JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 29 No. 10, pp. 207-218, October 2024
https://doi.org/10.9708/jksci.2024.29.10.207

# Development of a Model for Identifying Drug Organizations and Their Scale through Tweet Clustering

Jin-Gyeong Kim*, Eun-Young Park**, Da-Sol Kim**, Cho-Won Kim**, Jiyeon Kim***

*Graduate Student, Dept. of Computer Information Engineering, Graduate School, Daegu University, Korea
**Undergraduate Student, Dept. of Computer Engineering, Daegu University, Korea
***Professor, Dept. of Computer Engineering, Daegu University, Gyeongsan, Korea

## [Abstract]

In this paper, we propose a model for identifying drug trafficking organizations and assessing their scale by collecting drug promotional tweets from the social media platform 'X,' with a focus on investigating drug crimes that frequently occur among teenagers and young adults. Recently, various cyber crimes, such as drug distribution, illegal gambling, and sex offense, have been on the rise, exploiting the anonymity provided by social media. Drug trafficking organizations, in particular, operate in a decentralized cell structure, where each member receives anonymous instructions regarding only their specific role and is not directly connected to other members. To track these types of crimes, we designed experimental scenarios using various clustering algorithms, such as K-means Clustering and Spectral Clustering, alongside text embedding models like BERT (Bidirectional Encoder Representations from Transformers) and GloVe (Global Vectors for Word Representation). Furthermore, the clustering results derived from each scenario are validated using Jaccard Similarity and a full-scale investigation. We then analyze tweet clusters identified as the same drug organization across all scenarios, prioritizing the identification of high-priority accounts for cyber investigations.

▸ **Key words**: Cyber Investigation, Drug, Social Media, Tweet, Clustering

## [요 약]

본 논문은 10대와 청년층에서 빈번하게 발생하는 마약 범죄를 수사하기 위해 소셜미디어 플랫폼 'X'에서 마약 홍보 트윗을 수집하고, 이를 바탕으로 마약 유통 조직 및 규모를 식별하는 클러스터링 모델을 개발하는 것을 목표로 한다. 최근 소셜미디어의 익명성을 악용한 마약, 불법 도박, 성범죄 등 다양한 사이버 범죄가 증가하고 있으며, 특히 마약 유통 조직은 각 구성원이 자신의 역할에 대해서만 익명으로 지시를 받고, 다른 구성원들과 직접 연결되지 않은 점조직 형태로 운영되고 있다. 이러한 유형의 범죄를 추적하기 위해 BERT(Bidirectional Encoder Representations from Transformers), GloVe(Global Vectors for Word Representation)와 같은 텍스트 임베딩 모델 및 K-means Clustering과 Spectral Clustering 등 다양한 클러스터링 알고리즘을 활용하여 실험 시나리오를 설계하였다. 또한, 각 시나리오에서 도출된 클러스터링 결과를 자카드 유사도(Jaccard Similarity) 및 전수조사 기반으로 검증하고, 모든 시나리오에서 동일한 마약 조직으로 식별된 트윗 클러스터를 분석하여 사이버 수사 시, 추적 우선순위가 높은 계정을 식별한다.

▸ **주제어**: 사이버 수사, 마약, 소셜미디어, 트윗, 클러스터링

• First Author: Jin-Gyeong Kim, Corresponding Author: Jiyeon Kim
*Jin-Gyeong Kim (wlsrud1470@daegu.ac.kr), Dept. of Computer Information Engineering, Graduate School, Daegu University
**Eun-Young Park (pey6693@daegu.ac.kr), Dept. of Computer Engineering, Daegu University
**Da-Sol Kim (dasol0803@daegu.ac.kr), Dept. of Computer Engineering, Daegu University
**Cho-Won Kim (kcw37@daegu.ac.kr), Dept. of Computer Engineering, Daegu University
***Jiyeon Kim (jyk@daegu.ac.kr), Dept. of Computer Engineering, Daegu University
• Received: 2024. 09. 09, Revised: 2024. 09. 23, Accepted: 2024. 09. 30.


http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

## I. Introduction

페이스북, 'X(구 트위터)', 인스타그램 등과 같은 소셜미디어(Social Media)는 사용자들이 실시간으로 다양한 콘텐츠를 공유할 수 있게 하는 디지털 플랫폼으로서 전 세계의 약 50억 명 이상이 소셜미디어를 사용하고 있다[1]. 소셜미디어는 실시간 소통 및 다양한 콘텐츠 창작을 지원하고, 익명성 기반으로 자유로운 의사 표현 기회를 제공하는 장점이 있지만, 마약, 불법 도박, 성범죄 등과 같은 다양한 범죄의 수단으로 악용되는 사례도 증가하고 있다. 나아가 마약 범죄에 연루된 사람 중 대다수가 10-20대의 연령대로 나타났으며, 큰돈을 벌 수 있다는 유혹과, 마약에 보다 쉽게 접근할 수 있다는 점에 빠져 마약 밀수 및 거래에 가담하게 된다. 이는 향후 마약 투약자의 환각 운전이나 살인과 같은 사회 전체에 큰 손실을 초래하는 2차 피해로 이어질 수 있다[2]. 2024년에는 해외에서 마약을 밀반입하고 소셜미디어를 통해 이를 판매한 조직이 검거되었으며, 이들 중 판매책 27명은 11개의 텔레그램 채널을 통해 16명의 구매자와 투약자에게 마약류를 거래 및 판매하였다[3]. 또 다른 사건으로는 텔레그램 채널을 통해 마약류를 밀수입한 후 합성 마약을 제작하고 유통한 조직도 검거된 사례가 있다[4]. 이 사건들은 모두 최소한의 정보만을 공유하는 점조직 형태로 운영되었으며, 수사망을 피하기 위해 가상화폐를 사용하거나 텔레그램과 같은 익명성이 보장된 인스턴트 메신저를 통해 판매자와 접촉하는 방식으로 이루어졌다.

이러한 형태의 마약 범죄 조직을 검거하기 위해서는 신분을 위장하여 판매자에게 접근하는 위장 수사가 활용된다. 독일에서는 마약, 조직범죄, 아동 음란물 제작 및 배포 등과 관련된 범죄 수사에서 신분 위장 수사가 허용되며, 미국에서도 부패, 테러, 마약 범죄를 대상으로 한 가이드라인에 따라 위장 수사가 일반적인 수사 기법으로 효과적으로 활용[5]되고 있는 반면, 국내의 경우 아동·청소년 대상 디지털 성범죄 수사에만 위장 수사를 적용할 수 있다. 또한, 마약 범죄 수사에서는 경찰 신분을 밝히지 않고 접근하는 정도의 위장 수사만이 인정되어, 적극적인 선제 조치 및 신속한 검거에 어려움이 존재한다. 과거의 마약 범죄는 대면을 통해 직접 거래하는 방식이 주를 이루었으나, 현재의 소셜미디어를 통한 범죄 홍보 및 비대면 거래와 같은 방식에 대응하기 위해서는 변화된 마약 유통 및 거래 방식에 따른 수사가 필요한 실정이다. 최근 소셜미디어를 통한 마약 범죄 홍보는 하나의 조직이 단일 인스턴트 메신저 ID가 아닌 다수의 인스턴트 메신저 ID를 사용하여 거래를 진행하고 있으며, 이에 따라 수사관들은 이러한 점조직의 복잡한 구조와 다중 채널 활용을 파악하고 대응하기 위해 효과적인 수사 기법과 전략적 접근 기술 개발이 필수적이다. 본 연구에서는 소셜미디어 'X'에서 발생하는 마약 범죄를 대상으로, 수집된 트윗 데이터에서 거래를 위한 수단으로 사용되는 인스턴트 메신저 ID를 추출하고, 클러스터링 모델을 활용하여 마약 범죄 조직 및 규모를 식별하는 모델 개발을 목표로 한다. 이후, 제안된 시나리오별 클러스터링 모델을 통해 도출된 결과를 바탕으로 유사도 분석을 진행하여, 해당 모델의 유효성을 검증한다.

