國土計劃
pISSN: 1226-7147, eISSN: 2383-9171
https://doi.org/10.17208/jkpa.2025.04.60.2.25

Journal of Korea Planning Association Vol.60, No.2 (2025)
대한국토·도시계획학회지 「국토계획」 제60권 제2호 pp.25-40

서울시 범죄 토픽의 시공간 변화 분석
: 빅카인즈 뉴스 데이터와 KoBERTopic 모형을 활용하여*

Temporal and Spatial Analysis of Crime Topics in Seoul, Korea
: Using BigKinds News Data and the KoBERTopic Model

성우석** · 김혜빈*** · 이수기****
Seong, Wooseok · Kim, Hyebin · Lee, Sugie

Abstract

Crime negatively impacts daily activities and diminishes the quality of life for urban residents. Analyzing the spatial distribution and types of crime, along with formulating strategies to reduce crime, is essential for enhancing citizens' well-being. In Korea, crime location data is private, while crime rate data, categorized by type, is available at the city, county, and district levels. Recently accumulated digital news data provides details on crime locations, types, and targets. However, extracting relevant crime-related information from the extensive collection of news articles amassed over the past decades has posed challenges. This study examined the spatial distribution and changes in major crimes by analyzing news article data from 2000 to 2023, sourced from Big Kinds, utilizing the KoBERTopic text mining methodology and focusing on Seoul. Crime-related news was filtered using text mining technology, and location information tied to administrative districts was extracted from the articles to analyze crime concentration over time. Moreover, shifts in major crime topics were identified using the KoBERTopic methodology. The analysis revealed that the types of crimes and crime-concentrated areas in Seoul changed annually, influenced by significant factors. Furthermore, changes in crime topics over time were visually represented spatially. This study is notable for leveraging large-scale news data to analyze shifts in crime locations and concentrations, leading to policy recommendations for crime prevention.

주제어 범죄, 시공간 패턴 분석, 뉴스 데이터, 텍스트마이닝, KoBERTopic
Keywords Crime, Spatiotemporal Pattern Analysis, News Data, Text Mining, KoBERTopic

Ⅰ. 서 론

1. 연구의 배경 및 목적

오늘날 범죄로부터 안전한 삶은 도시민의 삶의 만족도를 결정하는 중요한 요인 중 하나이다. 대검찰청(2023a)의 "2023 범죄분석"에 따르면 지난 10년간 연도별 전체범죄 발생비는 2021년까지 감소하는 추세를 보이다가, 2022년에는 전년 대비 증가하는 것으로 나타났다. 하지만, 특별법범죄와 형법범죄의 발생비를 구분 지어 살펴보게 되면 특별범범죄는 2016년을 기점으로 꾸준히 감소하고 있지만 형법범죄는 큰 변동 없이 증가와 감소를 반복하다가 2022년에는 전년 대비 크게 증가하는 것으로 나타났다(그림 1). 이는 도로교통법(음주운전), 식품위생법 등과 같은 특별법범죄는 감소하고 있지만 재산범죄, 강력범죄 등에 대한 형법범

* 본 연구는 2024년 대한국토·도시계획학회 춘계산학학술대회에서 발표하여 우수논문상 수상 논문을 수정·보완하여 작성하였음.
** Master's Student, Department of Urban Planning & Engineering, Hanyang University (First Author: wsseong18@gmail.com)
*** Master's Degree, Department of Urban Planning & Engineering, Hanyang University (Co-author: khb723@naver.com)
**** Professor, Department of Urban Planning & Engineering, Hanyang University (Corresponding Author: sugielee@hanyang.ac.kr)

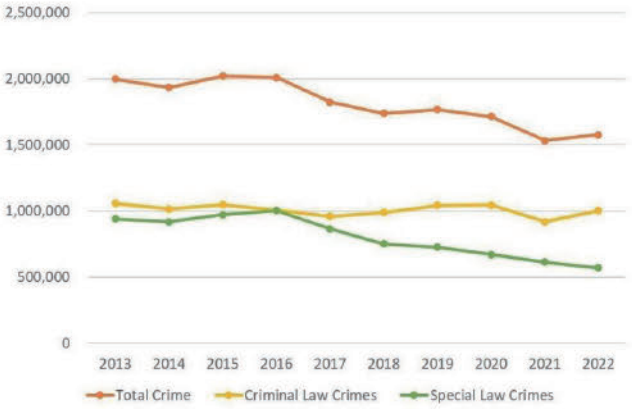


그림 1. 연도별 전체범죄, 형법범죄, 특별법범죄 발생 건수(2013-2022)
Figure 1. Total crimes, criminal law crimes, and special law crimes by year (2013-2022)
Source: 2023 Statistical Analysis on Crime (2013-2022)

죄는 꾸준히 발생하는 것으로 해석할 수 있다. 이에 따라 범죄 유형별 발생 현황과 변화 양상을 분석하고, 범죄 발생 원인에 대한 분석과 범죄 저감 방안 마련이 필요하다. 특히, 각 범죄 유형에 따른 발생 원인에 대한 분석과 이에 대한 맞춤형 대책을 제시하는 것은 범죄 예방 정책 수립에 중요한 기초자료로 활용될 수 있다.

그동안 범죄의 흐름을 파악하기 위해 범죄발생과 영향 요인을 분석하는 연구가 다수 진행되었지만, 경찰청에서 제공하는 범죄 관련 자료는 자치구 수준 이상의 큰 공간단위 분석만 가능하며 미시적인 특성차이를 분석한 연구는 미비한 실정이다. 범죄 데이터는 위치정보와 개인정보를 포함하고 있어 범죄에 대한 동네의 이미지, 주민들의 불안감 등을 형성할 수 있기 때문에 전체적인 범죄발생 수를 큰 공간단위 수준에서 합계하여 제공하고 있어 미시적 분석에 한계가 있다.

도시민들은 뉴스, Social Network Service(SNS)와 같은 미디어를 통해 일상생활에서 범죄에 대한 정보를 얻으며, 범죄 정책 형성에 간접적으로 영향을 미치고 있다(Garland, 2000). 뉴스 자료의 경우 과거부터 현재까지 온라인 텍스트 데이터가 축적되고 있으며, 최근 발생하고 있는 범죄 사건(절도, 강도, 살인, 성폭력 등)의 다양한 유형과 양상을 확인할 수 있다(박준영 외, 2016). 예를 들어, 비교적 최근에 발생한 '택시기사 폭행 사건(2020.11.06.)', '신림역 칼부림(2023.07.21.)', '신협 강도 사건(2023.08.18.)' 등의 강력범죄는 관련 뉴스 및 SNS를 통해 확산되며 도시민들의 범죄 두려움을 증가시키는 요인이 되기도 한다. 이처럼 범죄발생은 심리적인 요소와 연관되어 있으며 결과적으로 도시민의 기본적인 일상 활동에 부정적인 영향을 미치는 중요한 요소로 작용한다.

이와 같이 온라인 텍스트 데이터는 COVID-19의 발생 및 4차 산업혁명의 도래로 인해 비대면 온라인 플랫폼이 활성화되어 빠르게 축적되고 있다. 또한, 데이터 마이닝 기술의 발전으로 비정형 텍스트 데이터를 수집하고 가공하는 것이 용이하게 되었으며, 다수의 연구에서도 활용되고 있는 추세이다. 이러한 빅데이터와 분석기술의 발전은 소규모 데이터의 한계를 어느 정도 해결할 수 있으며, 현재 또는 미래에 발생할 수 있는 도시문제에 대응할 수 있는 가능성을 제공한다(하재현 외, 2019).

본 연구는 서울시를 대상으로 장기간(2000-2023)에 해당하는 시계열 데이터 구축을 통해 범죄 동향의 변화를 분석하고자 한다. 이를 위해 24년간 뉴스 데이터를 크롤링을 진행하였으며, 시·공간적으로 범죄 동향의 변화를 파악하고자 KoBERTopic을 활용한 시계열 토픽모델링과 핫스팟 분석을 진행하였다. 시간의 흐름에 따른 도시 내 공간 변화를 확인하기 위해 네트워크 분석방법론을 활용하여 범죄 키워드 사이의 관계를 도출하고자 한다. 본 연구의 결과는 범죄 유형과 동향을 키워드 도출을 통해 효율적으로 파악할 수 있고, 범죄 핫스팟 지역의 변화를 파악하여 이에 따른 안전한 도시를 위한 정책 형성에 유용하게 활용될 수 있다.

II. 선행연구 고찰

1. 범죄 시계열 분석 연구

범죄 유형과 발생 위치의 시간적 패턴을 식별하는 것은 범죄 감소에 필수적인 요소이다. 이러한 패턴을 파악하여 범죄가 가장 많이 발생하는 기간 또는 지역의 순찰을 강화하거나 범죄 예방 시설을 배치하여 범죄 활동을 사전에 방지할 수 있다(Yadav and Sheoran, 2018). 최근 빅데이터의 발전으로 다양한 시계열 데이터가 증가하고 있으며 이에 따라 다양한 범죄 연구가 다수 진행되었다.

시계열 데이터는 도시 범죄의 흐름과 패턴을 식별하고 미래의 범죄율 예측, 정책 효과 평가, 예방 시설 배치를 통해 안전한 도시를 만드는 데 도움을 줄 수 있다(Borges et al., 2018). Devi and Kavitha(2022)는 시계열 분석 알고리즘을 활용하여 향후 범죄 발생을 예측했으며, 범죄가 발생하기 쉬운 위치를 식별하였다. Towers et al.(2018)은 범죄의 시간적 추세에 대한 예측 분석의 정확성을 높이기 위해 공휴일과 기후(기온, 바람, 강수 등)를 활용하여 연구를 진행하였으며, 연구 결과, 공휴일과 기후를 포함한 범죄 예측의 정확성과 정밀도가 높아진 것으로 나타났다.

또한, 범죄 관련 시계열 자료는 범죄 예방의 정책 및 효과를 분석할 수 있다는 점에서 주목을 받고 있다. 윤우석(2015)은 범죄예방환경조성사업이 시계열 범죄 변화에 미치는 영향을 확인하였다. 분석 결과, 범죄예방 정책이 재산범죄 감소에 유의미한 감소 효과를 보이는 것으로 나타났지만, 폭행에서는 영향이 없는 것으로 나타났다. 따라서 범죄 예방 및 예측을 파악하기 위해서 시계열 패턴을 활용하게 범죄의 변화에 따라 발생하는 문제점에 대해 효율적으로 해결할 수 있다.

2. 텍스트 마이닝과 범죄 관련 연구

텍스트 마이닝을 통해 구조화되지 않은 비정형 데이터 세트에서 새로운 정보를 발견할 수 있으며, 텍스트 문서에서 주요 정보 및 키워드를 추출할 수 있다(Aggarwal and Zhai, 2012). Dasgupta et al.(2017)은 디지털 뉴스 기사 텍스트를 대상으로 피의자, 범죄유형, 위치 및 날짜와 같은 데이터를 포함한 사건을 식별하였다. 이와 더불어 Umair et al.(2020)은 범죄자의 네트워크의 행동을 분석하기 위해 Natural Language Processing(NLP)를 활용하여 뉴스 내 범죄 데이터의 정보를 추출하였다. 이를 통해 핫스팟 지역을 맵핑하고 범죄 특징 간의 패턴의 추세 및 관계를 식별하였다.

최근 다양한 소셜미디어 빅데이터가 축적되면서 감정분석과 도시공간 활용도를 분석하는 것이 용이해졌다. Twitter와 같은 SNS을 기반으로 한 범죄 연구가 다수 진행되었다. Aghababaei and Makrehchi(2018)은 Twitter의 데이터를 활용하여 범죄 동향 예측 가능성을 확인하였으며, 토픽, 정서 및 주제를 통합한 예측 모델을 제안하였다. 분석 결과 텍스트와 범죄 경향 간의 상관관계가 나타났으며 예측 정확도를 향상시키는 효과가 있음을 확인하였다. 또한, Twitter를 사용하여 범죄 분류, 시각화 및 시공간 분석을 통해 범죄 예측에 활용할 수 있다(Vivek and Prathap, 2023).

