온라인 상의 마약 거래 분석 및 예측을 통한 모니터링 시스템 구현

성균관대학교 글로벌융합학부 데이터사이언스융합전공

양지인 김유진 선해라 우다연

Background

법무부에 따르면 최근 5년간 10대, 20대 마약 소비량이 약 150% 증가하면서 젊은 층을 대상으로 한 마약 확산이 빠르게 이루어지고 있으며, 과거 대포통장이나 차명계좌를 통해 오프라인으로 마약을 거래하던 방식에서, 현재는 트위터, 텔레그램 등의 소셜네트워크서비스(이하 SNS)에서 마약이 활발하게 유통되는 방식으로 변화함.
이처럼 마약의 주 거래처가 온라인 공간에서 이루어지면서, 누구나 어디서든 쉽게 마약 거래를 할 수 있게 되었음. 인터넷의 익명성과 비대면이라는 특성으로 인해 접근성이 높아진 것에 비해 수사와 검거는 더 어려워진 추세임. 마약 거래자들이 SNS에 마약 판매 관련 게시글을 올리며 텔레그램 등을 통해 실제 거래로 유도한다는 것을 고려할 때, SNS상에서 마약 거래가 어떤 방식으로 노출되고 있는지 파악하는 것이 마약 수사에 필수적임.

따라서, 본 프로젝트에서는 젊은 층을 대상으로 마약 거래가 주로 이루어지는 트위터의 2021년 1월부터 2023년 3월까지 최근 3년간의 트윗을 수집하여 다음과 같은 분석 목표를 정함. 우선, 한글 은어 외에도 영문 은어를 파악해보는 등 SNS를 이용한 **마약 거래의 최근 경향성을 파악**해보고자 함. 또한, **마약 거래 게시글을 판별하는 모델을 구축**하는 것을 목표로 함. 최종적으로 트위터에 업로드 되는 마약 거래 게시글의 내용과 위치 등의 정보와 마약 거래의 동향을 한 눈에 보여주는 **실시간 모니터링 시스템을 구축**하고자 함.

Data

수집 정보
snscrape로 트윗 수집
2021.01.01~2023.03.31
검찰청 마약 분류, 선행 연구, 인터넷 기사, 실제 트윗 참고하여 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
   - 텔레그램 아이디 언급된 경우 우선 보존
   - 사회 이슈 관련된 트윗 제거
   - 일상적으로 활용되는 단어 제거
   - 성적인 목적의 트윗 제거
   - 하나씩 살펴보면서 제거
2. 텍스트 전처리
   1) 이모지, 특수문자 제거
      - replace() 함수 활용
      - 마침표, 쉼표, 콜론은 판매 정보와 관련되므로 보존
      - url 삭제
   2) 자모음, 대소문자 통합
      - hangul-utils 활용
      - lower() 함수 활용
   3) 아이디 띄어쓰기 제거
      - 정규표현식 활용
3. mecab 형태소 분석기로 토큰화

Analysis & Methods

Our Research questions:

1. **EDA 결과를 정제해서 한 곳에 모아볼 수 없을까?**
   1) 2021-2023년에 사용된 한글/영문 은어를 더 정확하게 추출하기 → **클러스터링 기반 워드클라우드**
   클러스터링을 통해, 워드클라우드에 '마약 거래'와 관련성이 높은 단어들만 나타내고자, 다음과 같은 과정을 거침.
   - mecab 텍스트 토큰들 중 'NNP(일반명사)', 'NNG(고유명사)', 'SL(외국어)'을 Word2Vec 분석을 거쳐 각 단어들 간의 유사도를 파악함.
   - PCA 주성분 분석으로 차원을 축소함.
   - k-means clustering의 k=5로 설정해서 마약과 직접적인 관련이 있는 군집 1개를 추출함.
   - 해당 군집 내에서 다시 최적 k=2로 설정하여 결과를 시각화 함.

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

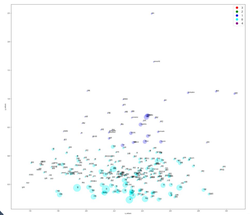

Figure 8. 클러스터링 결과

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

   2) EDA에서 추출한 **위치 데이터**(위도, 경도 데이터)를 지도에 **시각화** (Figure 10 지도)

2. **인공지능 모델이 마약 거래 트윗을 실시간으로 추적하고, 자동으로 분류 해줄 수 없을까?**
   1) 크롤링 자동화 및 DB 저장
   snscrape으로 트위터 데이터를 크롤링하는 파이썬 코드를 Linux crontab을 사용해 일정시간마다 자동으로 실행되게 스케줄링 함. 크롤링한 데이터는 goorm에서 구축한 DB 서버의 MySQL database에 데이터 저장함.