본 논문의 구성은 다음과 같다. 2장에서는 소셜미디어에서의 마약 범죄 수사와 클러스터링 모델을 활용한 다양한 연구를 살펴본다. 3장에서는 범죄 조직 및 규모를 식별하는 모델을 설계하고, 본 연구에서 제안하는 텍스트 임베딩 모델과 클러스터링 모델을 설명한다. 4장에서는 3장에서 제안한 모델을 기반으로 'X'에서 수집한 마약 범죄 트윗을 클러스터링하여 범죄 조직 및 규모를 식별한 결과를 설명한다. 5장에서는 동일 마약 조직으로 식별된 트윗 클러스터를 전수 검증하여 제안된 모델의 성능 및 수사 기술로서의 유효성을 검증한다. 마지막으로, 6장에서는 결론과 향후 연구 방향을 제시한다.

## II. Preliminaries

### 2.1 Cyber Investigation Studies on Social Media

소셜미디어에서는 마약 거래, 불법 도박, 성 착취물 유포 등 다양한 사이버 범죄가 발생하고 있으며 이러한 범죄는 사회에 여러 가지 악영향을 미치고 있다. 따라서, 이러한 범죄를 추적하기 위한 다양한 수사 기법들이 연구되고 있으며, 소셜미디어 상의 범죄를 탐지하고 패턴을 분석 및 예측하기 위한 연구들은 주로 텍스트 마이닝 및 머신러닝(Machine Learning) 기반으로 수행되었다.

먼저, 소셜미디어 상의 범죄를 탐지하는 연구로는 마약류 관련 단어의 언급 빈도 및 트렌드 데이터를 분석하거나[6], 소셜미디어와 신문 등에서 수집된 데이터를 통해 범죄 패턴을 식별하는 연구[7], N-gram 및 머신러닝을 통해 언어적 패턴을 식별하고[8], SVM(Support Vector Machine)과 CNN(Convolutional Neural Network)을 활용해 불법 마약 광고를 탐지하는 연구[9], 멀티모달 분석 기법을 통해 텍스트와 이미지 데이터를 결합하는 마약 거래를 탐지하거나[10], LLM(Large Language Model) 기반의 탐지 프레임워크를 제안하는 연구[11], 하이퍼그래프

대조 학습(Hypergraph Contrastive Learning)기법을 통해 마약 밀매 커뮤니티를 효과적으로 탐지하는 모니터링 연구들이 수행되었다[12].

범죄 예측 및 패턴 분석을 위한 연구로는 Decision tree, KNN(K-Nearest Neighbors), Naive Bayes classifier, SVM(Support Vector Machine) 등을 활용하여 범죄 발생을 예측하는 연구[13], 음이항 회귀 모델과 다수준 회귀 모델을 통해 약물 관련 트윗과 범죄 데이터와의 관계를 분석한 범죄 예측 연구[14], SNA(Social Network Analysis)를 활용하여 마약 관련 범죄 네트워크를 분석하고 재범 가능성을 예측하는 모델 연구[15], 범죄 현장 추적 및 시각화를 위해 범죄 사건의 시공간적 특성을 시각화하고, 트윗을 워드 클라우드 형태로 나타내어 수사에 도움을 주는 연구[16], 인스타그램에서 마약 판매 계정을 추적하기 위해 프로필 및 게시물의 유사도를 분석하는 연구들이 진행되었다[17].

또한, 텍스트마이닝 및 머신러닝 외에도 소셜미디어 상의 범죄 예방을 위한 정책적 연구도 진행되었다. 인터넷과 소셜미디어를 통한 마약 거래를 효과적으로 제한하기 위한 법적 개선 방안을 제시하거나[18], 청소년 마약 범죄 증가 추세를 파악하고 성인 범죄자로 발전하는 것을 예방하기 위한 연구들도 함께 수행되었다[19].

### 2.2 Clustering-based Data Analysis Studies

클러스터링(Clustering)은 서로 유사한 특성을 가진 데이터를 집합으로 묶는 과정으로서 데이터 패턴 분석을 통해 유사한 데이터 간의 관계를 도출하는 데에 활용된다.

클러스터링은 범죄 수사 외에도 다양한 분야에서 활용되고 있으며 GNN(Graph Neural Network)에서 그래프 풀링을 개선하기 위해 Spectral Clustering을 활용하는 연구[20], 학생들의 질문 유형을 분석하기 위해 문서 클러스터링을 활용한 연구[21], K-Means Clustering을 활용한 알약 분류에 도움을 주는 연구[22], 대중교통 이용자 통행 패턴 생성을 위해 K-Means Clustering을 활용한 연구[23], 계층 클러스터링을 사용해 실시간 데이터를 이용한 충돌 위험 평가에 대한 연구[24], K-Means Clustering, Hierarchical Clustering, DBSCAN Clustering 기법을 활용한 도로 링크별 취약성 평가에 관한 연구[25]들이 진행되고 있다.

범죄 수사에서 클러스터링이 활용된 연구로는 K-Means Clustering을 통해 범죄 발생 가능성이 높은 지역을 식별하고, 다양한 머신러닝 알고리즘을 적용하여 범죄 예측 모델의 정확도를 개선하는 연구[26], K-Means Clustering, DBSCAN Clustering 및 계층적 군집화 등의 방법론을 통한 공간 클러스터링 기반의 범죄 예측 알고리즘 개선 연구[27], K-Means 및 Agglomerative Clustering과 같은 알고리즘을 적용하여 특정 범죄 패턴 및 위치를 시각화하는 연구[28], 범죄 커뮤니티와 관련된 연구로는 갱단 구성원 간의 지리적 데이터와 사회적 데이터를 결합하여 Spectral Clustering을 통해 갱단 커뮤니티를 식별하는 연구[29], Bi-Spectral Clustering을 활용해 사용자의 타임라인을 분석하고, 이를 통해 행동과 인식을 파악하여 커뮤니케이션 패턴을 이해하는 연구[30] 등이 범죄 수사를 위해 진행되고 있다.

소셜미디어에서 텍스트 마이닝 및 머신러닝 기반으로 범죄를 분석한 기존 연구들은 주로 키워드 중심으로 범죄 발생 양상 및 패턴 분석을 하였고, 클러스터링 기반의 기존 연구들은 범죄 발생 시간 및 지역과 같은 사회적 범죄 데이터를 활용하여 범죄 유형 및 패턴을 식별하는 연구를 수행하였다. 반면, 본 논문은 국내 마약 범죄의 진화하는 양상에 맞추어 소셜미디어 범죄 조직이 다수의 인스턴트 메신저를 거래 수단으로 사용하는 특성에 따라 범죄 트윗을 수집하고, 수집된 트윗 데이터에서 인스턴트 메신저 ID를 추출하여 범죄 조직 및 규모를 식별하는 연구를 수행한다는 점에서 차별화된다.

## III. The Proposed Scheme

본 장에서는 소셜미디어 'X'에서 수집된 트윗 중, 인스턴트 메신저 ID 정보를 포함한 트윗을 활용하여 범죄 조직 및 규모를 식별하는 모델을 설계한다. 본 논문에서 제안하는 모델은 Fig. 1과 같은 과정을 통해 개발된다.

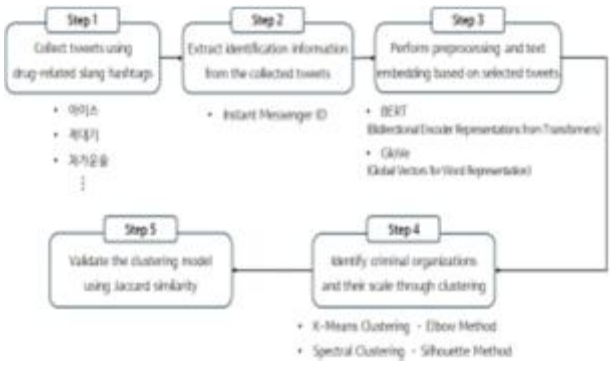


Fig. 1. A Procedure of developing a model for identifying Criminal Organizations and their Scale

Step 1에서는 마약 용어 해시태그인 '작대기', '아이스', '차가운술'을 검색 키워드로 사용하여 해당 키워드로 검색된 각 트윗에서 48개의 속성(트윗 ID, 작성자, 내용, 트윗 생성 시간, 해시태그 등)을 포함한 트윗 데이터를 수집하는 단계로 본 논문에서는 2024년 5월 13일부터 5월 20일, 6월 8일부터 6월 15일까지 수집된 총 16,232개의 트윗을 수집하였다. Step 2는 동일 마약 조직 식별을 위한 식별 정보를 포함하고 있는 트윗을 선별하는 단계이다. 본 논문에서는 동일한 인스턴트 메신저 ID가 포함된 트윗을 동일 마약 조직이 게시한 트윗으로 판단하며 16,232개의 트윗 중, 인스턴트 메신저 ID를 포함하고 있는 14,890개의 트윗을 선별하였다. 선별된 트윗을 분석한 결과, 14,890개의 트윗을 단 15개의 계정이 업로드한 것으로 분석되었다. Table 1은 수집된 트윗에서 추출한 인스턴트 메신저 ID를 보여준다.