텍스트 마이닝은 정형화 되어 있지 않은 텍스트문서 내에서 정보와 지식을 추출하는 과정을 말한다. 텍스트 데이터는 특정 분야에 대한 통찰력 있는 정보의 원천이며, 발행된 텍스트의 증가, 기술 발전 등의 추세와 함께 텍스트 마이닝은 다양한 분야에서 의사 결정 프로세스를 지원하는 방식으로 활용되고 있다(Zanini and Dhawan, 2015). 이 중 텍스트 사이의 관계와 패턴을 파악하는 토픽모델링 기법은 꾸준히 연구에서 활용되고 있다(김혜빈·이수기, 2024). 최근 임베딩 기반 토픽모델링이 관심을 받고 있으며, Doc2Vec을 기반으로 한 Top2Vec모델과 Sentence-BERT(SBERT)을 활용한 BERTopic 모델이 주목받고 있다(Angelov, 2020; Grootendorst, 2020). 이처럼 텍스트 마이닝과 네트워크 분석을 통해 도시문제를 분석 및 해결하는 데 활용될 수 있다는 점에서 관심을 받고 있다.

이러한 맥락에서 온라인 뉴스 데이터는 공개된 텍스트를 제공하고 있어 이를 활용하여 범죄 관련 사건을 추출할 수 있다는 장점이 있다. 이는 다양한 지역이나 범죄 종류, 범죄 활동을 모니터링, 비교 및 분석을 가능하게끔 한다(Khairova et al., 2023). 이러한 과정에서 텍스트 마이닝은 주요 분석 방법론으로 대두되었다.

텍스트 마이닝 기법은 범죄 네트워크의 분석(Al-Zaidy et al., 2011)이나 범죄와 범죄자 간의 상호 작용 분석(Elyezjy and Elhaless, 2015) 등에서 효과적인 것으로 확인되었다. 이처럼 텍스트 마이닝을 활용하여 범죄와 범죄자 간의 관계를 분석할 수 있다. 또한, 범죄 관련 용어에 대한 네트워크 접근법을 도입하여 범죄 수사에 활용할 수 있는 것으로 나타났다 (Tseng et al., 2012). 뉴스와 관련하여 박대민(2016)은 인용문, 연도, 매체, 지면별로 뉴스 내 텍스트 데이터를 자체적인 엑셀파일 형식으로 제공하는 뉴스 빅데이터 시스템인 BigKinds를 연구에 활용하였다. 해당 연구는 시계열 데이터를 축적하였으며, 이를 통해 경제지표 등 다양한 시계열 데이터와 비교연구를 진행할 수 있음을 보였다.

한편, 뉴스를 통한 범죄 정보 전달은 개인 또는 인식에 영향을 미칠 수 있으며, 범죄 두려움을 형성하는 데 영향을 미치는 것으로 나타났다(박상조·박승관, 2016; Mastrorocco and Minale, 2018). 이완수·송상근(2020)은 네이버 뉴스를 바탕으로 텍스트 구조분석을 진행하였으며, 국내 뉴스 기사의 텍스트는 범죄 인식에 대해 직접 영향을 미치는 것으로 나타났다.

3. 연구의 차별성

앞서 검토한 텍스트 데이터와 시계열 분석을 통해 도시 내 발생하는 범죄를 분석하는 연구들이 다수 진행되었다. 또한, 뉴스 데이터 중 BigKinds 내 자체 데이터를 활용한 시계열 분석도 다수 진행되었다. 하지만, BigKinds 자체 데이터를 활용한 기존 연구들은 기사 전체 본문이 아니라 BigKinds 측에서 제공하는 기사의 일부 데이터만 활용하였다. 이는 기사 본문에서 범죄 위치를 도출하는 데 어려움이 있다. 범죄 발생흐름에 대한 분석 연구는 공공데이터 또는 예외적인 경우로 경찰청에서 미시적인 단위의 자료를 제공받아 진행되었다. 하지만, 공공데이터는 자치구 단위의 데이터를 제공하고 있어 분석의 공간해상도가 낮아 구체성이 떨어지며, 미시적인 단위의 범죄 자료는 공개되어 있지 않다.

따라서 본 연구는 Python을 활용하여 BigKinds 기사의 원문 전체를 수집하여 KoBERTopic을 활용하여 도시 범죄 발생과 관련된 키워드 도출과 텍스트 데이터 중 위치를 추출하였다. 이를 통해 시간의 흐름에 따른 주요 범죄를 도출하고 범죄의 변화를 비교 분석하였다. 또한, 서울시 내 행정동 위치를 도출하였다는 차별성이 있으며, 과거부터 현재까지 범죄 발생 및 동향에 대해 보다 세밀한 범위에서의 시각적인 자료와 함께 시계열 자료를 구축하였다. 본 연구 결과는 향후 도시 범죄 연구에 있어 시계열 데이터와 범죄 유형을 분석하는 데 기초자료로 활용될 수 있을 것으로 기대된다.

Ⅲ. 연구 방법론

1. 분석 범위 및 데이터

1) 분석 범위

본 연구의 공간적 범위는 서울특별시로 설정하였다. 시간적 범

위는 2000년부터 2023년까지 총 24개 연도로 설정하였다. 수집한 데이터의 범위는 BigKinds에서 행정동 위치를 추출할 수 있는 텍스트 데이터이며, 분석의 공간적 범위는 서울특별시 내 426개의 행정동을 BigKinds에서 사용하는 245개의 행정동으로 변환하여 진행하였다.[1)]

2) 분석 자료

도시 내 발생하는 범죄의 동향을 확인하기 위해 BigKinds의 온라인 기사 텍스트 데이터를 수집한 후 전처리를 진행하여 연구 데이터를 도출하였다. BigKinds는 한국언론진흥재단에서 2016년 공개한 뉴스 빅데이터 분석 시스템이다. BigKinds는 현재 1990년 이후 41개 매체에 대한 뉴스 기사를 제공하고 있으며 종이신문 외에도 언론사 페이지에서 제공하는 기사를 모두 포함하는 뉴스 빅데이터를 제공한다(박대민, 2016). 또한, BigKinds는 기간, 언론사, 통합 분류 및 사건·사고 분류 등을 제공하여 구체적으로 분류를 설정하여 데이터를 수집할 수 있다는 장점이 있다. 이러한 맥락에서 본 연구는 장기 시계열 분석 및 범죄 관련 대규모 텍스트 데이터 수집을 위해 BigKinds의 뉴스 본문을 분석 자료로 설정하였다.

2. 분석 방법

1) 데이터 수집

본 연구는 〈그림 2〉와 같은 연구의 흐름을 가진다. 먼저 BigKinds의 기사 데이터를 수집한 뒤, 2단계에 걸쳐 데이터 전처리를 진행한다. 이후 2000년부터 2023년까지의 도시 범죄의 변화와 동향을 살펴보기 위해 히트맵과 토픽모델링 기법 중 하나인 KoBERTopic 방법론을 사용하였다. 이를 통해 주요 키워드와 토픽을 도출하여 범죄의 시계열적 변화를 파악하였으며, 구체적인 토픽을 분류하여 토픽의 변화를 확인하였다. 또한, 범죄의 공간 분포를 분석하기 위해 텍스트 데이터에서 서울시의 행정동을 추출하여 시각화를 진행하였다. 데이터 수집을 위해 Python의 selenium 4.1.0 패키지 중 하나인 Beautiful Soup 라이브러리를 활용하였다. 뉴스 데이터 중 '범죄'를 추출하기 위해, BigKinds 내 뉴스 검색 카테고리를 사용하여 검색하였다. 언론사별로 제공하는 뉴스 데이터에 차이가 있을 수 있어 '언론사'에서 모든 언론사를 선택하였다. 이후 '통합 분류'에서 '사회' 카테고리를 선택하여 사회에서 발생하는 범죄를 추출하고, '사건·사고 분류'에서 '범죄' 카테고리 내 '성범죄, 범죄일반' 카테고리를 선택하여 도시 내 발생하는 범죄가 아닌 '기업범죄, 정치'를 제외하였다. 수집한 데이터는 연도와 3개의 카테고리, 제목, 본문으로 구성되어 있으며, 총 1,921,607개를 수집하였다.

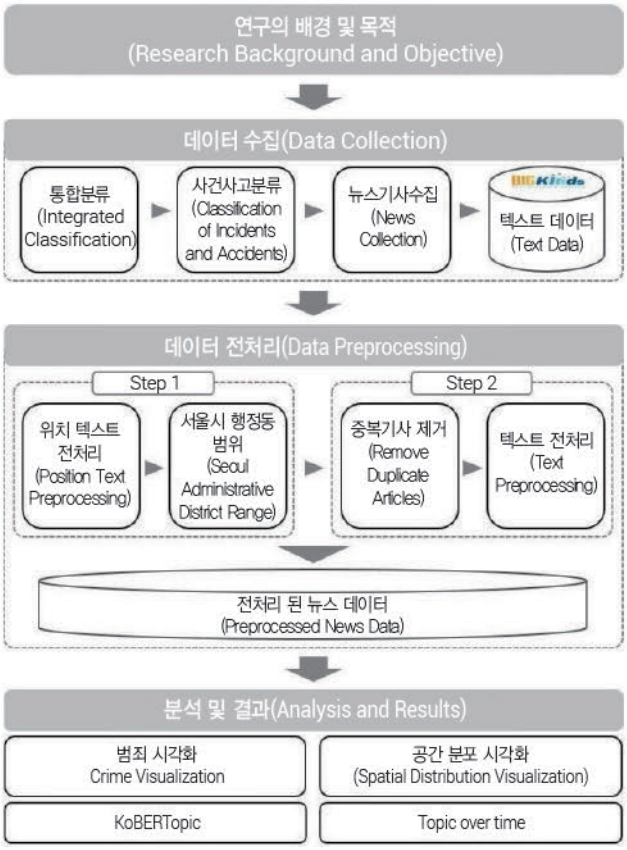


그림 2. 연구 프레임워크
Figure 2. Research framework

2) 데이터 전처리

본 연구는 뉴스의 본문 데이터를 사용하기 때문에 2단계의 전처리 과정을 진행하였다. 1단계는 본문에서 위치 텍스트 도출 이후, 2단계로 기본적인 전처리를 진행하였다. 구체적으로 1단계 과정에서는 뉴스 본문에서 범죄가 발생한 위치를 수집하기 위해 서울시 행정동 추출이라는 전처리를 진행하였다. 범죄 뉴스 데이터는 뉴스의 주제를 제목과 본문을 통해 특정 측면을 묘사, 강조하여 제공하고 있다(이완수·송상근, 2020). 또한, 범죄 발생 위치에 대한 정보도 같이 제공하고 있어, 연구의 공간적 단위를 바탕으로 2023년 서울시 기준 426개의 행정동 단위에 해당하는 데이터만 추출하였다.

2단계 과정에서는 중복된 기사 제거와 불용어 처리를 진행하였다. 뉴스 기사는 특정 키워드를 중심으로 제목과 본문의 텍스트를 구성된다(Tereszkiewicz, 2012). 이는 사회적 이슈에 대해 중복 기사가 발생할 수 있어 제목과 본문에서 중복되는 기사를 삭제하였다. 이후 토픽모델링을 활용하여 본문을 대상으로 전처리를 진행하였다. 토픽모델링은 텍스트 속에서 확률적으로 주제를 추출하는 텍스트 마이닝 방법 중 하나이다(Blei et al., 2003). 따라서 분석을 위해 596개의 Stopwords를 설정하고 해당 텍스트가 포함되어 있는 기사를 제거하였다. 또한, 불용어에 해당하는 조사, 접속사, 불분명한 어휘, 특정 단어 등을 제외한 뒤, 정규식 및

띄어쓰기 처리와 한국어 문장 분할(Korean Sentence Segmentation: KSS)을 사용하여 문장 분리를 진행하였다. 전처리 과정을 통해 도출한 최종 데이터의 개수는 〈표 1〉과 같다.

