 동향을 한 눈에 보여주는 **실시간 모니터링 시스템을 구축**하고자 한다.

Data

수집 정보
snscrape로 트윗 수집
2021.01.01~2023.03.31
검찰청 마약 분류, 선행 연구, 인터넷 기사, 실제 트윗에서 키워드 선정
11개 대분류, 52개 키워드
date, username, content, media, location 등 13개 feature
제외어 설정하여 수집 시간과 양 단축

1. 마약 거래와 무관한 트윗 필터링
   - 사회 이슈 관련된 트윗 제거
   - 일상적으로 활용되는 단어 제거
   - 성적인 목적의 트윗 제거
2. 텍스트 전처리
   1) 이모지, 특수문자 제거
      - 판매 정보 관련 특수기호 보존
      - url 삭제
   2) 자모음, 대소문자 통합
      - hangul-utils, lower() 활용
   3) 아이디 띄어쓰기 제거
3. mecab 형태소 분석기로 토큰화

Analysis & Methods

Our Research questions:

1. EDA 결과를 정제해서 한 곳에 모아볼 수 없을까?
   - Mecab 텍스트 토큰들 중 일반명사, 고유명사, 외국어를 Word2Vec 분석을 거쳐 각 단어들 간의 유사도를 파악한 후 PCA 주성분 분석으로 차원을 축소함. k-means clustering에서 최적의 k값을 5로 설정해서 마약과 직접적인 관련이 있는 군집 1개를 추출한 후, 해당 군집 내에서 다시 최적 k=2로 설정하여 '마약 거래'와 관련성이 높은 단어들만 시각화함.
   - EDA에서 추출한 위치 데이터를 지도에 시각화 (Figure 10 지도)

Figure 7. 모든 일반명사, 고유명사, 외국어 토큰에 대한 빈도 워드클라우드

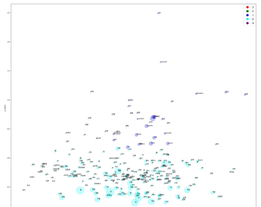

Figure 8. 클러스터링 결과

Figure 9. 클러스터링된 '마약 군집' 내에서 새롭게 형성한 워드클라우드

2. 마약 거래 트윗을 실시간으로 추적하고, 자동으로 분류 해줄 수 없을까?
   - 크롤링 자동화 및 DB 저장: snscrape 트윗 크롤링 코드를 Linux crontab을 사용해 일정시간마다 실행되게 스케줄링 함. 크롤링한 데이터는 goorm 내 구축한 DB 서버 MySQL database에 저장함.
   - 텍스트 classification 모델링: Huggingface의 'albert-kor-base' pre-trained model을 transfer learning 시킴. best accuracy를 보였던 epoch에서의 모델을 최종적인 classification 모델로 선정함.

Prototype Implementation

streamlit 활용 마약 거래 모니터링 시스템 구현

1. 마약 거래 동향

Figure 10. 프로토타입 마약 거래 동향 파악 스크린

2. 마약 거래 게시물 분류 결과

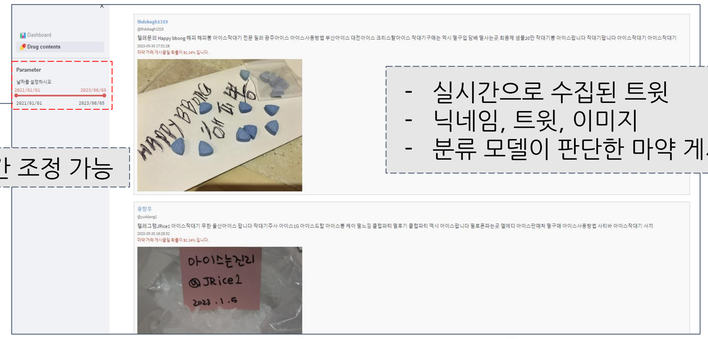

Figure 11. 프로토타입 마약 거래 게시물 분류 결과 스크린

EDA

1.NLP 분석: 트윗 내용 자체에 대한 분석

- 토큰화 결과 일반명사엔 '아이스' 등 마약 은어가, 고유명사엔 '텔레(마약 거래 플랫폼)' 및 지역명이, 외국어엔 마약 판매자 메신저 ID가 많았음(Figure 1).
- 검색 키워드를 제외한 단어들로 워드클라우드(Figure 2)를 형성하여 가시적인 효과를 높임. '캔디' 등 마약 은어가 가장 많았으며, '판매' 등 직접적으로 구매와 관련된 단어들도 빈번하게 등장함.