Table 1. Instant Messenger IDs extracted from tweets

| ID | Instant Messenger ID | ID | Instant Messenger ID |
|---|---|---|---|
| $U_1$ | A*******0 | $U_9$ | b**********9 |
| $U_2$ | E*****0 | $U_{10}$ | c******e |
| $U_3$ | H********g | $U_{11}$ | c****2 |
| $U_4$ | J****1 | $U_{12}$ | d**********9 |
| $U_5$ | K****4 | $U_{13}$ | i******n |
| $U_6$ | K****8 | $U_{14}$ | i******k |
| $U_7$ | M********7 | $U_{15}$ | i******s |
| $U_8$ | b********1 | - | |

Step 3에서는 텍스트 임베딩을 효과적으로 수행하기 위해 전처리 후 텍스트 임베딩을 수행하는 단계이다. 먼저, 텍스트 임베딩에 방해가 되는 요소인 특수 문자와 이모티콘을 제거하고, 영어로 작성된 모든 문자를 소문자로 변환하는 등 트윗 홍보 게시물의 특성을 고려하여 전처리를 진행하였다. 다음으로는 전처리 된 트윗을 기반으로 기계 학습 모델이 이해할 수 있는 수치 벡터로 변환하는 과정인 텍스트 임베딩을 수행하였다. 본 연구에서는 트윗을 효과적으로 임베딩하기 위해 두 가지 텍스트 임베딩 모델을 사용하여 비교 분석을 진행하였다. 먼저, 문맥을 고려하여 단어를 임베딩하고 문서의 특징을 자동으로 추출할 수 있는 BERT(Bidirectional Encoder Representations from Transformers) 모델과 전역적인 통계 정보를 기반으로 단어를 벡터로 표현하여 단어 간의 의미적 관계를 포착하는 GloVe(Global Vectors for Word Representation) 모델을 사용하여 트윗의 텍스트 데이터를 임베딩하였다.

Step 4에서는 Step 3에서 추출된 각 모델의 텍스트 임베딩 데이터를 기반으로 최적의 클러스터 수를 결정하고 클러스터링을 수행하는 단계이다. 본 논문에서는 효과적인 범죄 조직 및 규모 식별을 위해 두 가지 클러스터링 모델을 사용하였다. 첫 번째로, K-Means Clustering 모델의 비지도 학습 알고리즘으로 주어진 데이터 집합을 $k$개의 클러스터로 분할하는 방법이다. 해당 알고리즘은 클러스터 내의 데이터 포인트와 클러스터 중심 간의 거리를 최소화하여 클러스터를 형성한다[31]. 또한, 데이터 포인트 간의 거리를 기반으로 작동하기 때문에 복잡한 연산을 요구하지 않지만, 클러스터 개수를 미리 지정해야 하므로 데이터의 특징을 잘 모를 경우, 분석에 어려움이 있을 수 있다. 따라서, 본 연구에서는 K-Means Clustering의 최적 클러스터 수를 결정하기 위해 엘보우 기법(Elbow Method)을 사용하였다. 엘보우 기법은 전체적인 응집도를 고려하여 클러스터 내 데이터의 응집도를 평가하기 때문에 적절한 클러스터 수를 판단하는 데에 유용하며 특히 K-Means Clustering과 같은 거리 기반 클러스터링 알고리즘에서 효과적으로 사용할 수 있다. 두 번째 클러스터링 모델로 본 논문에서는 Spectral Clustering 모델을 사용하였다. Spectral Clustering은 그래프 이론에 기반한 방법으로 고유벡터를 사용해 데이터를 저차원 공간으로 매핑한 후, 이를 전통적인 클러스터링 방법(K-Means Clustering, Hierarchical Clustering 등)을 통해 클러스터링하는 방식이다[32]. 이는 복잡한 데이터 구조를 다루는 데 효과적이며, 개별 데이터 포인트 간의 관계뿐 아니라, 데이터 집합 전체에서 어떻게 클러스터링 될 수 있는지를 반영한다. Spectral Clustering의 최종 단계에서는 클러스터 수를 미리 지정해야 하며, 고유벡터 과정에서 계산 비용이 높아 대규모 데이터셋에 적용하기에는 어려움이 있을 수 있다. 본 연구에서는 Spectral Clustering의 결과를 평가하기 위해 실루엣 기법(Silhouette Method)을 적용하였다. 실루엣 기법은 클러스터 내 데이터의 밀집도와 클러스터 간 분리도를 함께 고려하여 클러스터링 결과의 품질을 직관적이고 포괄적으로 평가할 수 있으며 Spectral Clustering이 데이터의 구조를 얼마나 잘 반영했는지를 판단할 수 있다. Step 5는 Step 4에서 진행한 클러스터링 결과를 기반으로 클러스터 내 트윗들에 대한 유사도 분석을 통해 모델의 유효성을 검증하는 단계이다. 이 단계에서는 먼저 각 클러스터 내 트윗들의 유사성을 자카드 기반으로 분석하여 트윗 간의 유사도를 판단한다. 또한, 클러스터 내 트윗들의 해시태그, 키워드 등을 전수 조사하여 조직의 고유 식별자를 비교함으로써 클러스터링 결과에 대한 유효성을 검증한다.

## IV. Experimental Results

본 장에서는 3장에서 제안한 모델을 기반으로 인스턴트 메신저 ID를 중심으로 선별된 14,890개의 트윗에 대해 텍스트 임베딩, 클러스터 수 최적화, 클러스터링 모델을 각각 적용한 실험 시나리오를 설계하고, 설계된 시나리오 기반으로 클러스터링을 수행한다. 본 논문에서 제안하는 실험 시나리오는 Table 2과 같다.

Table 2. 4 Types of Clustering Scenarios

| Scenario | Text Embedding Model | Optimization Model | Clustering Model |
|---|---|---|---|
| 1 | BERT | Elbow | K-means Clustering |
| 2 | GloVe | Elbow | K-means Clustering |
| 3 | BERT | Silhouette | Spectral Clustering |
| 4 | GloVe | Silhouette | Spectral Clustering |

### 4.1. Scenario 1

시나리오 1에서는 BERT로 임베딩된 텍스트 데이터를 엘보우 기법으로 분석하여 6개의 최적 클러스터 수를 도출하였고, 이를 K-means Clustering 모델에 적용하여 Fig. 2와 같은 클러스터링 결과를 도출하였다.

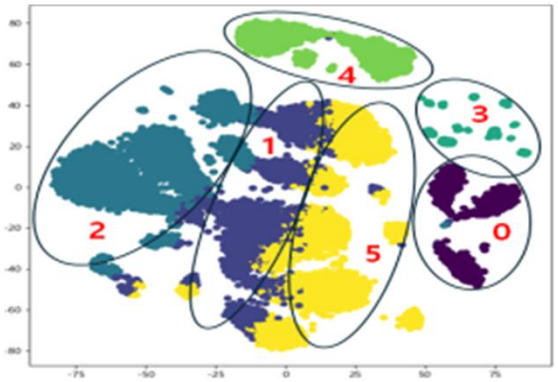


Fig. 2. Clustering results of Scenario 1

Fig. 2의 각 클러스터에 포함된 트윗의 인스턴트 메신저 ID는 Table 3과 같다.