3) KoBERTopic

Sentence-BERT(SBERT)의 프레임워크를 사용한 클러스터링 기술과 클러스터 기반 TF-IDF의 클래스 기반 변형을 활용하는 토픽모델링 기법을 BERTopic이라 정의한다(Grootendorst, 2022). 구조화되지 않은 텍스트에서 BERTopic은 임베딩 기반 토픽모델링 기법 중 가장 큰 잠재력을 가지고 있으며, 넓은 임베딩 모델 지원 및 모듈식 특성을 통해 연구자가 필요한 모델 구축이 가능하다(Egger and Yu, 2022; Grootendorst, 2023; 김혜빈·이수기, 2024). BERTopic은 크게 총 4단계로 구성된다. 먼저, 문서 임베딩을 진행한 뒤, 차원 축소 기법을 통해 문서 임베딩의 차원을 줄인다. 이후 의미적으로 유사한 문서를 분류하기 위하여 클러스터링을 진행하여, 각 주제별 중요도를 도출하는 클래스 기반 TF-IDF(cTF-IDF)를 사용하여 토픽을 도출한다(Grootendorst, 2022).

표 1. 연도별 최종 분석 데이터 개수
Table 1. Number of final analysis data by year

구분 (Division)	전처리 전 (Before pretreatment)	전처리 후 (After pretreatment)	
	전체 뉴스 개수 (Total news)	행정동 추출 (Administrative 'Dong' boundary)	최종 추출 (Final extraction)
2000	20,186	2,849	1,084
2001	23,075	3,190	890
2002	23,593	1,953	643
2003	26,333	3,815	1,004
2004	29,424	5,523	1,121
2005	28,568	2,208	706
2006	26,701	2,313	808
2007	30,522	3,019	961
2008	45,049	4,122	1,177
2009	60,276	4,178	1,352
2010	68,460	4,036	1,262
2011	78,532	5,902	1,931
2012	102,307	8,885	2,817
2013	105,477	7,931	3,219
2014	105,078	8,101	3,326
2015	139,211	10,983	2,394
2016	112,014	5,466	1,532
2017	101,095	5,620	1,149
2018	122,581	5,327	1,335
2019	147,702	10,108	2,225
2020	117,194	3,930	1,197
2021	127,982	3,719	964
2022	107,826	4,716	1,019
2023	172,421	12,799	3,244

BERTopic은 임베딩 기반 토픽모델링 기법 중 가장 큰 잠재력을 제공하며 넓은 임베딩 모델을 지원하기에 여러 연구에서 활용되고 있는 기법이다(김혜빈·이수기, 2024). 특히 BERTopic은 모듈식 특성을 지원하기에 각 단계에 맞춰 원하는 기법을 활용하여 데이터에 적절한 모델을 구축할 수 있다(Grootendorst, 2023). 임베딩, 차원 축소, 클러스터링, Vectorizer, cTF-IDF 등을 커스텀화 할 수 있다는 것이 특징이며 구체적으로 문서 임베딩의 경우 SBERT와 SpaCy 등을 지원하고, 차원 축소에서는 UMAP, PCA 등을 클러스터링에서는 HDBSCAN, k-Means 등을 지원하고 있다.

이외에도 BERTopic은 여러 토픽모델링 시각화, 분석 기법을 제공하는데, 이러한 기법에는 Topics over time, Text Generation/LLM 등이 있다. Topics over time은 Dynamic Topic Modeling의 일종으로, 시계열 데이터가 포함된 데이터셋에 대하여 토픽모델링을 진행하고 시간에 따른 주제의 빈도 변화를 시각화하는 기법이다. 또한, BERTopic은 토픽명 도출에 있어 선택적으로 Large Language Mode(LLM)을 적용하여 상대적으로 연구자의 주관성이 배제된 토픽명 도출이 가능하다(김혜빈·이수기, 2024). 현재는 Llama2, OpenAI의 GPT API, Transformers 등을 지원하고 있다. 추가적으로 형태소 분석에서는 KoNLPy에 있는 Okt를 사용하였으며, 도출된 토픽 사이의 중심 주제를 주관적이 아닌 객관적으로 도출하기 위해 OpenAI의 ChatGPT4.0을 사용하였다.

최종적으로 본 연구에서는 다국어 지원이 되는 SBERT를 사용하여 문서 임베딩을 진행하는 KoBERTopic 모델을 사용하였다. 차원 축소에서는 UMAP, 클러스터링에서는 HDBSCAN, 토픽 중요도 도출에서는 cTF-IDF를 활용하였으며 도출 과정은 〈그림 3〉과 같다. 이는 BERTopic에서 가장 널리 쓰이는 기본 제공 모듈이다.

Ⅳ. 분석 결과

1. 시계열 토픽 변화

1) KoBERTopic 토픽모델링

범죄 기사를 대상으로 KoBERTopic 분석을 진행하였으며 24개 년도를 2개년도씩 묶어 총 12개의 그룹을 분석하였다. 각 그룹마다 총 30개, 24개, 1개, 4개, 1개, 43개, 35개, 66개, 2개, 21개, 1개, 1개의 토픽을 추출하였다. 서울시 도시 범죄의 이슈를 설명하기 위해 공통적으로 핵심적인 Top 5개의 토픽을 나열하였다. 해당

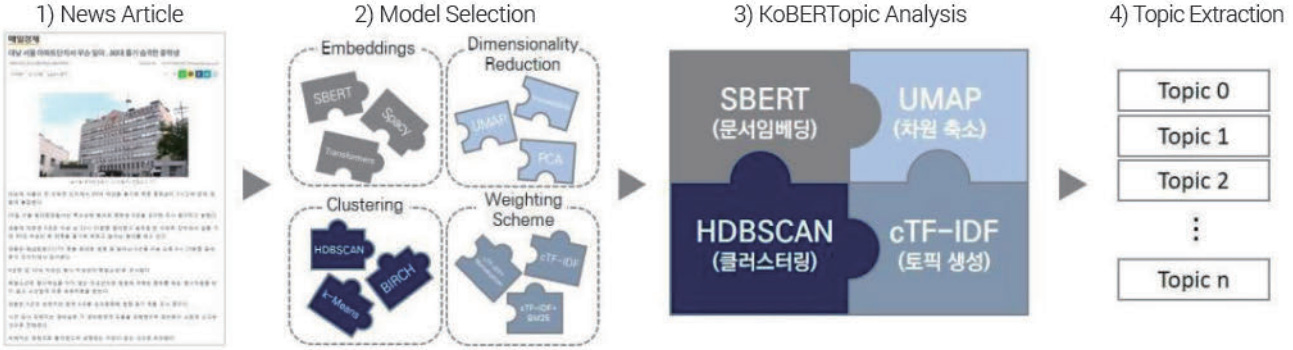


그림 3. BERTopic 도출 과정
Figure 3. BERTopic extraction process

분석은 그룹별로 데이터셋을 분리하여 분석을 진행하였기에 각 시기의 주요 토픽을 독립적으로 확인하는 데에 활용될 수 있다.

〈표 2〉는 KoBERTopic을 통해 추출된 5개의 도시 범죄에 대한 토픽과 밀접한 관련이 있는 10개의 단어와 각 토픽에 해당하는 뉴스 수를 제시하고 있다. Topic 1에서부터 순서대로 주요한 토픽 순이며, 세로 방향으로는 시계열 순서이다. 해당 빈도는 도시 내 범죄에 대한 중요성을 평가하는 데 사용되지 않았으며, 추가적인 '토픽 시계열 분석'을 통해 결과를 분석하였다.

모든 연도의 토픽 1은 각각 '서울 법적 사건', '금전 범죄', '외국인 폭력 범죄', '택시 살인', '교통사고', '폭력', '절도', '살인', '성폭력', '영상 유출 성범죄', '외국인 범죄', '서울 범죄 증가'로 공통적으로 도시 내에서 발생하는 '폭력'과 같은 단어들이 주된 토픽을 구성하고 있으며, '살인', '절도', '외국인 범죄' 등과 같은 단어들이 그 뒤를 잇는 것으로 나타났다.

각 연도별 주요 토픽을 살펴보면, 2000-2001년의 토픽 2는 절도와 관련이 있었으며 토픽 4, 5는 성범죄와 관련된 범죄로 확인되었다. 구체적으로 토픽 2는 절도와 관련이 높은 '현금', '절도', '빼앗다' 등과 같은 단어들이 확인되었다. 이와 대비 되어 토픽 4, 5는 성범죄와 직접적인 관련성이 높은 '성폭행', '강간', '성관계' 등과 같은 단어들이 확인되었다. 추가적으로 토픽 4는 '남편', '이혼소송'

표 2. KoBERTopic 도출 결과
Table 2. KoBERTopic extraction results

Division	Count				
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
2000-2001	**서울 법적 사건(Seoul legal cases)** (서울 Seoul, 혐의 Charge, 만원 Ten thousand won, 경찰 Police, 경찰서 Police Station, 신청 Application, 구속영장 Arrest warrant, 늘다 Increase, 모 씨 Mr. Anonymous person, 지난 Last) [N=93]	**절도 범죄(Theft crimes)** (현금 Cash, 만원 Ten Thousand won, 절도 Theft, 경찰 Police, 빼앗다 Rob, 금품 Valuables, 혐의 Charge, 경찰서 Police Station, 강도 Robbery, 지난 Last) [N=89]	**의료기관 범죄 (Healthcare-related crimes)** (환자 Patient, 병원 Hospital, 의원 Clinic, 진료 Treatment, 폐업 Closure, 청문 Hearing, 간판 Signboard, 의사 Doctor, 지하 Basement, 되다 Become) [N=78]	**가정 내 성범죄 (Domestic sexual crimes)** (남편 Husband, 성폭행 Sexual Assault, 혐의 Charge, 말다 Stop, 강간 Rape, 서울 Seoul, 이혼소송 Divorce Lawsuit, 변태 Pervert, 여성 Woman, 흉기 Weapon [N=63]	**인터넷 범죄(Cyber sexual crimes)** (통해 Through, 인터넷 Internet, 채팅 Chat, 청소년 Youth, 비디오 Video, 성관계 Sexual intercourse, 혐의 Charge, 사이트 Site, 서울 Seoul, 복제 Copy) [N=38]
2002-2003	**금전 범죄(Monetary crimes)** (만원 Ten thousand won, 서울 Seoul, 혐의 Charge, 수사 Investigation, 늘다 Increase, 사건 Case, 검찰 Prosecution, 지난 Last, 특수 Special, 서울시 Seoul City) [N=92]	**가정 내 살인(Domestic homicide)** (숨다 Hide, 살인 Murder, 경찰 Police, 남편 Husband, 숨지다 Die, 늘다 Increase, 살해 Kill, 혐의 Charge, 박 씨 Mr. Park, 범행 Crime) [N=91]	**법률 범죄(Legal violations)** (검찰 Prosecution, 변호사 Lawyer, 늘다 Increase, 검사 Prosecutor, 사건 Case, 사회 Society, 수사 Investigation, 공짜 Free, 되다 Become, 보다 See) [N=75]	**금융 관련 범죄(Finance-related crimes)** (채권 Bond, 만원 Ten thousand won, 특권 Privilege, 절도 Theft, 혐의 Charge, 경찰 Police, 서울 Seoul, 침입 Intrusion, 경찰서 Police Station, 범행 Crime) [N=49]	**차량 범죄(Vehicle-related crimes)** (혐의 Charge, 경찰 Police, 차량 Vehicle, 서울 Seoul, 네거리 Intersection, 만원 Ten thousand won, 경찰서 Police Station, 구속영장 Arrest warrant, 춘추관 Press Room, 신청 Application) [N=44]