   2) 텍스트 classification 모델링
   트랜스포머 모델 및 학습 스크립트를 제공하는, Huggingface의 'albert-kor-base' pre-trained model을 transfer learning 시킴. best train, validation accuracy를 보였던 epoch에서의 모델을 우리의 최종적인 classification 모델로 선정함.

Prototype Implementation

streamlit 활용 마약 거래 모니터링 시스템 구현

1. 마약 거래 동향

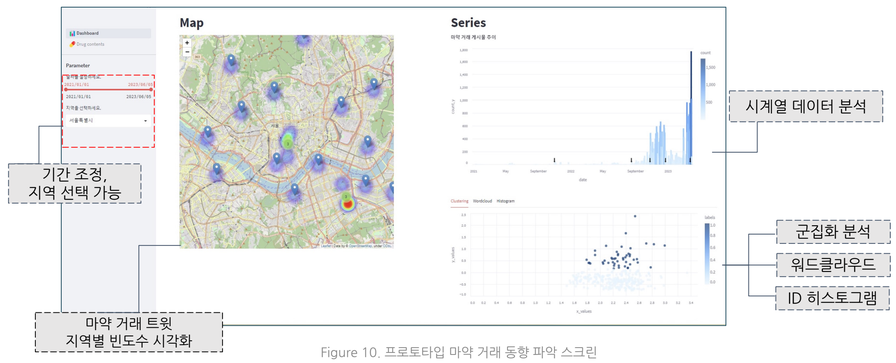

Figure 10. 프로토타입 마약 거래 동향 파악 스크린

2. 마약 거래 게시물 분류 결과

Figure 11. 프로토타입 마약 거래 게시물 분류 결과 스크린

EDA

트위터에서 이루어지는 마약 거래 동향 파악을 위해 NLP 분석, 위치 및 시계열 데이터 분석 등의 EDA를 진행함.

1. NLP 분석: 트윗 내용 자체에 대한 분석
- '허브' 등 마약을 뜻하는 총 52개의 키워드를 검색해 수집한 후, 전처리 및 필터링을 거친 40,970 행의 데이터 활용함.
- Mecab 형태소 품사 태깅에 따르면 NNG(일반명사), NNP(고유명사), SL(외국어) 순으로 많았음.

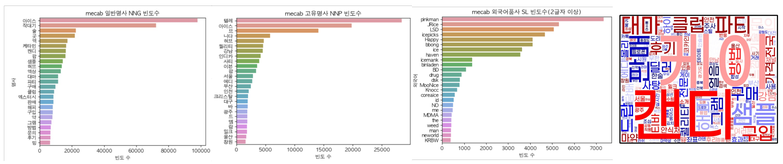

Figure 1. 일반명사, 고유명사, 외국어 품사 각각에 대하여 많이 사용된 단어 빈도를 나타낸 히스토그램

Figure 2. 워드 클라우드

- 일반명사의 경우, '아이스', '작대기' 등 마약 은어가 많았으며, 고유명사에선 '텔레'와 같은 마약 거래 수단, 혹은 '강남' 등의 지역명이 빈번하게 등장. 외국어 중에선 'pinkman' 등의 마약 판매자 메신저 아이디가 많았음 (Figure 1).
- 히스토그램에서 검색 키워드를 제외한 단어들의 빈도로 워드클라우드 (Figure 2)를 형성하여 가시적인 효과를 높임.
- '케이', '캔디' 등 마약 은어가 가장 많았으며, '판매', '샘플' 등 직접적으로 구매와 관련된 단어들도 빈번하게 등장함.

2. 위치 데이터 분석: 마약 거래가 이루어지는 지역 파악
- '울산아이스' 등 지역명을 포함하고 있는 데이터에서 지역명을 추출했고, '전국행정동리스트' 데이터를 이용해 모든 지역명을 리스트화 한 후, 마약 키워드 대분류를 기준으로 각각의 지역명 빈도를 시각화 함.
- 지역명 언급량이 가장 많았던 키워드 메스암페타민과 MDMA, 대마 모두 '강남', '서울', '부산'이 가장 많이 언급되었고, 그 외 '울산', '수원', '용인', '성남'도 언급량이 많았음 (Figure 3).