Table 3. Distribution of Instant Messenger IDs by Cluster of Scenario 1

| Cluster | Instant Messenger ID |
|---|---|
| Cluster 0 | $U_1$, $U_2$, $U_5$, $U_6$ |
| Cluster 1 | $U_3$, $U_4$, $U_7$, $U_8$, $U_9$, $U_{10}$, $U_{13}$, $U_{14}$, $U_{15}$ |
| Cluster 2 | $U_3$, $U_4$, $U_5$, $U_7$, $U_8$, $U_{10}$, $U_{13}$, $U_{14}$, $U_{15}$ |
| Cluster 3 | $U_8$, $U_{11}$, $U_{14}$ |
| Cluster 4 | $U_3$, $U_{14}$ |
| Cluster 5 | $U_3$, $U_4$, $U_5$, $U_7$, $U_8$, $U_{10}$, $U_{12}$, $U_{13}$, $U_{14}$, $U_{15}$ |

$U_3$, $U_4$, $U_5$, $U_7$, $U_8$, $U_{10}$, $U_{13}$, $U_{14}$, $U_{15}$가 클러스터 2와 클러스터 5에 모두 속해 있음을 알 수 있다. 클러스터 2에 속한 트윗들은 '빠른 거래', '안전한 거래 보장'과 같은 마약 거래 홍보 메시지와 다양한 마약 용어를 사용하는 특징을 보였고, 클러스터 5에 속한 트윗들은 '대구아이스'와 같은 특정 지역 정보를 포함한 마약 거래와 '허브', '사끼' 등의 마약 용어를 사용하는 특징을 보였다. 이때, 클러스터 2에 속한 9개의 인스턴트 메신저 ID는 '작대기', '떨판매'와 같은 간결한 문구를 사용하여 마약 거래를 강조하면서 구매를 유도하는 내용을 주로 포함하였고, 클러스터 5에 속한 9개의 인스턴트 메신저 ID는 '전국 배송 가능', '샘플 제공' 등 구체적인 거래 정보와 함께 구매를 유도하는 내용이 포함된 것을 확인하였다. 이 밖에도 클러스터 0은 '사칭주의', '안전한 거래'와 같은 신뢰성을 강조하는 문구를 사용하여 구매자들의 신뢰를 확보하려는 트윗들로 구성된 특징이 있으며, 클러스터 1은 주로 '대구아이스'와 같은 마약 거래를 위한 특정 지역 정보와 '샘플', '20만'과 같은 거래 금액 정보를 제공하는 트윗으로 주로 구성된 것을 확인하였다. 클러스터 3에서는 대다수의 트윗 게시물이 '뽕라덴복귀'라는 특정 용어를 사용하고 있으며 주로 다양한 마약 유형을 홍보하는 트윗 내용임을 확인하였다. 클러스터 4에서는 '차가운술' 및 '작대기'가 빈번하게 언급되며 특정 마약 제품의 거래를 강조하는 트윗들로 구성된 것을 확인하였다.

### 4.2. Scenario 2

시나리오 2에서는 GloVe로 임베딩된 텍스트 데이터를 엘보우 기법으로 분석하여 4개의 최적 클러스터 수를 도출하였고, 이를 K-means Clustering 모델에 적용한 결과는 Fig. 3과 같다.

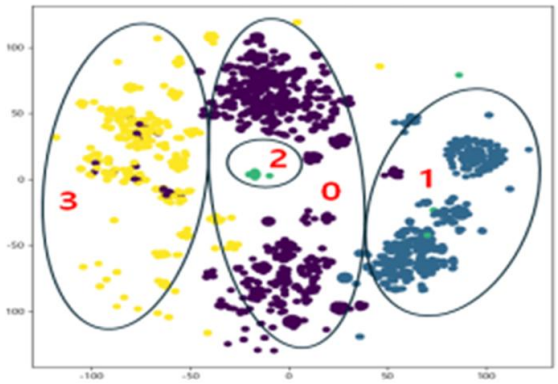


Fig. 3. Clustering results of Scenario 2

Fig. 3의 각 클러스터에 포함된 트윗의 인스턴트 메신저 ID는 Table 4와 같다.

Table 4. Distribution of Instant Messenger IDs by Cluster of Scenario 2

| Cluster | Instant Messenger ID |
|---|---|
| Cluster 0 | $U_3$, $U_4$, $U_{11}$, $U_{13}$, $U_{14}$, $U_{15}$ |
| Cluster 1 | $U_1$, $U_2$, $U_5$, $U_6$, $U_7$, $U_8$, $U_9$, $U_{10}$, $U_{12}$ |
| Cluster 2 | $U_7$, $U_8$, $U_{10}$ |
| Cluster 3 | $U_4$, $U_{14}$, $U_{15}$ |

$U_7$과 $U_8$이 클러스터 1과 클러스터 2에 모두 속해 있음을 알 수 있다. 클러스터 1에서는 '착한딜러', '사기 걱정 NO'와 같은 거래의 신뢰성을 강조하는 문구와 '아이스 0.5:40만'과 같은 특정 마약의 금액 정보를 담은 거래 중심의 트윗들이 게시되었고, 클러스터 2에서는 클러스터 1과 유사하게 다양한 마약 용어를 포함하면서도 거래의 속도와 안전성을 강조하는 문구를 포함하는 트윗이 주로 게시되었다. 이때, 클러스터 1에 속한 2개의 인스턴트 메신저 ID는 상세한 금액 정보와 거래 조건을 제공하는 것이 특징이며 클러스터 2에 속한 2개의 인스턴트 메신저 ID는 더 다양한 마약 유형과 거래의 안전성을 강조하는 트윗들을 게시한 것을 확인하였다. $U_4$, $U_{14}$, $U_{15}$는 클러스터 0과 클러스터 3에 속해 있으며 클러스터 0에는 '아이스작대기', '아이스', '사끼'와 같은 다양한 마약 용어가 포함된 트윗이 포함된 것을 확인하였다. 클러스터 3도 다양한 마약 용어를 사용하지만, 클러스터 0에는 나타나지 않은 마약 및 샘플 거래에 대한 상세한 정보가 포함된 것을 확인하였다. 이때, 클러스터 0에 속한 3개의 인스턴트 메신저 ID는 마약 거래에 집중하여 상세한 거래 정보와 제품의 품질을 강조하는 내용을 포함하였다. 반면, 클러스터 3에 속한 3개의 인스턴트 메신저 ID는 클러스터 0과 유사하게 다양한 마약 용어를 포함한 트윗을 게시하지만, 특정 마약의 거래를 강조하는 데에 초점을 맞추고 있다는 점에서는 차이가 있음을 확인하였다.

### 4.3. Scenario 3

시나리오 3에서는 BERT로 임베딩된 텍스트 데이터를 실루엣 기법으로 분석하여 4개의 최적 클러스터 수를 도출하였고, 이를 Spectral Clustering 모델에 적용한 결과는 Fig. 4와 같다.

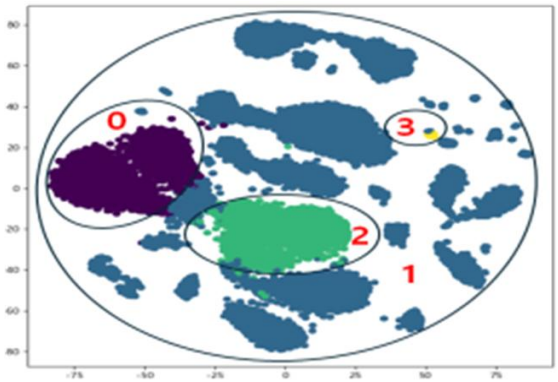


Fig. 4. Clustering results of Scenario 3

Fig. 4의 각 클러스터에 포함된 트윗의 인스턴트 메신저 ID는 Table 5와 같다.

Table 5. Distribution of Instant Messenger IDs by Cluster of Scenario 3

| Cluster | Instant Messenger ID |
|---|---|
| Cluster 0 | $U_3$, $U_4$, $U_9$, $U_{10}$, $U_{13}$, $U_{14}$, $U_{15}$ |
| Cluster 1 | $U_1$, $U_2$, $U_3$, $U_4$, $U_5$, $U_6$, $U_7$, $U_8$, $U_{10}$, $U_{11}$, $U_{12}$, $U_{13}$, $U_{14}$, $U_{15}$ |
| Cluster 2 | $U_4$, $U_{13}$, $U_{14}$, $U_{15}$ |
| Cluster 3 | $U_6$ |

클러스터 0에 속하는 트윗은 '작대기팝니다'와 같이 구매자가 특정 마약 유형과 마약 거래를 유도하는 문구를 포함하는 특징을 가지고, 클러스터 1은 다양한 마약 유형과 함께 '클럽캔디', '최음제 클럽파티'와 같은 낮은 연령층 대상의 마약 홍보글을 주로 게시하는 것을 확인하였다. 클러스터 0과 클러스터 1에 모두 속하는 $U_3$, $U_4$, $U_{10}$, $U_{13}$, $U_{14}$, $U_{15}$는 클러스터 0과 클러스터 1에서 다른 특징으로 트윗을 게시하는 특징을 보였다. 먼저, 클러스터 0에 속한 6개의 인스턴트 메신저 ID는 마약 거래를 위한 신속성 및 효율성을 강조하여 구매자의 신뢰를 얻을 수 있는 문구를

주로 사용하였고, 클러스터 1에 속한 6개의 인스턴트 메신저 ID는 간접적으로 마약 거래를 유도하거나 파티 문화와 연관 지어 마약을 홍보하는 방식으로 트윗을 게시한 것을 확인하였다. 이 밖에도 클러스터 2에서는 다양한 마약 유형과 함께 '떨느낌', '아이스사용후기'와 같이 마약의 효과 및 쓰임을 언급하는 특징을 가진 트윗들로 구성된 것을 확인하였고, 클러스터 3은 '아이디로 문의주세요!', '오실장' 등의 직급을 표현하거나 다른 트윗에서 사용되지 않는 특정 문구를 포함하는 특징을 가진 것으로 확인되었다.