다음 페이지에 계속(Continue on next page)

Division	Count				
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
2004-2005	**외국인 폭력 범죄 (Violent crimes by foreign nationals)** (일본인 Japanese, 어린이 Child, 박 씨 Mr. Park, 학교 School, 일본 Japan, 경찰 Police, 둔기 Blunt Weapon, 남자 Man, 치료 Treatment, 때리다 Hit) [N=11]	n/a	n/a	n/a	n/a
2006-2007	**택시 살인(Taxi-related homicides)** (택시 Taxi, 경찰 Police, 살해 Murder, 서울 Seoul, 차량 Vehicle, 오전 Morning, 혐의 Charge, 승용차 Sedan, 범행 Crime, 조사 Investigation) [N=57]	**병원 치료 관련 의료 이슈(Hospital treatment-related medical issues)** (병원 Hospital, 치료 Treatment, 입원 Hospitalization, 환자 Patient, 암치료 Cancer Treatment, 늘다 Increase, 아버지 Father, 진단 Diagnosis, 의학 Medicine, 광혜원 Gwanghye Hospital) [N=26]	**폭행 및 법적 문제 (Assault and legal issues)** (회장 President, 경찰 Police, 폭행 Assault, 피해자 Victim, 수사 Investigation, 첩보 Intelligence, 서강대 Sogang University, 삼성 Samsung, 외대 Hankuk University of Foreign Studies, 사건 Case) [N=20]	**공공질서 사건(Public order incidents)** (경찰 Police, 차량 Vehicle, 혐의 Charge, 시위 Protest, 기아차 Kia Motors, 만원 Ten thousand won, 사고 Accident, 로체 Lotze, 되다 Become, 김재현 Kim Jae-hyun) [N=13]	n/a
2008-2009	**교통사고 (Traffic accidents)** (사고 Accident, 경찰 Police, 택시 Taxi, 내다 Happen, 강인 Strong, 서울 Seoul, 승용차 Sedan, 차량 Vehicle, 혐의 Charge, 운전자 Driver) [N=79]	n/a	n/a	n/a	n/a
2010-2011	**폭력(Violence)** (말다 Stop, 혐의 Charge, 경찰 Police, 때리다 Hit, 경찰서 Police Station, 취해 Drunk, 술값 Bar bill, 서울 Seoul, 늘다 Increase, 조사 Investigation) [N=128]	**화재(Fire)** (화재 Fire, 사고 Accident, 피해 Damage, 건물 Building, 소방 Firefighting, 방화 Arson, 지르다 Set fire, 오늘 Today, 병원 Hospital, 재산 Property [N=109]	**살인(Homicide)** (살해 Murder, 인력당 Manpower Party, 선고 Sentence, 살인 Murder, 시신 Corpse, 재판 Trial, 이유 Reason, 사건 Case, 현아 Hyun-a, 무기징역 Life imprisonment) [N=68]	**흉기 살인(Homicide with a weapon)** (살해 Murder, 흉기 Weapon, 찌르다 Stab, 여자친구 Girlfriend, 어머니 Mother, 살인 Murder, 생모 Birth mother, 늘다 Increase, 경찰 Police, 자수 Surrender) [N=63]	**택시 기사 폭행(Taxi driver assault)** (택시 Taxi, 기사 Driver, 박씨 Mr. Park, 경찰서 Police Station, 폭행 Assault, 요금 Fare, 승객 Passenger, 늘다 Increase, 경찰 Police, 서울 Seoul) [N=59]
2012-2013	**절도(Theft)** (경찰 Police, 절도 Theft, 혐의 Charge, 범행 Crime, 만원 Ten thousand won, 현금 Cash, 금품 Valuables, 경찰서 Police Station, 상당 Worth, 서울 Seoul) [N=430]	**폭행(Assault)** (말다 Stop, 경찰 Police, 폭행 Assault, 혐의 Charge, 취해 Drunk, 서울 Seoul, 조사 Investigation, 경찰서 Police Station, 구속 Arrest, 때리다 Hit) [N=223]	**프로포폴 남용(Propofol abuse)** (프로포폴 Propofol, 병원 Hospital, 투약 Administration, 장미인애 Jang Mi-in-ae, 시술 Procedure, 박시연 Park Si-yeon, 이승연 Lee Seung-yeon, 김지훈 Kim Ji-hoon, 치료 Treatment, 서울 Seoul) [N=205]	**연예인 사건(Celebrity cases)** (재판 Trial, 박시후 Park Si-hoo, 혐의 Charge, 변호사 Lawyer, 고영욱 Go Young-wook, 피해자 Victim, 류시원 Ryu Si-won, 대해 About, 기소 Indictment, 사건 Case) [N=143]	**화재(Fire)** (사고 Accident, 소방 Firefighting, 화재 Fire, 피해 Damage, 재산 Property, 원인 Cause, 연기 Smoke, 어젯밤 Last night, 불길 Flames, 대피 Evacuation) [N=94]

다음 페이지에 계속(Continue on next page)

Division	Count				
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
2014-2015	**살인(Homicide)** (살해 Murder, 경찰 Police, 범행 Crime, 늘다 Increase, 서초동 Seocho-dong, 살인 Murder, 사건 Case, 숨지다 Die, 조사 Investigation, 발견 Discovery) [N=324]	**화재(Fire)** (화재 Fire, 승객 Passenger, 도곡역 Dogok Station, 방화 Arson, 사고 Accident, 대피 Evacuation, 전동차 Subway train, 소방 Firefighting, 연기 Smoke, 지르다 Set fire) [N=143]	**폭행(Assault)** (차량 Vehicle, 운전 Driving, 택시 Taxi, 운전자 Driver, 보복 Retaliation, 경찰 Police, 사고 Accident, 늘다 Increase, 오전 Morning, 주차장 Parking lot) [N=95]	**절도(Theft)** (범행 Crime, 경찰 Police, 혐의 Charge, 현금 Cash, 금품 Valuables, 서울 Seoul, 강도 Robbery, 경찰서 Police Station, 절도 Theft, 만원 Ten thousand won) [N=93]	**외국인 범죄(Crimes committed by foreign nationals)** (서방 Seobang, 직원 Employee, 혐의 Charge, 경찰 Police, 조선족 Joseonjok, 살해 Murder, 청부 Contract killing, 브로커 Broker, 김태촌 Kim Tae-chon, 늘다 Increase) [N=77]
2016-2017	**성폭행(Sexual assault)** (성폭행 Sexual assault, 경찰 Police, 서울 Seoul, 오후 Afternoon, 경찰서 Police Station, 임세영 Lim Se-young, 개포동 Gaepo-dong, 식사 Meal, 고소장 Complaint, 배우 Actor) [N=25]	**음주운전(Drunk driving)** (단속 Crackdown, 경찰 Police, 강제추행 Forcible molestation, 음주운전 Drunk driving, 폭행 Assault, 김미나 Kim Mi-na, 음주 Drunk, 혐의 Charge, 경찰서 Police Station, 서울 Seoul) [N=14]	n/a	n/a	n/a
2018-2019	**영상 유출 성범죄(Sexual crimes involving non-consensual video distribution)** (양상 Situation, 신림동 Sillim-dong, 여성 Woman, 경찰 Police, 강간미수 Attempted rape, 남성 Man, 촬영 Filming, 들어가다 Enter, 체포 Arrest, 공개 Publicize) [N=78]	**살인 사건(Homicide cases)** (살해 Murder, 경찰 Police, 흉기 Weapon, 살인 Murder, 범행 Crime, 숨지다 Die, 늘다 Increase, 가정폭력 Domestic violence, 이혼 Divorce, 피해자 Victim) [N=75]	**프로포폴 남용(Propofol abuse)** (병원 Hospital, 환자 Patient, 프로포폴 Propofol, 구하라 Goo Hara, 메르스 MERS, 의료 Medical, 감염 Infection, 성형 Plastic surgery, 치료 Treatment, 경찰 Police) [N=70]	**경찰 여성 제압(Police restraint of woman)** (여경 Female police officer, 경찰관 Police officer, 여성 Woman, 경찰 Police, 제압 Subdue, 남성 Man, 영상 Video, 대응 Response, 시민 Citizen, 논란 Controversy) [N=68]	**연예인 폭행 범죄 (Assault crimes against celebrities)** (구하라 Goo Hara, 폭행 Assault, 남자친구 Boyfriend, 경찰 Police, 여성 Woman, 남성 Man, 논현동 Nonhyeon-dong, 경찰서 Police Station, 카라 Kara, 신고 Report) [N=66]
2020-2021	**외국인 범죄(Crimes committed by foreign nationals)** (대사 Ambassador, 벨기에 Belgium, 부인 Wife, 옷가게 Clothing Store, 레스쿠이 L'Escuyer, 직원 Employee, 폭행 Assault, 때리다 Hit, 환경미화원 Street Cleaner, 용산구 Yongsan-gu) [N=23]	n/a	n/a	n/a	n/a
2022-2023	**서울 범죄 증가(Increase in Seoul crimes)** (서울 Seoul, 늘다 Increase, 경찰 Police, 범행 Crime, 피해자 Victim, 혐의 Charge, 사건 Case, 되다 Become, 살해 Murder, 납치 Kidnapping) [N=41]	n/a	n/a	n/a	n/a

이 토픽 5는 '인터넷 청소년', '복제'로 성범죄 내에서도 가정 내 성범죄와 인터넷 성범죄에 대한 세부 주제로 구분되어 나타났다.

2002-2003년 토픽 1, 4는 주로 절도, 금융 범죄와 관련된 토픽이 도출되었다. 이에 따라 '절도', '채권', '수사', '특수'와 같은 단어들이 나타났다. 추가적으로 토픽 2에서는 '살인', '남편', '살해'와 같은 단어들이 도출되었으며 이에 따라 살인 토픽이 나타났다. 2004-2005년에는 외국인 폭력 범죄 토픽 하나가 도출되었다. '일본인', '어린이', '둔기', '때리다'와 같은 단어들이 도출되었으며, 이는 2003년 9,338건이었던 외국인 범죄 수가 2004년 12,821건으로 증가하면서 외국인 범죄가 증가한 결과와 일치하며 외국인 범죄에 대한 심각성을 시사한다.

2006-2007년 살인, 폭행에 대한 범죄가 토픽 1, 3에서 나타났다. 각각 '살해', '혐의', '택시와 폭행', '피해자', '수사'와 같은 단어들이 확인되었다. 2008-2009년에는 '택시', '사고', '혐의'와 같은 단어들이 나타나면서 교통사고 토픽이 도출되었다. 전체적으로 2000년대에서는 절도에 대한 범죄가 많이 발생하였으며 '폭력', '살인', '성범죄'가 그 뒤를 잇는 것을 알 수 있다.

2010-2019년에는 전체적으로 다양한 도시 내 범죄가 발생하였다. 2010-2011년 토픽 1에서 토픽 4까지 서로 다른 도시 범죄가 확인되었다. 각각 '폭력', '화재, '살인', '흉기 살인'과 같은 토픽들이 나타났으며 토픽 5는 토픽 1과 같은 폭력 범죄로 나타났다. 2012-2013년에는 토픽 1과 토픽 2는 '절도'와 '폭력'이 도출되었지만, 도시 내 범죄 중 '마약'이 토픽 3에 새롭게 도출되었다. 토픽 3은 마약과 관련이 높은 '프로포폴', '병원', '투약'과 같은 단어들이 도출되었으며, 이는 마약범죄에 대한 범죄 예방이 필요함을 시사한다.