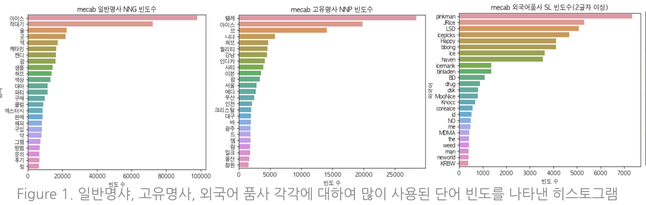

Figure 1. 일반명사, 고유명사, 외국어 품사 각각에 대하여 많이 사용된 단어 빈도를 나타낸 히스토그램

Figure 2. 워드 클라우드

2. 위치 데이터 분석: 마약 거래가 이루어지는 지역 파악

- '울산아이스' 등의 데이터에서 지역명 추출함. '전국행정동리스트' 데이터를 이용해 모든 지역명을 리스트화 한 후 마약 키워드 대분류 별 빈도를 시각화 함.
- 메스암페타민과 MDMA, 대마 모두 '강남', '서울', '부산'이 가장 많이 언급되었고, 그 외 '울산', '수원', '성남'도 언급량이 많았음(Figure 3).

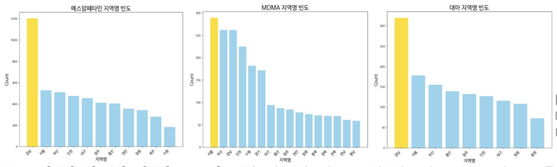

Figure 3. 메스암페타민, MDMA, 대마 마약 대분류별 언급된 지역명 빈도

3. 시계열 데이터 분석: 마약 거래 트윗이 업로드 되는 추이 파악을 위한 분석

- 전체 트윗 추이(Figure 4)와 마약 대분류 별 추이(Figure 5) 분석 결과 2022년부터 SNS 상 마약 거래가 급증하였음을 확인함. 일부 마약류는 이전 연도들과 달리 2023년에 폭증하는 추세를 보이기도 함.

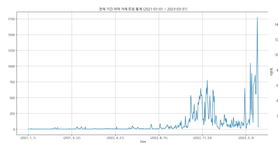

Figure 4. 마약 거래 트윗 업로드 추이 (2021.01.01~2023.03.31)

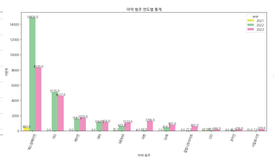

Figure 5. 마약 대분류 연도별 마약 거래 트윗 개수

Figure 6. ID 히스토그램

4. ID 히스토그램: 마약 거래 트윗 작성이 잦은 사용자 파악을 위한 분석

- 마약 거래 트윗을 많이 작성한 순으로 작성자의 트위터 ID의 빈도를 히스토그램으로 나타냄(Figure 6).

Conclusion

본 프로젝트에서는 마약 거래 글이 다량 업로드되는 트위터 데이터를 수집하여 최근 3년간의 마약 거래 동향을 파악하였고, 실시간으로 마약 거래 정황을 예측할 수 있는 모니터링 시스템을 구축함.

군집 분석 결과, 마약 거래와 관련된 토큰은 판매자 군집과 은어&지역 군집으로 나누어짐. 또 언급된 지역명의 빈도를 파악하여 시각화함으로써 마약 거래 분포 전반을 파악함. 모니터링 시스템 구축을 위해 데이터 수집부터 EDA까지의 모든 분석 과정을 자동화하여, 특정 시기에만 국한되지 않은 프로토타입을 제시함. 마약 거래 게시글 분류 모델을 모니터링 시스템과 연결하여 기간별로 수집된 트윗 데이터에 대해서 분류 할 수 있음

본 프로젝트를 통하여 수사 기관 등의 공공 기관에 마약 거래 현황 및 위치 등의 정보 실시간 제공하고 SNS가 마약 거래 수단으로 전락하는 것을 방지하며 마약에 대한 경각심 제고 및 잠재적 호기심을 사전에 차단할 수 있을 것으로 기대함.

Reference

- 검찰청 검찰활동 마약범죄수사 https://www.spo.go.kr/site/spo/02/10202030200002018100811.jsp
- 최은정 외 5인. (2021). 'SNS 빅데이터 및 검색포털 트렌드와 마약류 사건 통계간의 비교 및 의미분석 연구'