### 4.4. Scenario 4

시나리오 4에서는 GloVe로 임베딩된 텍스트 데이터를 실루엣 기법으로 분석하여 9개의 최적 클러스터 수를 도출하였고, 이를 Spectral Clustering 모델에 적용한 결과는 Fig. 5와 같다.

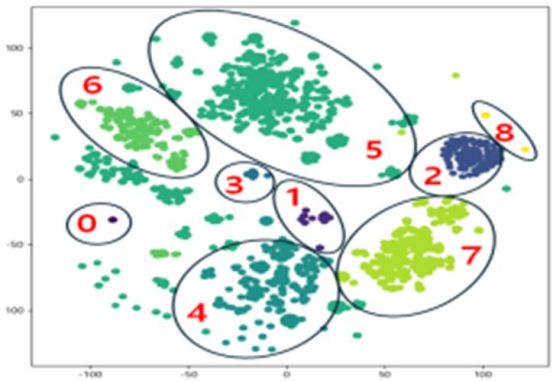


Fig. 5. Clustering results of Scenario 4

Fig. 5의 각 클러스터에 포함된 트윗의 인스턴트 메신저 ID는 Table 6과 같다.

Table 6. Distribution of Instant Messenger IDs by Cluster of Scenario 4

| Cluster | Instant Messenger ID |
|---|---|
| Cluster 0 | $U_{14}$ |
| Cluster 1 | $U_3$ |
| Cluster 2 | $U_1$, $U_8$ |
| Cluster 3 | $U_8$ |
| Cluster 4 | $U_{13}$ |
| Cluster 5 | $U_3$, $U_4$, $U_5$, $U_6$, $U_9$, $U_{10}$, $U_{11}$, $U_{13}$, $U_{14}$, $U_{15}$ |
| Cluster 6 | $U_{15}$ |
| Cluster 7 | $U_7$, $U_{10}$, $U_{12}$ |
| Cluster 8 | $U_2$, $U_6$ |

$U_3$은 클러스터 1과 클러스터 5에 모두 속해 있는 것을 볼 수 있다. 클러스터 1은 특정 마약 제품이나 사용 방식에 대해 더 구체적인 정보를 제공하는 특징을 가지고, 클러스터 5는 특정 마약의 샘플, 후기, 판매처 등을 강조하여 거래를 유도하는 트윗들이 주로 게시된 것을 확인하였다. 이때, 클러스터 1에 속한 $U_3$은 '아이스', '작대기' 등의 다양한 마약 용어가 주로 사용되며 클러스터 5에 $U_3$은 클러스터 1과 유사한 마약 용어를 포함하지만, '천안아이스'와 같이 특정 지역과 마약 용어를 함께 포함하는 트윗으로 구성된 것을 확인하였다. 또한, 클러스터 2는 '전국', '샘플', '전국 실시간 좌표'와 같은 용어를 주로 포함하고, 클러스터 3은 다양한 마약 용어를 포함하는 트윗으로 구성된 특징을 가진다. $U_8$은 클러스터 2와 클러스터 3에 모두 속한 것을 볼 수 있다. 클러스터 2에 속한 $U_8$은 거래 정보와 마약 홍보를 강조하는 특징을 가지고, 클러스터 3에 속한 $U_8$은 마약 거래뿐 아니라, 마약의 유형 및 클럽과 같은 장소를 포함하는 특징을 가진다. 이 밖에도 클러스터 0은 다양한 마약 용어들을 포함하는 트윗들로 구성되었으며, 클러스터 4는 '여러분에 안식처 하이븐'이라는 특정 문구를 포함하는 트윗들로 구성된 것을 확인하였다. 클러스터 6은 주사기를 통해 투약하는 필로폰의 은어인 '작대기얼음'을 'ㅈㄷㄱㅇㄹ'과 같이 초성으로 표현하거나, 코로 흡입하는 마약류인 '후리'를 'ㅎㄹ'로 줄여 사용하는 등 마약 용어의 변형어로 트윗을 게시한 것을 확인하였다. 또한, 클러스터 7에서는 마약 거래의 품질과 신뢰성을 강조하는 정보가 중심이 되어, '아이스 샘플: 20만'과 같은 마약 금액과 '후리하는 방법', '작대기 효능' 등 마약의 쓰임과 효과를 강조하여 구매자들의 호기심을 자극하는 특징을 보였다. 마지막으로 클러스터 8은 '오실장'과 같은 직급을 사용하거나, '사칭주의'라는 용어를 사용하여 거래의 신뢰성을 확보하려는 문구를 사용하는 특징이 있다는 것을 전수조사를 통해 확인하였다.

## V. Discussion

### 5.1. Analysis of Jaccard Similarity-based Clustering Effectiveness

본 절에서는 클러스터링 시나리오 기반의 범죄 조직 및 규모 식별 모델의 유효성을 검증하기 위하여 자카드 유사도(Jaccard Similarity)를 기반으로 시나리오 1부터 4까지 각 클러스터 내 트윗들의 평균 유사도를 계산하였다. 자카

드 유사도는 두 개의 집합 사이의 유사성을 측정하는 통계적 방법으로서 두 집합의 교집합 크기를 합집합 크기로 나눈 값으로 계산된다. 즉, 두 집합이 얼마나 유사한지를 0과 1 사이의 값으로 나타내며 1에 가까울수록 두 집합의 유사성이 높고, 0에 가까울수록 유사성이 낮다는 것을 의미한다. 시나리오별 클러스터 내 트윗의 자카드 유사도 평균 결과는 Table 7과 같다.

Table 7. Average Jaccard Similarity Results of Tweets in Each Cluster by Scenario

| Scenario | Cluster | Jaccard Similarity |
|---|---|---|
| 1 | Cluster 0 | 0.4003 |
| | Cluster 1 | 0.2435 |
| | Cluster 2 | 0.2079 |
| | Cluster 3 | 0.9889 |
| | Cluster 4 | 0.9769 |
| | Cluster 5 | 0.2222 |
| 2 | Cluster 0 | 0.2826 |
| | Cluster 1 | 0.1554 |
| | Cluster 2 | 0.9831 |
| | Cluster 3 | 0.3400 |
| 3 | Cluster 0 | 0.3351 |
| | Cluster 1 | 0.1968 |
| | Cluster 2 | 0.3738 |
| | Cluster 3 | 0.8571 |
| 4 | Cluster 0 | 0.4459 |
| | Cluster 1 | 0.9772 |
| | Cluster 2 | 0.4210 |
| | Cluster 3 | 0.9920 |
| | Cluster 4 | 0.3874 |
| | Cluster 5 | 0.2264 |
| | Cluster 6 | 0.3947 |
| | Cluster 7 | 0.1957 |
| | Cluster 8 | 0.7564 |