2014-2015년의 토픽 1은 '살인'으로 나타났으며 '화재', '폭행', '절도', '외국인 범죄'가 그 뒤를 이었다. 토픽 1에서는 '살해', '살인', '숨지다'와 같은 단어들이 도출되었으며 서초동이 도출되었는데 이는 살인 중에서도 2015년에 발생한 '서초동 세 모녀 일가족 살인사건'이 가장 큰 사건이었음을 암시한다. 추가적으로 각각 '화재', '방화', '지르다', '보복', '택시', '금품, '강도', '조선족', '청부' 등과 같은 단어들이 도출되었다. 2016-2017년 성폭행, 음주운전과 관련된 범죄 이슈가 나타났다. 각각 토픽 1에서는 '성폭행', '개포동', '고소장' 등과 같은 단어들이 토픽 2에서는 '단속', '음주', '강제추행' 등의 단어들이 도출되었다.

2018-2019년에는 다시 성범죄, 살인, 마약, 폭행과 같은 다양한 도시 내 범죄 이슈가 나타났다. 이에 따라 토픽 1은 '신림동', '강간미수', '촬영' 등과 같은 단어가 토픽 2에서는 '살해', '흉기', '숨지다', 토픽 3에서는 '프로포폴', '병원', '성형', 토픽 5에서는 '폭행', '신고', '남자친구'와 같은 단어들이 도출되었다. 전체적으로 2010년도에는 도시 내 다양한 범죄가 발생하였으며 각 시대에 따라 주요 범죄가 변화하는 것으로 나타났다. 이는 형범 범죄가 꾸준히 발생하고 있으며 큰 변동이 나타나지 않는 〈그림 1〉의 결과와 일치하며 형법 범죄에 대한 범죄 예방이 필요함을 시사한다.

2020-2021년에는 하나의 토픽이 도출되었으며 토픽 1에서 '폭행', '벨기에', '직원', '때리다' 등과 같은 단어들이 나타났으며, 이에 따라 외국인 범죄 토픽이 도출되었다. 이는 COVID-19로 인한 '사회적 거리두기'로 외부활동이 감소하게 되면서 전체적인 범죄 수가 감소한 것으로 볼 수 있다. 하지만 2022-2023년에는 서울 범죄 증가 토픽이 도출되었으며 '늘다', '범행', '피해자', '혐의'와 같은 단어들이 나타났다. 이는 '사회적 거리두기'가 해제됨에 따라 외부 활동이 증가하면서 도시 범죄 증가로 인한 결과로 예상해 볼 수 있다.

전체적으로 2000년대에는 특정 범죄가 많이 발생하였지만 2010년대에 들어서면서 다양한 범죄 유형이 꾸준히 발생하는 것으로 나타났다. 하지만 2020년대에는 COVID-19를 기점으로 범죄 발생이 감소하였지만, 2022년 '사회적 거리두기' 해제되며 일상활동이 복구됨에 따라 다시 서울시 범죄가 증가하는 것으로 나타났다. 이는 〈그림 1〉의 전체적으로 형법범죄의 변화가 나타나지 않았으며 2020년에서 2021년으로 전체범죄 및 형법범죄가 감소, 2022년 범죄가 다시 증가하는 결과와 일치한다. 따라서 서울시에서의 범죄 발생 양상이 시대별로 변화하고 있음을 고려할 때, 범죄 유형에 따른 맞춤형 범죄 예방 및 대응 전략을 강화하는 것이 필요함을 시사한다. 이를 통해 범죄 발생을 사전에 예방하고, 범죄 발생 시 신속하고 효과적으로 대응할 수 있는 체계를 구축함으로써 시민들의 안전을 보다 확실하게 보장할 수 있다.

2) Topics over time 시각화

BERTopic의 Topics over time을 기반으로 시계열로 정리한 Topic 변화는 다음 〈그림 4〉와 같다. 해당 분석은 전체 시간적 범위 내 공통적으로 존재하는 범죄 토픽의 시계열 변화를 확인할 수 있는 기법이다. 본 연구는 전체 시기 내 지속적으로 존재한 범죄 토픽에 대하여 그 시계열 변화를 확인하기 위하여 Topics over time을 진행하였으며, 시각화는 연구의 전체 시간적 범위에 대해 상위 10개 토픽에 대해서 진행하였다. 전반적으로 토픽들은 2000년에는 빈도가 적었으나, 2023년으로 갈수록 높은 빈도를 보였다.

먼저 Topic 1(금융 범죄)의 경우, 2000년 초반 증가하다 2008년 경제 위치를 기점으로 크게 증가하는 것으로 나타났으며, 이후 COVID-19를 기점으로 크게 감소하는 것으로 나타났다. 매년 꾸준히 빈도를 보이다 2015년 초반에 급증한 것으로 나타났으며, 그 이후에도 2015년에 높은 빈도를 보이며 한 해 동안 지속적으로 발생한 것으로 나타난다. Topic 2(살인)의 경우, 2004년에는 굉장히 낮은 빈도를 보이다 2010년부터 차츰 증가하다 2016년 최대 빈도를 기록하였다. 이후 범죄의 빈도가 감소하다, 2020년을 기점으로 크게 감소하는것으로 나타났다. 이는 살인에 대한 범죄 예방이 꾸준히 진행되면서 범죄율이 감소한 것으로 예상할 수 있다.

Topic 3(폭력)의 경우 2000년 초반 크게 증가하는 경향을 보이

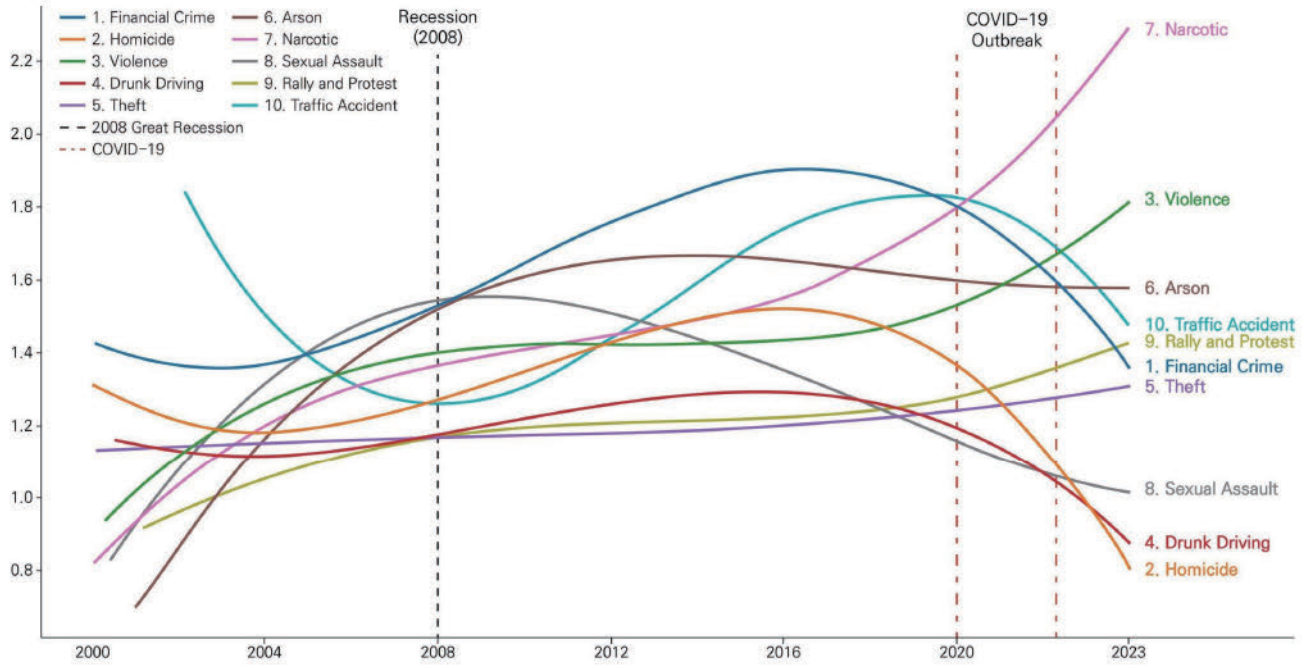


그림 4. 연도별 Topic 변화
Figure 4. Yearly topic changes

다 2008년부터 2016년까지는 크게 변화가 없는 것으로 나타났다. 하지만 2020년 크게 증가하는 모습이 나타났으며 2023년 폭력 범죄의 빈도가 가장 높게 나타났다. 이는 매년 꾸준히 높은 수치를 기록하며 폭력 범죄의 심각성을 확인할 수 있다. Topic 4(교통사고 및 음주운전), Topic 10(교통사고)는 감소하는 추세를 보이다 2010년도 증가하는 것으로 나타났다. 하지만, COVID-19를 기점으로 크게 감소하는 것으로 나타났는데, 이는 〈그림 1〉의 특별법범죄가 꾸준히 감소하고 있다는 결과와 일치한다. 결과적으로 교통사고에 대한 범죄 예방이 효과가 있는 것을 알 수 있으며, '사회적 거리두기'로 인해 외부 활동이 감소하면서 교통사고 범죄가 감소한 것으로 판단된다.

Topic 5(절도)는 2000년부터 2016년까지 전반적으로 큰 변화가 나타나지 않았지만, COVID-19를 기점으로 점차 증가하는 것으로 나타났으며 2023년 높은 빈도를 보였다. Topic 6(화재)은 꾸준히 증가하는 추세를 나타내다 2008년 가장 높은 빈도를 나타냈다. 이는 2008년 2월에 발생한 '숭례문 방화 사건'과 더불어 '고시원 방화'와 같은 화재 사건으로 나타난 결과로 예상할 수 있다.

Topic 7(마약)은 2000년대부터 증가하는 추세가 나타났으며, 2018년과 COVID-19 발생 시기에 가장 크게 증가한 것으로 나타났다. 이는 2018년부터 2022년까지 마약 범죄가 꾸준히 증가한다는 대검찰청(2023b)의 '2023 Drug Control in Korea' 결과와 일치하며, 마약 범죄에 대한 예방이 필요함을 시사한다. Topic 8(성범죄)는 2000년부터 증가하는 추세를 보이다 2009년을 기점으로 꾸준히 감소하는 것으로 나타났으며, Topic 9(집회 및 시위)는 다른 Topic 범죄 유형과는 다르게 큰 변화를 나타내지 않는 것으로 나타났다.

2. 범죄 공간분포 시각화

1) 범죄 시각화

범죄의 공간 분포를 확인하기 위해 전체범죄의 공간적, 시간적 범위에서 수집한 데이터를 시각화하였다(그림 5). 이를 통해, 범죄 패턴 및 추세를 식별할 수 있으며 범죄의 군집화, 무작위화 분산형으로 범죄율의 공간적 분포를 확인하였다. 각 연도마다 수집된 기사와 범죄 발생 건수가 다르기 때문에 각각 상위 10% 이상, 10%~20%, 20~30%, 30~40%, 40% 미만 총 5개의 범례로 분류하여 진행하였다.

뉴스 기사 중 일부는 가락1동, 가락2동, 논현1동, 논현2동 등과 같이 분리된 행정동이 아닌 가락동, 논현동 등과 같이 하나의 행정동으로 위치를 제공하고 있다. 따라서 2023년 기준 서울시 426개 행정동에 대해 전처리를 진행하였다. 최종적으로 245개의 행정동을 도출하였으며 시각화는 ArcGIS Pro를 사용하였다. 분석 결과 2000년부터 2023년까지 범죄율은 특정 지역에서 점차 증가했으며, 범죄의 지리적 분포는 시간이 지남에 따라 변화하고 주변 지역으로 확장하는 것으로 나타났다.

2) 서울시 내 범죄 공간 분포

〈그림 5〉를 살펴보게 되면 색이 진할수록 해당 연도의 범죄 발생이 높은 지역을 나타내며 색이 연할수록 범죄 발생이 낮은 지역을 나타낸다. 범죄의 추세를 살펴보게 되면 강남, 한남, 홍대, 신림, 성북, 장안을 중심으로 범죄가 집중되는 것으로 나타났다.