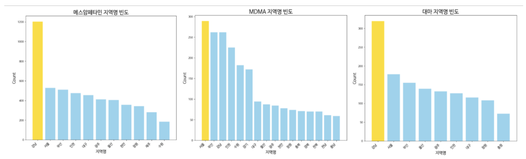

Figure 3.
메스암페타민, MDMA, 대마
마약 대분류별 언급된 지역명 빈도

3. 시계열 데이터 분석: 마약 거래 트윗이 업로드 되는 추이 파악을 위한 분석
- 2건 업로드 되었던 2021년 1월 1일과 비교할 때 일 1,750건 이상의 트윗이 업로드 되는 2023년의 증가 추세가 돋보임 (Figure 4).
- 2023년 데이터가 3월까지로 국한된 것을 고려하면 모든 마약 대분류에 있어 마약 거래 수치가 2022년부터 급증했으며, 아편, 알킬니트라이트의 경우 이전 연도들과 달리 2023년에 마약 거래가 폭증하였음 (Figure 5).
- 최근 SNS를 중심으로 하는 마약 거래가 늘고 있음을 파악할 수 있으며, 최근 동향에 따른 연구의 필요성이 강조됨.

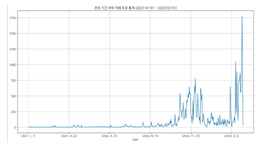

Figure 4. 마약 거래 트윗 업로드 추이
(2021.01.01~2023.03.31)

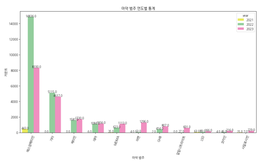

Figure 5. 마약 대분류 연도별 마약 거래 트윗 개수

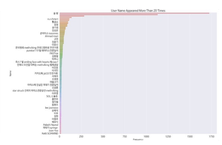

Figure 6. ID 히스토그램

4. ID 히스토그램: 마약 거래 트윗 작성량이 많은 사용자 빈도를 파악하기 위한 분석
- 마약 거래 트윗을 많이 작성한 순으로 작성자의 트위터 ID의 빈도를 히스토그램으로 나타냄 (Figure 6).
- 마약 거래 확산 방지를 위해 우선적으로 추적할 사용자를 정하는 데 도움이 될 것으로 기대할 수 있음.

Conclusion

본 프로젝트에서는 마약 거래를 유도하는 게시글이 많이 업로드되는 트위터 데이터를 수집하여 최근 3년간의 마약 거래 동향을 새롭게 파악하였고, 나아가 실시간으로 마약 거래 정황을 예측할 수 있는 모니터링 시스템을 구축함.

- 군집 분석 결과, 마약 거래와 관련된 토큰은 판매자 관련 군집과 은어 및 지역 군집 2개로 나누어짐. 판매자 군집에서 아이디를 포함한 영문 단어들도 새롭게 발견함. 또한, 2차 클러스터링을 통해 마약 군집만을 추출하여 분석을 진행하였기 때문에 마약 거래와 관련성이 낮은 '섹트', '만남' 등의 성적 단어들은 제외할 수 있었음.
- 마약 거래 게시글에서 언급된 지역명의 빈도를 파악하여, 지도에 시각화함으로써 전반적인 마약 거래 분포를 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 제공
- SNS가 마약 거래 수단으로 전락하는 것을 방지
- 마약에 대한 사람들의 경각심 제고 및 잠재적 마약 관련 범죄 동기나 호기심을 사전에 차단할 수 있을 것으로 기대함.

Reference

- "'요즘 마약거래'어떻게 이뤄지나… 국내 마약 실태는?",이경욱, 국토일보, 2022.12.20, https://www.ikld.kr/news/articleView.html?idxno=266838
- '1020마약, 5년새 150% 증가…"상황 심각, 예방교육 확대", 뉴시스, 2023.01.29, https://www.newsis.com/view/?id=NISX20230129_0002172993
- 검찰청 검찰활동 마약범죄수사 https://www.spo.go.kr/site/spo/02/10202030200000201810081 1.jsp
- 최은정 외 5인. (2021). 'SNS 빅데이터 및 검색포털 트렌드와 마약류 사건 통계간의 비교 및 의미분석 연구'