시나리오 1에서 클러스터 내 자카드 유사도 평균이 가장 높은 클러스터는 클러스터 3으로서 유사도는 0.9889이다. 이 클러스터에는 인스턴트 메신저 ID $U_8$, $U_{11}$, $U_{14}$가 포함되어 있으며, 다양한 마약 용어와 '뽕라덴복귀'라는 다른 트윗에서 사용되지 않는 특정 문구를 사용하는 다수의 트윗으로 구성된 것을 확인하였다. 두 번째로 높은 자카드 유사도 평균은 0.9769로서 클러스터 4에 해당된다. 이 클러스터에서는 인스턴트 메신저 ID $U_3$, $U_{14}$가 속하며 '작대기팝니다'와 '아이스팝니다'와 같은 구매를 유도하는 문구를 사용하는 트윗들로만 구성되어 높은 유사도를 가진 것으로 확인되었다. 시나리오 2에서 클러스터 내 자카드 유사도 평균이 가장 높은 클러스터는 클러스터 2이며 유사도 평균은 0.9831이다. 이 클러스터에는 인스턴트 메신저 ID $U_7$, $U_8$, $U_{10}$이 속하며 다양한 마약 용어와 '뽕라덴복귀'라는 특정 문구가 사용되는 트윗들로 구성된 것을 확인되었다. 시나리오 3에서 가장 높은 클러스터 내 자카드 유사도 평균은 0.8571로 클러스터 3에 해당된다. 이 클러스터에서는 인스턴트 메신저 ID $U_6$을 중심으로 '오실장', '부산작대기', '사칭주의'와 같은 직급 및 특정 지역과 마약 용어를 함께 사용하는 트윗들이 다수 존재하고, 이러한 트윗들은 동일한 문구와 구조를 패턴으로 사용하여 유사도가 높은 것으로 확인되었다. 시나리오 4에서 가장 높은 클러스터 내 자카드 유사도 평균은 0.9920으로서 클러스터 3에 해당된다. 이 클러스터에 속한 인스턴트 메신저 ID는 $U_8$이며 '뽕라덴복귀'와 같은 다른 트윗에서는 사용되지 않는 특정 용어 및 문구로 트윗을 구성하고 있는 것으로 확인되었다. 다음으로 높은 자카드 유사도 평균은 0.9772로 클러스터 1에 해당된다. 이 클러스터의 인스턴트 메신저 ID는 $U_3$이 중심을 이루고 있으며, $U_3$에서 주로 사용되는 '텔 레 문 의'와 같은 특정 문구를 다수의 트윗이 포함하는 것으로 확인되었다. 또한, 클러스터 8의 자카드 유사도 평균은 0.7564로 높은 클러스터 유사도와 상대적으로 낮은 자카드 유사도 평균을 보였다. 이 클러스터는 인스턴트 메신저 ID $U_2$와 $U_6$으로 구성되어 있으며 '오실장', '최고급 제품 및 서비스로 모시겠습니다'와 같은 직급 및 마약 판매 서비스의 품질을 강조하는 문구가 트윗에 공통적으로 나타나는 것을 확인할 수 있었다. 따라서, 모든 시나리오에서 높은 자카드 유사도 평균을 보인 클러스터들의 공통점은 특정 형식과 고유한 문구인 '뽕라덴복귀', '오실장' 등이 사용되었으며 트윗들이 매우 유사한 문장 구조를 사용하는 것으로 분석되었다. 또한, 각 클러스터에서는 고유한 인스턴트 메신저 ID를 중심으로 트윗이 반복적으로 업로드되었고, 특정 직급 및 서비스 품질을 강조하는 문구가 일관되게 사용되면서 유사한 트윗들의 클러스터링이 효과적으로 이루어진 것으로 보여진다. 반면, 자카드 유사도 평균이 가장 낮은 클러스터들을 분석한 결과, 다양한 마약 용어를 중심으로 트윗들이 군집 된 것으로 나타났지만, 마약 용어 외에 홍보 방식이나 문장 구조의 패턴이 달라, 클러스터 내 유사도가 낮게 측정되었다. 이러한 패턴은 시나리오 1부터 시나리오 4까지 반복적으로 나타났으며 유사한 마약 용어가 사용되었음에도 각 트윗의 홍보 방식과 문구가 다양하게 나타나 클러스터링이 적절히 이루어지지 않은 것으로 분석되었다.

### 5.2. Brute force-based Analysis of Novel Indentifiers of Criminal Organizations

본 절에서는 수집된 트윗을 전수조사하여 동일한 범죄 조직이 사용하는 해시태그, 키워드 등 고유 식별자를 분석하고 제안된 클러스터링 모델의 유효성을 검증한다. 범죄 조직의 고유 식별자로는 트윗의 인스턴트 메신저 ID 표기 방식, 게시물의 구조, 마약 금액 제시 여부, 특정 지역 언급, 마약 사용 방법 및 효과를 설명하는 문구, 그리고 마약 용어의 포함 여부를 분석하였다.

먼저, 수집된 인스턴트 메신저 ID를 중심으로 트윗 전수조사를 통해 분석하여 동일 조직으로 분류한 결과, $U_4$, $U_{13}$, $U_{14}$, $U_{15}$는 검열을 회피하기 위해 'a p p l e'과 같이 인스턴트 메신저 ID를 띄어쓰기를 통해 표기하였으며 트윗 구성, 마약 유형 및 거래 금액 정보, 해시태그 내 도메인 포함 등의 공통된 특징을 보였다. $U_3$ 또한 인스턴트 메신저 ID 띄어쓰기를 통해 검열을 회피한 것을 볼 수 있었지만, 홍보 트윗 내에 거래 금액이나 특정 지역명이 언급되지 않고 마약 정보만 포함되어 있어, 앞서 언급된 인스턴트 메신저 ID와의 동일 조직으로 판단하지 않았다. $U_2$와 $U_6$은 '오실장', '거물하부대리'와 같은 직급을 나타내는 단어와 '최고급 제품 및 서비스로 모시겠습니다'라는 문구를 사용하여 두 인스턴트 메신저 ID는 동일 조직인 것으로 판단하였다. $U_7$과 $U_{10}$은 유사한 홍보 구조를 보였으며 주로 '엑스터시효능', '천안아이스'와 같은 마약 제품의 효능과 특정 지역 정보를 포함한다는 점에서 동일 조직으로 판단하였다. 반면, $U_8$이 게시한 트윗에는 다른 계정이 게시한 트윗에는 포함되지 않는 '뽕라덴복귀'라는 문구가 일관되게 사용되었고, 거래 방식이나 지역 언급보다는 마약 유형을 언급하는 특징을 가지고 있어 독립적인 조직으로 판단하였다. $U_1$은 '구매 시 트위터 보고 왔다 하시면 ' 또는 '텔레 유저들 추천으로 트위터에 왔습니다'와 같은 문구를 사용하였으며 이는 다른 트윗에서는 사용되지 않은 표현이기 때문에 해당 인스턴트 메신저 ID도 독립적인 조직으로 판단하였다. $U_5$는 '사기 걱정 NO', '안전한 거래' 등의 문구와 함께 다양한 마약 홍보 및 '착한 딜러 샘플부터 주문주세요'라는 표현을 사용하였으며 이 문구 또한 다른 트윗에서는 사용되지 않았기 때문에 해당 인스턴트 메신저 ID 또한 독립적인 조직으로 판단하였다. 마지막으로 $U_9, U_{11}, U_{12}$는 게시한 트윗의 수가 매우 적고, 마약 유형, 지역 언급, 홍보 방식에서 일관성이 부족하여 단일 조직으로 판단하였다.

### 5.3. Analysis of Instant Messenger IDs Clustered into Common Clusters in all the Scenarios

본 절에서는 사이버 수사를 위한 추적 대상 우선순위를 책정하기 위하여 모든 시나리오의 클러스터링 결과, 동일 조직으로 판단된 인스턴트 메신저 ID를 분석하였다.

본 연구에서 선별된 15개의 인스턴트 메신저 ID를 4개의 클러스터링 시나리오 기반으로 분석한 결과, 2개 이상의 인스턴트 메신저 ID가 조합된 집합은 총 4,928개로 확인되었다. 이 중, 하나의 시나리오에서만 조합된 집합은 2,830개, 두 개의 시나리오에서 공통적으로 조합된 집합은 901개, 세 개의 시나리오에서 공통적으로 조합된 집합은 202개, 네 개의 시나리오에서 공통적으로 조합된 집합은 35개로 확인되었다. 즉, 모든 시나리오에서 공통된 조직으로 판단된 35개 조직이 동일 조직일 확률이 높다고 판단할 수 있으며 이 중, 특정 조직이 다른 조직의 부분 집합으로 표현되는 조직을 하나의 조직으로 판단하여 최종적으로 Table 8과 같이 7개의 조직을 추출하였다.