2000-2001년에는 서울 중심부의 일부 지역에서 주로 범죄가 많이 발생하는 것으로 나타났다. 특히 명동, 을지로, 강남, 서초

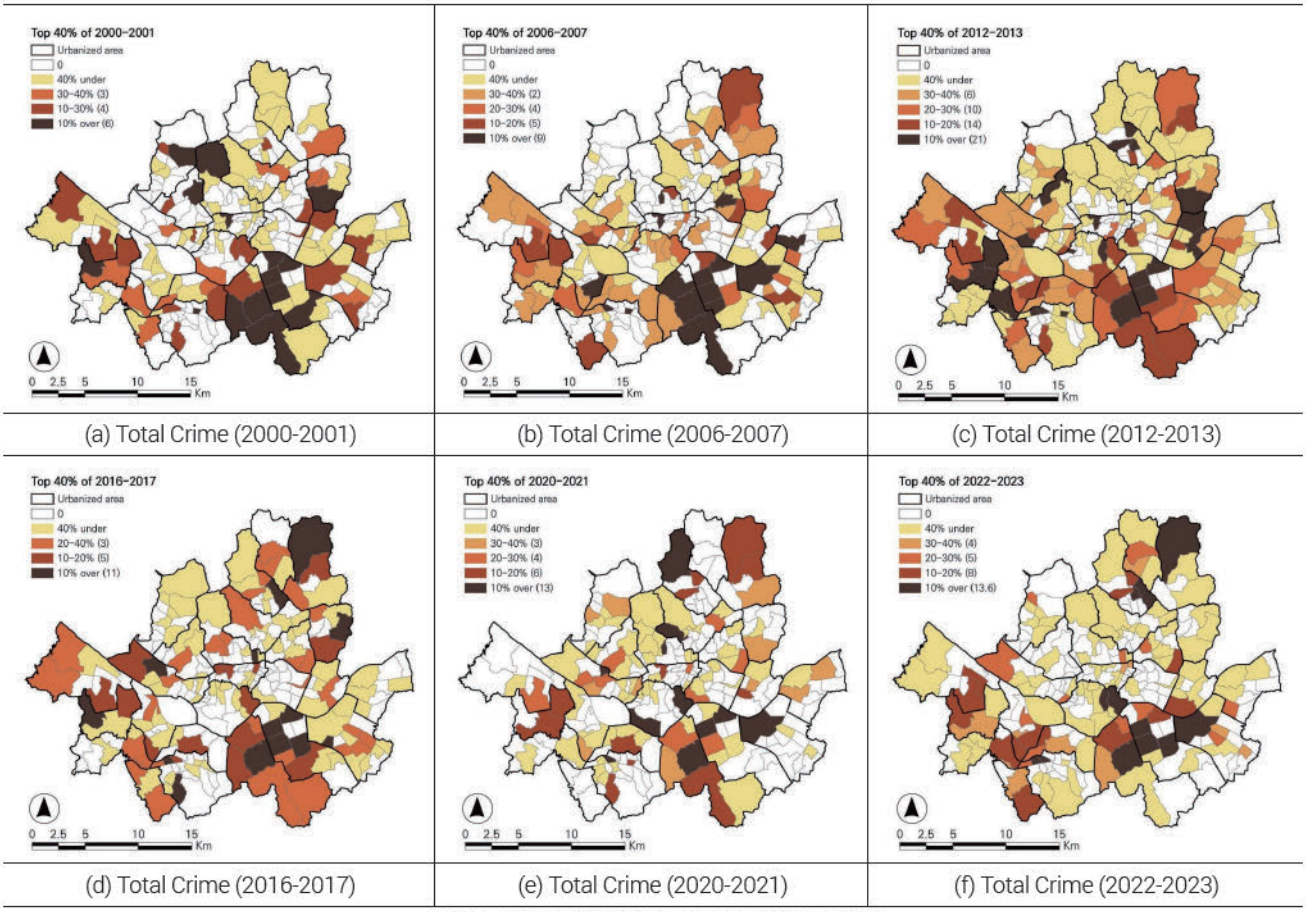


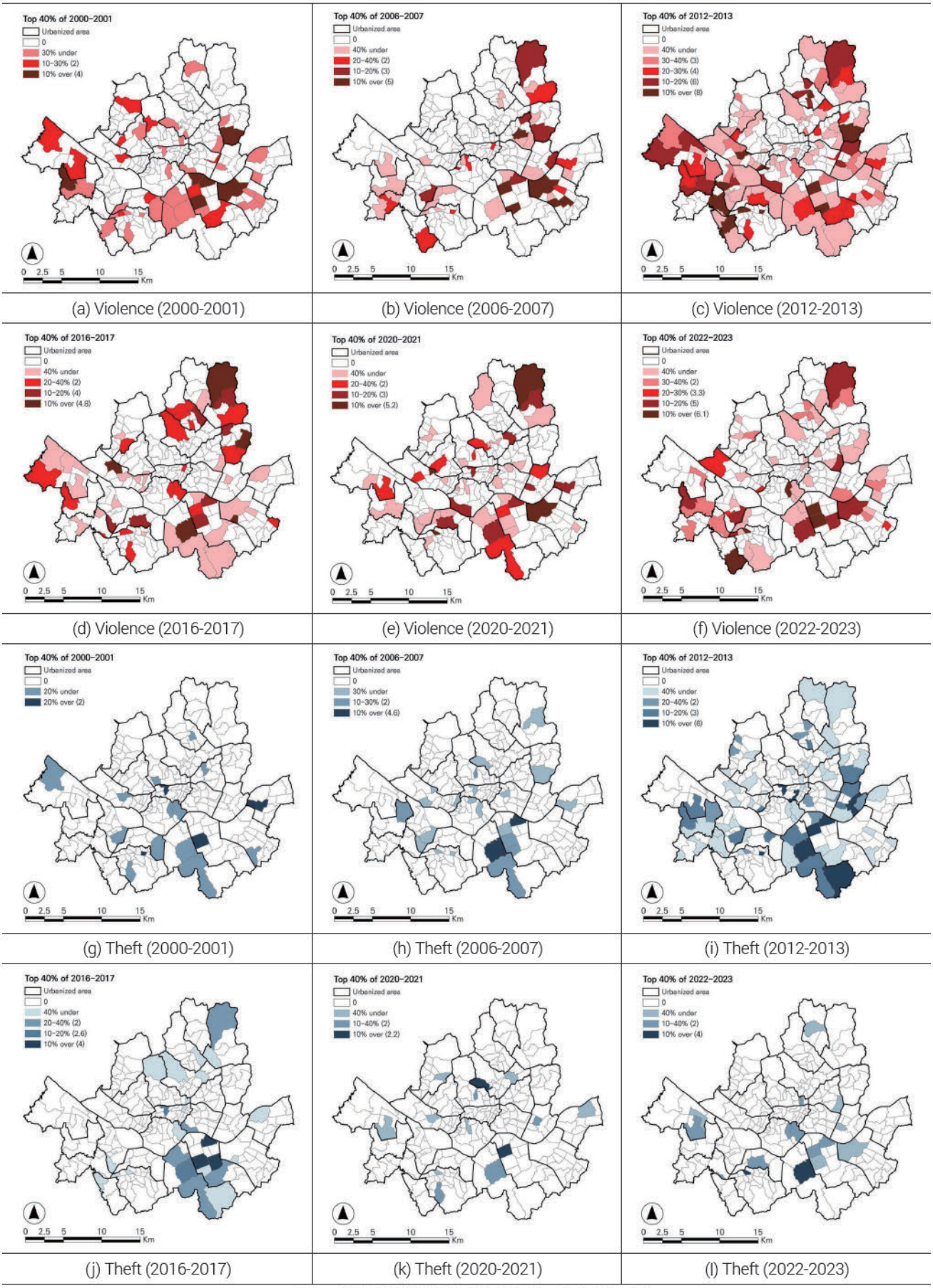
그림 5. 서울시 행정동별 연도별 전체범죄 건수 비율
Figure 5. Annual percent of total crimes in administrative dongs in Seoul

에서 범죄 발생이 집중되어 있는 것으로 나타났다. 이는 서울 중심부의 상업 밀집 지역이며, 인구 밀도와 활동량이 높은 지역으로 범죄 발생이 높은 것으로 예상할 수 있다. 2006-2007년에는 전체적인 범죄가 중심부에서 남서쪽, 남동쪽으로 확산되기 시작했으며 송파구, 동작구, 노원구 등 주변 지역으로 확산되는 것으로 나타났다. 하지만 강남과 서초에 대한 범죄의 변화는 크게 나타나지 않았으며 범죄가 기존의 높은 범죄율 지역에서 주변으로 확산되는 것을 알 수 있다. 2012-2013년에는 범죄율이 서울 전역으로 확산하는 것으로 나타났으며, 특히 송파구, 종로구, 영등포구 등 지역의 범죄 분포가 확산된 것으로 나타났다. 이는 〈그림 4〉의 2012년 전반적으로 주요 토픽들이 증가하는 결과와 일치한다. 2016-2017년에는 범죄 발생이 더욱 다양한 지역으로 확산되었다. 강남과 중구, 종로구를 포함한 중심부에는 여전히 높은 범죄 발생률을 유지하고 있으며 이외에도 강동구, 서대문구, 은평구 등의 지역에서도 범죄율이 상승하는 것으로 나타났다. 이는 서울시 전역에서 다양한 경제적 및 사회적 활동이 확대되면서 범죄의 지리적 분포가 넓어진 것으로 예상할 수 있다. 2020-2021년에는 범죄가 전반적으로 감소하는 것으로 나타났는데, 이는 COVID-19로 인해 사회적 활동이 제한되면서 범죄 발생률이 감소한 결과로 볼 수 있다. 그러나 강남구, 서초구, 송파구 등 일부 지역에서는 여전히 상대적으로 높은 범죄율을 유지하는 것으로 나타났다. 2022-2023년에는 다시 서울시 전역의 범죄가 증가하는 경향을 나타냈다. 이는 '사회적 거리두기' 해제에 따른 경제 회복과 사회적 활동이 재개되면서 범죄율이 다시 증가하고 있는 것으로 해석할 수 있다.

종합적으로 범죄 발생은 초기 강남, 중구와 같은 서울의 중심부와 외각지역에 집중되어 있었으나 시간이 흐름에 따라 주변 지역으로 분산되는 것으로 나타났으며, 최종적으로 서울시 전역으로 범죄 발생이 증가한 것으로 나타났다. 또한, COVID-19와 같은 사회적 이슈로 인해 범죄가 일부분 감소하는 것으로 나타났다. 이는 범죄가 특성 지역에 집중되는 경향을 보이지만 주변 지역에도 영향을 미치며 범위가 증가하는 것으로 해석할 수 있다.