Table 8. Instant Messenger IDs identified as a Common Organization with High Probability

| Organization | Instant Messenger IDs |
|---|---|
| (a) | $U_2$, $U_6$ |
| (b) | $U_5$, $U_6$ |
| (c) | $U_9$, $U_{10}$ |
| (d) | $U_{11}$, $U_{14}$ |
| (e) | $U_5$, $U_{10}$ |
| (f) | $U_7$, $U_{10}$, $U_{12}$ |
| (g) | $U_3$, $U_4$, $U_{13}$, $U_{14}$, $U_{15}$ |

Table 8의 (a)에서 $U_2$와 $U_6$는 모든 시나리오에서 동일 조직으로 판단된 것을 볼 수 있다. 시나리오 1부터 시나리오 3의 경우에는 두 계정 외에도 다른 계정이 함께 클러스터에 포함되었지만, 시나리오 4에서는 두 계정만 하나의 클러스터로 형성되었다. 두 계정의 자카드 유사도 평균은 0.7564로 높게 나타났으며 전수조사 분석 결과에서도 유사한 홍보 문구를 사용하여 동일 조직으로 식별되었다. Table 8의 (f)에 속한 $U_7$, $U_{10}$, $U_{12}$도 모든 시나리오에서 동일 조직으로 판단되었지만, 자카드 유사도 평균은 0.1957로 낮게 측정되었다. 그러나, 시나리오 2의 클러스터 2에 속하는 $U_7$, $U_8$, $U_{10}$을 자카드 유사도 기반으로 분석한 결과, 유사도 평균이 0.9831로 높게 나타났으며 $U_7$과 $U_{10}$이 게시한 트윗은 '아이스 샘플: 20만'과 같은 마약 금액과 '후리하는 방법', '작대기 효능' 등 마약의 쓰

임과 효과를 강조하는 동일한 패턴을 보이는 것을 확인하여 $U_7$과 $U_{10}$은 동일한 조직으로 보는 것이 타당하다.

Table 8의 (b), (c), (d), (e), (g)의 경우에는 자카드 유사도 평균 및 전수조사 결과, (a)와 (f)에 비해 동일 조직으로 판단할 수 있는 근거가 부족하여 총 15개 계정 중, $U_2$와 $U_6$, 그리고 $U_7$과 $U_{10}$은 높은 확률로 동일한 조직으로 판단할 수 있다. 즉, 위 4개 계정은 사이버 수사 계획 수립 시, 우선순위가 높은 추적 대상으로 선정할 수 있다.

## VI. Conclusion

소셜미디어를 통해 확산되는 마약 범죄가 점점 더 조직화되고 진화하면서, 운반책이나 관리책을 검거하더라도 범죄 조직을 근본적으로 제거하지 못하는 경우가 많다. 따라서 마약 유통을 근절하고 범죄 조직의 뿌리를 제거하려면 조직의 구성과 규모를 정확히 파악하는 것이 필수적이다.

본 논문에서는 10대부터 청년층 사이에서 마약 범죄가 빈번하게 발생하는 소셜미디어 'X'에서 마약 관련 트윗 데이터를 수집하고, 클러스터링 모델을 활용하여 마약 범죄 조직 및 규모를 식별하기 위한 모델을 제안하였다. 먼저 BERT와 GloVe 모델 기반으로 수집된 트윗 텍스트를 임베딩하고, 엘보우 및 실루엣 기반으로 최적의 클러스터 수를 도출한 후, K-means Clustering과 Spectral Clustering 모델 기반으로 동일 범죄 조직에서 게시한 것으로 판단되는 트윗을 클러스터링하였다. 본 논문에서 소셜미디어 'X'에서 수집한 마약 범죄 관련 트윗은 총 16,232개로서 모든 트윗은 단 15개의 계정에서 게시한 것으로 확인되었다. 또한, 적용된 텍스트 임베딩 모델 및 클러스터링 모델의 조합으로 실험 시나리오를 개발하여 클러스터링을 수행한 결과, 트윗에 포함된 마약 관련 키워드, 장소, 문구 등의 패턴에 의해 동일 조직이 클러스터링 되었음을 확인할 수 있었다. 또한, 제안된 모델의 유효성을 판단하기 위하여 자카드 유사도 및 전수조사 기반으로 클러스터 내의 모든 트윗의 유사성을 확인한 결과, 자카드 유사도 평균이 높을수록 전수조사에서 도출한 동일 조직의 고유 식별자도 유사함을 확인하였다. 마지막으로 모든 시나리오에서 동일 조직으로 판단된 계정을 식별하여 자카드 유사도 및 전수조사 결과와 비교함으로써 높은 확률로 동일 조직으로 판단되는 4개의 계정을 도출하여 사이버 수사 계획 수립 시, 우선적으로 추적해야 하는 대상을 선정하였다.

소셜미디어 범죄 데이터를 활용한 기존 연구들은 주로 텍스트 마이닝 및 머신러닝 기반으로 범죄 유형을 분류하거나 범죄 패턴을 분석하는 연구를 수행한 반면, 본 논문은 범죄 추적을 위해 필요한 범죄 조직 및 규모를 식별하는 연구를 진행하였다는 점에서 차별성을 갖는다.

본 논문에서 제안한 모델은 소셜미디어 상에서 발생하는 마약 범죄를 체계적으로 추적하기 위한 사이버 수사 기술로 활용될 수 있으며 마약 범죄뿐 아니라, 조직적으로 이루어지는 불법 도박, 디지털 성범죄 등 다양한 범죄 수사에도 활용될 수 있을 것이다. 향후에는 트윗에서 수집된 인스턴트 메신저 ID뿐 아니라, 이메일, 가상화폐 지갑 주소 등 다양한 식별 정보를 연동하여 더욱 효과적인 범죄 단서를 확보하는 연구를 진행할 계획이다.

## ACKNOWLEDGEMENT

This work was supported by 'Tech. Challenge for Future Program Policing(www.kipot.or.kr)' funded by Ministry of Science and ICT(MSIT, Korea) & Korean National Police Agency(KNPA, Korea). [Project Name : Development of Active Dark Web Information Collection, Analysis and Tracking Technology to Prevent Dark Web Crime / Project Number : RS-2023-00244362]

This work was supported by Korea Foundation for Women In Science, Engineering and Technology (WISET) grant funded by the Ministry of Science and ICT(MSIT) under the team research program for female engineering students.

## REFERENCES

[1] Digital 2024 Global Overview Report, https://datareportal.com/reports/digital-2024-global-overview-report

[2] Y. Kim, J. Choi, and J. Shin, "Social Losses Due to Drug Addiction Calculated in Monetary Terms... [Investigation K] ['Weak' Society, Talking About Drugs]," KBS NEWS, https://news.kbs.co.kr/news/pc/view/view.do?ncd=7714951.

[3] D. Lee, "Drugs Smuggled Overseas, Sold via Cryptocurrency and SNS... Organization Arrested (Comprehensive)," NEWSIS, https://www.newsis.com/view/NISX20240423_0002709777.

[4] Y. Lee, "70 Arrested for Distributing Drugs Nationwide via Telegram,", Herald Economy, https://biz.heraldcorp.com/view.php?ud=20240709050189.

[5] S. Y. Shin, "[Exclusive] Undercover Investigations in Drug Crimes

Like in 'New World'," Seoul Newspaper, https://www.seoul.co.kr/news/society/2022/08/18/20220818010010.

[6] E. Choi, S. Lee, H. Kwon, M. Kim, I. Lee, and S. Lee, "A Study on the Comparison and Semantic Analysis between SNS Big Data, Search Portal Trends and Drug Case Statistics," Journal of Digital Convergence, vol. 19, no. 2, pp. 231-238, Feb. 2021. DOI: 10.14400/JDC.2021.19.2.231

[7] S. Degadwala, D. Vyas, M. R. Hossain, A. R. Dider, M. N. Ali, and P. Kuri, "Location-Based Modelling And Analysis Of Threats By Using Text Mining," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 1940-1944, Coimbatore, India, Aug. 2021. DOI: 10.1109/ICESC51422.2021.9532825.

[8] D. Petrou, V. Martinez-Gil, F. Castillo, C. Tunc, and R. Bryce, "Twitter Account Analysis for Drug Involvement Detection," 2023 3rd Intelligent Cybersecurity Conference (ICSC), pp. 9-16, San Antonio, TX, USA, Oct. 2023. DOI: 10.1109/ICSC60084.2023.10349992.