3) 범죄 유형별 공간 분포

범죄 유형별 공간 분포 분석에서는 폭력 범죄와 절도 범죄를 중심으로 시각화를 진행하였다. 이는 살인 등 기타 강력 범죄의 경우 발생 빈도가 상대적으로 낮아, 공간 분포 비교에 한계가 있기 때문이다. 따라서 본 연구에서는 모든 범죄 유형을 대상으로 하기보다는, 비교적 빈번하게 발생하여 공간적 분석이 가능한 폭력 및 절도 범죄에 한정하여 시각화를 수행하였으며, 그 결과는 〈그림 6〉과



(a) Violence (2000-2001) (b) Violence (2006-2007) (c) Violence (2012-2013)

(d) Violence (2016-2017) (e) Violence (2020-2021) (f) Violence (2022-2023)

(g) Theft (2000-2001) (h) Theft (2006-2007) (i) Theft (2012-2013)

(j) Theft (2016-2017) (k) Theft (2020-2021) (l) Theft (2022-2023)

그림 6. 서울시 행정동별 연도별 폭력, 절도 범죄 건수 비율

Figure 6. Annual percent of violent and theft crimes in administrative dongs in Seou

같다. 먼저 2000-2001년 폭력 범죄는 서울의 중심부와 강남 지역에 집중된 것으로 나타났으며, 특히 중구, 강남구, 송파구에서 높은 범죄율이 나타났다. 2006-2007년에는 전체적인 폭력 범죄 발생 지역이 감소하였지만 인구 밀도가 높으며 활동량이 높은 강남과 송파로 범죄가 집중되는 것으로 나타났다. 추가적으로 노원구와 구로구 지역에서 새로운 폭력 범죄가 발생한 것으로 나타났다. 2012-2013년에 들어서면서 기존의 서울 중심 구역에서 벗어나 서울 외곽 지역에서도 범죄율이 증가한 것으로 나타났다. 특히 강서구, 서초구, 영등포구, 금천구 등에서 눈에 띄게 증가하였으며 이는 경제 활동의 증가와 상업지역의 확장이 주된 원인으로 예상할 수 있다. 2016-2017년에는 주변으로 확산되었던 범죄가 다시 중심 지역으로 밀집하는 것으로 나타났다. COVID-19 시기인 2020-2021년에는 전체범죄와 달리 폭력 범죄의 분포가 축소되지는 않았지만 서울의 외곽인 양재동과 상계동 등에서 높은 범죄의 발생률을 나타냈으며 강남과 영등포구의 범죄는 과거에 비해 범죄 발생률이 감소하는 것으로 나타났다. 이후 2022-2023년 다시 서울의 중심인 강남과 서초구, 영등포구의 범죄 발생률이 증가하는 것으로 나타났다. 이는 COVID-19 이후 일상 활동이 복구됨에 따라 다시 인구 이동과 사회적 활동이 활발한 지역으로 범죄가 이동한 것으로 볼 수 있다.

마지막으로 서울시에서 발생한 절도 범죄를 살펴보게 되면, 2000-2001년 서울의 중심부 강남, 명동, 을지로, 서초 등과 같은 지역에서 범죄 발생률이 높게 나타났다. 이는 전체범죄, 폭력과 마찬가지로 인구 밀도가 높으며 다양한 활동이 이루어지는 지역이다. 2006-2007년에는 범죄의 집중도가 이전보다 약간 넓어진 것으로 나타났으며, 중심 상업지역에서는 여전히 높은 범죄율을 보였다. 2012-2013년으로 넘어오게 되면서 서울 중심부(강남, 서초)의 범죄 발생률이 증가했으며, 특히 중구와 종로구, 양천구 등에서 높은 범죄율과 서울의 외곽 지역에서도 범죄 발생률이 증가한 것으로 나타났다. 이는 지속적인 상업의 성장과 더불어 높은 인구 유동성으로 인한 결과로 볼 수 있다. 2016-2017년에는 서울의 전체적인 범죄가 감소하면서 다시 서울의 중심지역으로 밀집하는 것으로 나타났다. 이는 단속과 CCTV와 같은 보안 시설의 확장으로 인해 범죄가 감소한 것으로 해석할 수 있지만, 여전히 강남과 서초구와 같은 지역은 높은 범죄율을 나타내고 있다. 2020-2021년에는 COVID-19로 인해 전반적인 인구 유동 및 외부 활동 감소로 인해 절도 범죄가 감소한 것으로 나타났으며, 서울 중심부의 범죄율도 감소한 것으로 나타났다. 2022-2023년에는 전체범죄, 폭력과 다르게 '사회적 거리두기' 해제에 따른 범죄율의 증가가 나타나지 않았으며 일부 지역에서 소폭 증가하는 것으로 나타났다. 이를 통해 서울시의 절도 범죄율이 안정화되었음을 알 수 있다.

전반적으로 서울에서 발생하는 폭력과 절도 범죄는 유동 인구가 많으며 사회적 활동이 활발한 지역에서 주로 발행하는 것으로 나타났다. 시계열적인 측면에서 살펴보면, 절도 범죄의 경우 2012-2013년까지는 대체로 증가하는 양상을 보이다가 2016-2017년부터는 점차 감소하는 추세를 나타내고 있다. 폭력 범죄 역시 2012-2013년까지는 증가하는 양상을 보였으나, 절도 범죄와는 달리 이후 뚜렷한 감소 추세는 보이지 않았다. 또한, COVID-19로 인해 범죄율이 일부분 감소하는 것으로 나타났지만, 범죄 유형별로 시간이 흐름에 따라 변화가 상이한 것으로 나타났다. 이는 범죄 유형별로 사회적 이슈에 따른 분석이 필요하며 안전한 도시를 위한 예방 정책이 필요함을 시사한다.

V. 결 론

본 연구는 서울시에서 발생한 범죄 기사를 대상으로 텍스트 마이닝을 진행하여 범죄 토픽을 도출하였다. 이후 범죄의 시계열 분석과 시각화를 통해 시공간적으로 범죄의 동향을 살펴보았으며 이에 따른 주요 결과는 다음과 같다.

첫째, 본 연구는 2000년부터 2023년까지 총 24년간의 시계열 범죄 기사 대상으로 텍스트 분석을 진행하였다. 뉴스는 일상생활에서 발생하는 주요 범죄에 대한 정보를 포함하고 있다는 점에서 중요하며, 이를 바탕으로 서울시에서 발생하는 범죄 이슈를 도출하여 도시 범죄 변화분석을 진행하였다. 또한, 뉴스 데이터는 범죄 발생 양상을 파악할 수 있을 뿐만 아니라, 도시민의 범죄 인식에 직접 영향을 미치는 역할로 작용하기도 한다(이완수·송상근, 2020). 이러한 맥락에서 본 연구의 분석 결과는 뉴스 보도 자료를 기반으로 도시민이 체감하는 범죄 인식 및 공간적 분석 연구의 기반이 될 수 있다. 본 연구는 장기 시계열 데이터를 구축하였으며, 과거에서 현재까지 행정동 수준의 공간단위에서 범죄 유형별 발생 건수를 도출하고 연도별 범죄토픽의 변화를 분석하였다는 점에서 의의가 있다.

둘째, 본 연구는 가장 우수한 성능을 가진 딥러닝 기반 토픽모델링 기법인 KoBERTopic을 사용하여 주요 토픽을 도출하였다. 구조화되지 않은 뉴스 텍스트에서 각 연도마다의 최적의 토픽 개수를 도출하여 토픽모델링을 진행하였다. 이후, LLM 모델을 적용하여 토픽명을 도출하였으며, 이러한 일련의 과정들을 통해 시대에 따른 범죄 비교 분석하는 과정에서 토픽 개수에서의 객관성을 확보할 뿐만 아니라 토픽의 중요도 변화를 확인하였다.

셋째, 전체 시간적 범위에서 범죄 기사별 토픽을 도출하여 시간적 범죄의 동향을 10가지 유형으로 분류하였다. 전체적으로 마약, 폭력, 절도, 교통사고, 화재, 살인 등과 같은 범죄 토픽이 도출되었다. 서울시에서 발생하는 교통사고, 금융범죄, 방화 등의 범죄는 2016년 이후로 꾸준히 감소하였지만, 마약, 폭력, 절도 등의 범죄는 증가하는 것으로 나타났다. 이는 대검찰청(2023a)의 결과와 일치하는 것을 알 수 있으며, 시간의 흐름에 따라 형법범죄의 범죄율이 증가하였으며 범죄 유형별로 변화가 상이한 것으

로 나타났다. 이중 절도의 경우 COVID-19기점으로 크게 증가하는 것으로 나타났다. 이는 '사회적 거리두기'로 시행으로 거리의 사람들이 줄어들게 되며 절도 범죄가 증가한 것으로 해석할 수 있으며, Jacobs (1961)의 '거리의 눈' 효과 감소와 관련이 있을 것으로 판단된다. 이를 통해 주요 범죄 유형 분석과 심각성을 확인할 수 있으며, 이는 시대별 주요 범죄에 대한 예방 정책을 시행하는 데 중요한 근거로 활용될 수 있을 것으로 판단된다.

마지막으로, 공간적 위치가 결합 되지 않은 텍스트 기반 빅데이터를 공간적 단위로 가공하는 방법론적 프레임워크를 제시하였다. 이를 통해 과거부터 현재까지 범죄의 공간적 분포를 확인하였다. 범죄는 강남, 한남, 홍대, 신림과 같은 지역을 중심으로 집중된 것으로 나타났으며, 이는 인구 밀집 지역 또는 도시 활력이 높은 지역에서 범죄 발생 확률이 높다는 결과와 맥락을 같이한다(김선재 외, 2022; Jiang et al, 2023).

2005년에는 범죄 발생이 집중된 지역이 감소하며 분포가 넓어졌지만, 2015년에 들어서면서 범죄 집중 지역이 증가하였으며 범죄 비율이 동시다발적으로 증가하였다. 이와 대비되어 2020년도에는 금융범죄, 교통사고, 화재와 같은 특별법범죄 발생 분포가 감소하였지만, 2023년 전체적으로 범죄 발생의 분포가 증가하는 것으로 나타났다. 이는 사회적 이슈였던 '사회적 거리두기'로 인해 외부활동이 감소하면서 일시적으로 범죄가 감소하는 것으로 볼 수 있다. 하지만 강남과 한남, 신림과 같은 범죄 집중 지역은 큰 변화를 나타내지 않으며 높은 범죄 발생 비율을 보이는 것으로 나타났다. 이는 '환경범죄학', '합리적 선택' 이론에 따라 장소와 장소에 대한 이미지가 범죄 발생에 중요한 영향을 미치는 것으로 해석할 수 있으며(Brantingham and Brantingham, 1981; Cornish and Clarke, 1989), 이러한 도시 내 범죄 발생은 주민들의 불안감을 증가시키는 동시에 주민들의 삶의 질을 저하시키는 요인으로 작용할 수 있다. 따라서 본 연구의 분석 결과를 바탕으로 강남, 한남, 신림과 같은 범죄 다발 지역의 경우, 장소 이미지 개선을 통한 범죄 예방 전략이 필요하다고 판단된다. 또한, 범죄는 범죄 집중 지역 주변 또는 범죄 예방 지역 주변으로 범죄가 확산되는 '풍선효과(Balloon Effect)'의 특징을 가지고 있으므로(Gilad et al, 2019) 범죄 발생 지역뿐만이 아닌 주변 지역을 포함하여 범죄 유형별로 범죄 예방 효과가 있는 방범시설을 배치하여 전체적인 범죄발생 저감시킬 필요가 있다. 이와 더불어 서울시 도시 데이터를 기반으로 실시간 인구 밀집 지역을 확인하여 범죄 유형별 강화된 모니터링 전략이 필요하다. 이처럼 범죄 발생 집중지역에 대한 범죄 예방 시설 배치와 방범 조치를 실시하여 이에 따른 범죄의 변화에 대한 모니터링을 진행함으로써 범죄로부터 안전한 도시를 만드는 데 기여할 수 있다.