[9] F. Zhao et al., "Computational Approaches to Detect Illicit Drug Ads and Find Vendor Communities Within Social Media Platforms," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 19, no. 1, pp. 180-191, 3 2022. DOI: 10.1109/TCBB.2020.2978476.

[10] C. Hu, M. Yin, B. Liu, X. Li, and Y. Ye, "Detection of Illicit Drug Trafficking Events on Instagram: A Deep Multimodal Multilabel Learning Approach," Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), pp. 3838-3846, Virtual Event, Australia, 2021. DOI: 10.1145/3459637.3481908.

[11] C. Hu, B. Liu, X. Li, and Y. Ye, "Unveiling the Potential of Knowledge-Prompted ChatGPT for Enhancing Drug Trafficking Detection on Social Media," arXiv preprint arXiv:2307.03699, 7 2023. DOI: https://doi.org/10.48550/arXiv.2307.03699

[12] T. Ma, Y. Qian, C. Zhang, and Y. Ye, "Hypergraph Contrastive Learning for Drug Trafficking Community Detection," 2023 IEEE International Conference on Data Mining (ICDM), pp. 1205-1210, Shanghai, China, Dec. 2023. DOI: 10.1109/ICDM58522.2023.00149.

[13] N. Shah, N. Bhagat, and M. Shah, "Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention," Visual Computing for Industry, Biomedicine, and Art, vol. 4, no. 9, pp. 1-14, Apr. 2021. DOI: 10.1186/s42492-021-00075-z.

[14] Y. Wang, W. Yu, S. Liu, and S. D. Young, "The Relationship Between Social Media Data and Crime Rates in the United States," *Social Media + Society*, vol. 5, no. 1, Mar. 2019. DOI: 10.1177/2056305119834585.

[15] F.-C. Tsai, M.-C. Hsu, C.-T. Chen, and D.-Y. Kao, "Exploring drug-related crimes with social network analysis," Procedia Computer Science, vol. 159, pp. 1907-1917, Oct. 2019. DOI: 10.1016/j.procs.2019.09.363.

[16] P. Siriaraya, Y. Zhang, Y. Wang, Y. Kawai, M. Mittal, P. Jeszenszky, and A. Jatowt, "Witnessing Crime through Tweets: A Crime Investigation Tool based on Social Media," Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '19), pp. 568-571, New York, NY, USA, Nov. 2019. DOI: 10.1145/3347146.3359082.

[17] E.-Y. Park, J. Kim, and C.-H. Kim, "A Tracking Method of Same Drug Sales Accounts through Similarity Analysis of Instagram Profiles and Posts," Journal of The Korea Society of Computer and Information, vol. 29, no. 2, pp. 109-118, Feb. 2024. DOI: 10.9708/jksci.2024.29.02.109

[18] .H. J. Park, "A Study on Drug trading countermeasures via internet and sns," Journal of the Korea Information Assurance Society, Vol. 18, No. 1, pp. 93-102, Mar. 2018.

[19] H.-j. Song and S.-y. Oh, "An Analysis of the Progress of Youth Drug Crimes Using Cryptocurrency," Korean Journal of Convergence Science, vol. 12, no. 11, pp. 133-145, Nov. 2023. DOI: 10.24826/KSCS.12.11.9.

[20] F. M. Bianchi, D. Grattarola, and C. Alippi, "Spectral Clustering with Graph Neural Networks for Graph Pooling," Proceedings of the 37th International Conference on Machine Learning, vol. 119, pp. 874-883, Jul. 2020. Available: https://proceedings.mlr.press/v119/bianchi20a.html.

[21] Lee, Seul-ki, "Analysis of Question Types by Student Writers Using Clustering in Writing Using Generative Artificial Intelligence," The Journal of Korean Language and Literature Education, Vol. 84, pp. 69-105, Feb. 2024.

[22] S.-Y. Ihm, "A Study on Clustering-based Color Extraction Method for Pill Classification," Journal of Big Data Service, vol. 1, no. 1, pp. 79-84, Jul. 2023. DOI: 10.61241/KBDSS.01.01.07.

[23] Inmook Lee, Jaehong Min, Kyoungtae Kim, and Seung-Young Kho, "Generating Travel Patterns of Public Transportation Users Using a k-means Clustering Based on Smart Card Data," Journal of the Korean Society for Railway, Vol. 23, No. 3, pp. 204-215, Mar. 2020. DOI: 10.7782/JKSR.2020.23.3.204

[24] D.-T. Vu and J.-Y. Jeong, "Collision Risk Assessment by using Hierarchical Clustering Method and Real-time Data," Journal of the Korean Society of Marine Environment and Safety, vol. 27, no. 4, pp. 483-491, June 2021. DOI: 10.7837/kosomes.2021.27.4.483.

[25] J.-H. Tak, J.-Y. Hong, and D.-J. Park, "A Study on Road Link Vulnerability Assessment Based on Clustering Analysis for Disaster Situations," Journal of Korean ITS Society, vol. 22, no. 2, pp. 29-43, Feb. 2023. DOI: 10.12815/kits.2023.22.2.29

[26] G. Hajela, M. Chawla, and A. Rasool, "A Clustering Based Hotspot Identification Approach For Crime Prediction,"

*Procedia Computer Science*, vol. 167, pp. 1462-1470, Apr. 2020. DOI: 10.1016/j.procs.2020.03.357.

[27] D. Y. Kim and S. W. Jung, "A Preliminary Study on the Application of Spatial Clustering Techniques for Crime Prediction," in *Proceedings of the Korean Society of Spatial Information Science Conference*, vol. 2020.6, pp. 111-114, Jun. 2020.

[28] A. A. Alkhaibari and P.-T. Chung, "Cluster analysis for reducing city crime rates," 2017 IEEE Long Island Systems, Applications and Technology Conference (LISAT), pp. 1-6, Farmingdale, NY, USA, May 2017. DOI: 10.1109/LISAT.2017.8001983.

[29] Y. van Gennip, B. Hunter, R. Ahn, P. Elliott, K. Luh, M. Halvorson, S. Reid, M. Valasik, J. Wo, G. E. Tita, A. L. Bertozzi, and P. J. Brantingham, "Community detection using spectral clustering on sparse geosocial data", arXiv preprint arXiv:1206.4969, Jun. 2012. DOI: https://doi.org/10.48550/arXiv.1206.4969

[30] K. Joseph, R. J. Gallagher, and B. F. Welles, "Who Says What with Whom using Bi-Spectral Clustering to Organize and Analyze Social Media Protest Networks," Nov. 2020. DOI: 10.5117/CCR2020.2.002.JOSE

[31] R. Tibshirani, G. Walther, and T. Hastie, "K-Means Clustering and Related Algorithms," Technical Report, Stanford University, 2004.

[32] U. von Luxburg, "A Tutorial on Spectral Clustering," Technical Report No. MPIK-TR-149, Max Planck Institute for Biological Cybernetics, Aug. 2007. DOI: 10.1007/s11222-007-9033-z

## Authors

Jin-Gyeong Kim entered the Department of Computer Engineering at Daegu University, Gyeongsan, South Korea, in 2024 and received a bachelor's degree in Computer and Information Engineering.

She is currently pursuing a master's degree in the Department of Computer and Information Engineering at the Graduate School of Daegu University. Her research interests include cybersecurity, cybercrime, and artificial intelligence.

Eun-Young Park is an undergraduate student in the Department of Computer Engineering, Daegu University, Gyeongsan, Korea, since 2021. Her research interests include cybersecurity, internet of things, and artificial intelligence.

Da-Sol Kim is an undergraduate student in the Department of Computer Engineering, Daegu University, Gyeongsan, Korea, since 2023. Her research interests include cybersecurity, Darkweb of things, and artificial intelligence.

Cho-Won Kim is an undergraduate student in the Department of Computer Engineering, Daegu University, Gyeongsan, Korea, since 2023. Her research interests include cybersecurity and artificial intelligence.

Jiyeon Kim received the B.S. and Ph.D. degrees in information security engineering from Seoul Women's University, Seoul, South Korea, in 2007 and 2013, respectively. Dr. Kim was a Postdoctoral Research Associate in the Department of Electrical and Computer Engineering, Carnegie Mellon University, United States, from 2014 to 2017. She is currently an Assistant professor in the Department of Computer Engineering, Daegu University, Gyeongsan, South Korea. Her research interests include cybersecurity, cybercrime investigation, cloud computing, artificial intelligence, and critical infrastructure protection.