본 연구는 한계점은 다음과 같다. 첫째, 뉴스 기사는 모든 범죄 발생에 대하여 제공하고 있지 않으며 중복된 사건에 대한 보도가 발생할 수 있다. 이는 정확한 범죄 발생을 확인하기에는 어려움이 있다. 하지만 전반적인 범죄 발생과의 관계를 파악할 수 있으며, 이는 주민, 경찰, 연구자 등 일상생활 또는 방범, 연구를 진행하는 데 도움을 줄 수 있다. 둘째, 시간 및 데이터의 가용성으로 인해 제한된 데이터셋을 활용하였다. 본 연구는 총 1,921,670개의 범죄 기사를 수집하였으며 전처리 과정을 통해 37,360개의 기사를 추출하였다. 해당 데이터는 특정 뉴스 사이트에서 추출되었기 때문에 범죄 발생 건수에 불확실성이 있을 수 있으며, 전처리 과정에서 기사의 중복 및 누락이 있을 수 있다. 이는 향후, 블로그, Twitter, Facebook 등과 같은 다양한 텍스트 데이터를 수집하여 분석할 필요성이 있으며, 추가적으로 범죄 키워드에 따른 감정분석을 통해 범죄 발생과 지역에 대해 지역 주민들이 느끼는 두려움에 대한 분석이 진행될 필요가 있다. 셋째, 〈표 2〉의 폭행과 〈그림 4〉의 교통사고같이 각각 '폭력', '폭행', '교통사고 및 음주운전', '교통사고'에 대한 키워드를 도출하는 과정에서 '폭행', '교통사고'라는 키워드가 아닌 별개로 집계되어 분석되는 한계가 있다. 이는 향후, 단어의 형태는 다르지만 본질적인 의미가 일치하는 단어들을 동일한 단어로 처리하는 과정을 거쳐 단어에 대한 표준화를 진행할 필요가 있음을 시사한다. 마지막으로, 범죄 발생 시간대인 주간, 야간 등을 세분화하여 분석하는 데 한계가 있었다. 범죄는 시간대에 따라 상이한 발생 패턴을 보이나, 본 연구에서는 구체적인 시간대 정보를 수집하는 데 제약이 있어 시간대별 분석이 이루어지지 못하였다. 이는 향후 연구에서 정량 데이터 보완 및 텍스트 데이터 전처리 과정의 개선을 통해 '오전', '밤', '새벽' 등과 같은 시간 관련 키워드를 추출함으로써 데이터의 해상도를 높이고 보다 정밀한 분석을 수행할 필요성을 시사한다.

주1. 연도의 변화에 따라 각 연도별 행정구역의 변화로 인해 2000-2023년의 모든 행정동을 뉴스기사에서 사용하는 행정동 단위로 변환 진행하였다.

인용문헌
References

1. 김선재·조월·이수기, 2022. "도시환경 특성과 범죄발생의 연관성 분석: 도시 빅데이터와 공간더빈 모형을 활용하여", 「도시설계」, 23(3): 143-162.
   Kim, S., Cao, Y., and Lee, S., 2022. "Analysis of the Association between Urban Environmental Characteristics and Crime Incidence: Using Urban Big Data and Spatial Durbin Model", *Journal of The Urban Design Insitute of Korea Urban Design*, 23(3): 143-162.
2. 김혜빈·이수기, 2024. "COVID-19 전·후 우울감 완화를 위한 도시공간 활용의 변화 분석: 소셜미디어 빅데이터와 KoBERTopic 모형의 적용", 「도시설계」, 25(2): 109-125.
   Kim, H. and Lee, S., 2024. "Analysis of Changes in Urban

Space Usage for Alleviating Depression Before and After COVID-19: Application of KoBERTopic Model with Social Media Bigdata", *Journal of The Urban Design Insitute of Korea Urban Design*, 25(2): 109-125.

3. 대검찰청, 2023a. 「2023 범죄분석」.
Supreme Prosecutors' Office, 2023. *2023 Statistcal Analysis on Crime*.

4. 대검찰청, 2023b. 「2023 Drug Control in Korea」.
Supreme Prosecutors' Office, 2023. *2023 Drug Control in Korea*.

5. 박대민, 2016. "장기 시계열 내용 분석을 위한 뉴스 빅데이터 분석의 활용 가능성: 100만 건 기사의 정보원과 주제로 본 신문 26년", 「한국언론학보」, 60(5): 353-407.
Park, D., 2016. "Automated Time Series Content Analysis with News Big Data Analytics: Analyzing Sources and Quotes in One Million News Articles for 26 Years", *Korean Journal of Journalism & Communication Studies*, 60(5): 353-407.

6. 박상조·박승관, 2016. "외국인 범죄에 대한 언론 보도가 외국인 우범자 인식의 형성에 미치는 영향", 「한국언론학보」, 60(3): 145-177.
Park, S.J. and Park, S.G., 2016. "Are Foreigners Really Crime-prone?: How Crime News Coverage Shapes Perceptions of Foreign Criminality in South Korea", *Korean Journal of Journalism & Communication*, 60(3): 145-177.

7. 박준영·채명수·정성관, 2016. "실시간 범죄 예측을 위한 랜덤포레스트 알고리즘 기반의 범죄 유형 분류모델 및 모니터링 인터페이스 디자인 요소 제안", 「정보과학회 컴퓨팅의 실제 논문지」, 22(9): 455-460.
Park, J.Y., Chae, M.S., and Jung, S.K., 2016. "Classification Model of Types of Crime based on Random-forest Algorithms and Monitoring Interface Design Factors for Real-time Crime Prediction", *KIISE Transactions on Computing Practices*, 22(9): 455-460.

8. 윤우석, 2015. "시계열분석을 통한 범죄예방환경 조성사업의 범죄억제효과 분석: 구미시 사례를 중심으로", 「한국범죄학」, 9(3): 131-164.
Yun, W., 2015. "Testing the Crime Deterrence Effects of Creating Crime Preventive Environment by Using Time-Series Analysis: Focusing on Gu-mi City Case", *Journal of Korean Criminological Association*, 9(3): 131-164.

9. 이완수·송상근, 2020. "범죄기사는 어떻게 의미구성되는가?: 포털 사이트〈네이버〉에 실린 범죄뉴스 텍스트 구조분석을 통해", 「방송과 커뮤니케이션」, 21(4): 5-46
Lee, W.S. and Song, S.K., 2020. "How Is Crime News Semantic Structured?: Through the Text Structure Analysis of Major Crime News on the Korean Portal Site 〈Naver〉", *Broadcasting & Communacation*, 21(4): 5-46

10. 하재현·기동환·이수기·안동욱, 2019. "4차 산업혁명 요소 기술을 통해 대응 가능한 서울시 도시문제 및 이슈 도출: 텍스트 마이닝 분석과 델파이 조사의 적용을 중심으로", 「서울도시연구」, 20(4): 1-21.
Ha, J., Ki, D., Lee, S., and An, D., 2019. "Identification of Urban Problems and Issues using 4th Industrial Revolution Element Technology in Seoul, Korea: Focusing on the Application of Text Mining Analysis and Delphi Survey", *Seoul Studies*, 20(4): 1-21.

11. Aggarwal, C.C. and Zhai, C., 2012. *Mining Text Data*, Springer International Publishing, 429-455.

12. Aghababaei, S. and Makrehchi, M., 2018. "Mining Twitter Data for Crime Trend Prediction", *Intelligent Data Analysis*, 22(1): 117-141.

13. Al-Zaidy, R., Fung, B.C., and Youssef, A.M., 2011. "Towards Discovering Criminal Communities from Textual Data", In Proceedings of the 2011 ACM Symposium on Applied Computing, 172-177, TaiChung Taiwan.

14. Angelov, D., 2020, "Top2vec: Distributed Representations of Topics", arXiv Preprint (unpublished material).

15. Blei, D.M., Ng, A.Y., and Jordan, M.I., 2003. "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, 3(Jan): 993-1022.

16. Borges, J., Ziehr, D., Beigl, M., Cacho, N., Martins, A., Araujo, A., Bezerra, L., and Geisler, S., 2018. "Time-Series Features for Predictive Policing", In Proceedings of the 2018 IEEE International Smart Cities Conference, Kansas City, USA.

17. Brantingham, P.J. and Brantingham, P.L., 1981. *Environmental Criminology*, CA, Sage Publications.

18. Cornish, D.B. and Clarke, R.V., 1989. "Crime Specialisation, Crime Displacement and Rational Choice Theory", In *Criminal Behavior and the Justice System: Psychological Perspectives*, 103-117.

19. Dasgupta, T., Naskar, A., Saha, R., and Dey, L., 2017. "Crime-profiler: Crime Information Extraction and Visualization from News Media", In Proceedings of the International Conference on Web Intelligence, 541-549, Leipzig, Germany.

20. Devi, J.V. and Kavitha, K.S., 2022. "Time Series Analysis and Forecasting on Crime Data", In Proceedings of 2nd International Conference on Artificial Intelligence: Advances and Applications: ICAIAA 2021, 281-297, Jaipur, India.

21. Egger, R. and Yu, J. 2022, "A Topic Modeling Comparison between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts", *Frontiers in Sociology*, 7: 886498.

22. Elyezjy, N.T. and Elhaless, A.M., 2015. "Investigating Crimes Using Text Mining and Network Analysis", *International Journal of Computer Applications*, 126(8): 19-25.

23. Garland, D., 2000. "The Culture of High Crime Societies", *British Journal of Criminology*, 40(3): 347-375.

24. Gilad, M., Gutman, A., and Chawaga, S.P., 2019. "The Snowball Effect of Crime and Violence: Measuring the Triple-C Impact", *Fordham Urban Law Journal*, 46(1): 1-72.

25. Grootendorst, M., 2022. "BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure", arXiv preprint arXiv:2203.05794.

26. Jacobs, J., 1961. *Tire Death and Life of Great American Cities*, New York: Vintage Books.

27. Jiang, X., Zheng, Z., Zheng, Y., and Mao, Z. 2023. "Spatio-temporal Distribution and Influencing Factors of Theft During the Pre-COVID-19 and COVID-19 Periods: A Case

Study of Haining City, Zhejiang, China", *ISPRS International Journal of Geo-Information*, 12(5): 189.

28. Khairova, N., Mamyrbayev, O., Rizun, N., Razno, M., and Galiya, Y., 2023. "A Parallel Corpus-Based Approach to the Crime Event Extraction for Low-Resource Languages", *IEEE Access*, 11: 54093-54111.
29. Mastrorocco, N. and Minale, L., 2018. "News Media and Crime Perceptions: Evidence from a Natural Experiment", *Journal of Public Economics*, 165: 230-255.
30. Tereszkiewicz, A., 2012. "Lead, Headline, News Abstract?–Genre Conventions of News Sections on Newspaper Websites", *Studia Linguistica Universitatis Jagellonica Cracoviensis*, 129: 211-224.
31. Towers, S., Chen, S., Malik, A., and Ebert, D., 2018. "Factors Influencing Temporal Patterns in Crime in a Large American City: A Predictive Analytics Perspective", *PLoS One*, 13(10): e0205151.
32. Tseng, Y.H., Ho, Z.P., Yang, K.S., and Chen, C.C., 2012. "Mining Term Networks from Text Collections for Crime Investigation", *Expert Systems with Applications*, 39(11): 10082-10090.
33. Umair, A., Sarfraz, M.S., Ahmad, M., Habib, U., Ullah, M.H., and Mazzara, M., 2020. "Spatiotemporal Analysis of Web News Archives for Crime Prediction", *Applied Sciences*, 10(22): 8220.
34. Vivek, M. and Prathap, B.R., 2023. "Spatio-temporal Crime Analysis and Forecasting on Twitter Data Using Machine Learning Algorithms", *SN Computer Science*, 4(4): 383.
36. Yadav, R. and Sheoran, S.K., 2018. "Crime Prediction Using Auto Regression Techniques for Time Series Data", In 2018 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE), 1-5, Jaipur, India.
36. Zanini, N. and Dhawan, V., 2015. "Text Mining: An Introduction to Theory and Some Applications", *Research Matters*, 19: 38-45.
37. Grootendorst, M., 2020. "BERTopic: Leveraging BERT and c-TF-IDF to Create Easily Interpretable Topics", Zenodo, Version v 0, 9. https://github.com/MaartenGr/BERTopic
38. Grootendorst, M., 2023. "The Algorithm. BERTopic", https://maartengr.github.io/BERTopic/algorithm/algorithm.html

Date Received	2024-05-29
Reviewed(1st)	2024-10-04
Date Revised	2025-01-14
Reviewed(2nd)	2025-03-18
Date Accepted	2025-03-18
Final Received	2025-04-15