JDFR Journal of Data Forensics Research
데이터포렌식연구, 3권 1호, 2026년 6월, 27-54면
Journal of Data Forensics Research, Volume 3, Issue 1, June 2026, Pages 27-54.
https://doi.org/10.12972/JDFR.2026.3.1.2
Print ISSN 3058-9460
Online ISSN 3058-9479

원저

토픽 모델링 평가를 통한 다크웹 데이터 분류와 위협 대응방법

김순욱[1], 오소정[2], 최민영[1], 김형빈[1], 양주은[1], 강우현[1], 김기범[3]

[1]성균관대학교 과학수사학과 디지털포렌식전공 석사과정, [2]성균관대학교 과학수사학과 디지털포렌식전공 박사과정, [3]성균관대학교 과학수사학과 교수

교신저자: 김기범, freekgb02@gmail.com

요약

다크웹은 해킹 도구, 금융사기 정보, 불법 콘텐츠 등 다양한 범죄 관련 정보가 유통되는 익명 네트워크 공간으로 방대한 데이터의 비정형적 특징으로 인해 효과적인 분류 및 분석에 어려움이 존재한다. 본 연구는 다크웹 데이터의 체계적인 분류를 통해 위협 유형을 식별하고 대응 전략 수립에 활용하기 위해 토픽 모델링 기법 성능을 비교·평가하였다. 연구를 위해 약 8개월간 수집한 다크웹 HTML 데이터에 대해 중복 제거 및 본문 추출, 언어 필터링, 정규화 및 불용어 제거 등의 전처리를 수행하여 분석용 데이터셋을 구축하였다. 이후 단어 빈도(BoW) 기반의 LDA와 NMF, 임베딩 기반의 BERTopic과 Top2Vec, 대규모 언어 모델(LLM) 기반의 TopicGPT 프레임워크를 적용하여 토픽 분류를 수행하였으며, 정량적·정성적 비교 분석을 병행하여 각 모델의 분류 특성과 적용 가능성을 비교하였다. 분석 결과, 다크웹 데이터는 압수·폐쇄 사이트, 금융사기, 암호화폐, 마켓플레이스, 해킹, 음란물 등 주요 범주로 구분되었으며 모델에 따라 토픽 분리 수준과 세부 범주 식별 능력에서 차이를 보였다. 특히 임베딩 기반 모델은 의미적 유사성을 반영하여 세부 범주를 상대적으로 더 세분화하는 경향을 보였으며 단어 빈도 기반 모델은 높은 해석 가능성을 제공하였다. 이를 통해 다크웹 내 주요 위협 유형의 분포와 특성을 파악할 수 있었으며, 토픽 모델링이 대규모 다크웹 데이터의 자동 분류와 사이버 위협 인텔리전스 구축에 활용될 수 있음을 확인하였다.

주제어

다크웹, 토픽 모델링, 단어 빈도(BoW), 임베딩, LLM(Large Language Model)

Open Access

Received: June 08, 2026
Revised: June 30, 2026
Accepted: June 30, 2026
Published: June 30, 2026

Original Article

Dark web data classification and threat response methods through topic modeling evaluation

Soonwook Kim[1], Sojung Oh[2], Minyoung Choi[1], Hyungbin Kim[1], Jueun Yang[1], Woohyun Kang[1], Gibum Kim[3]

[1]Master's Program in Digital Forensics, Department of Forensic Sciences, Sungkyunkwan University, Republic of Korea
[2]Ph.D. Program in Digital Forensics, Department of Forensic Sciences, Sungkyunkwan University, Republic of Korea
[3]Professor, Department of Forensic Sciences, Sungkyunkwan University, Republic of Korea

Corresponding Author: Gibum Kim, freekgb02@gmail.com

ABSTRACT

The dark web is an anonymous network in which criminal-related information, including hacking tools, financial fraud data, and illicit content, is actively distributed. Owing to its large-scale and unstructured nature, effective classification and analysis remain challenging. This study compared and evaluated topic-modeling techniques to identify threat categories and support threat-response strategies. A dataset was constructed from dark web HTML data collected over eight months. The data were preprocessed, including duplicate removal, content extraction, language filtering, normalization, and stop-word removal. Subsequently, topic classification was performed using Bag-of-Words (BoW)-based models (latent dirichlet allocation and non-negative matrix factorization), embedding-based models (BERTopic and Top2Vec), and the large language model-based topic generative pre-trained transformer framework. The analysis identified major categories, such as seized or closed sites, financial fraud, cryptocurrency-related activities, marketplaces, hacking services, and pornographic content. Embedding-based models distinguished more fine-grained categories through semantic understanding, whereas BoW-based models provided higher interpretability. These findings demonstrate the potential of topic modeling for large-scale dark web-data classification and cyber-threat intelligence applications.

KEYWORDS

Dark web, Topic Modeling, Bag-of-Words(BoW), Embedding, Large Language Model(LLM)

I. 서론

과학기술의 발전은 역설적이게도 범죄 진입 장벽을 쉽게 낮추는 환경을 제공하였다. 특히, AI는 쉽게 가짜영상을 제작해 거짓정보를 사회에 전파하여 혼란을 야기하고, 다크웹은 신원 추적을 방지하고, 안전하게 범죄활동을 할 수 있도록 암시장 공간을 마련해주었다. 2010년대 비트코인의 등장으로 마약, 랜섬웨어, 아동성착취물 등 불법정보 거래가 성행되는 범죄수익의 이전 및 관리에 악용될 수 있는 환경이 확대되었다. 대표적으로 자신의 해킹기술을 제품화하고, 무단 침입하여 불법적으로 지득한 정보를 포트폴리오처럼 게시하여, 판매하는 범죄가 자주 발생한다[1,2]. 다크웹에서 유통되는 데이터의 종류는 매우 다양하며, 양도 방대하다. Europol에 따르면 다크웹 기반 사이버 범죄는 매년 그 규모와 복잡성이 증가하고 있으며, 국경을 초월한 초국가적 범죄 생태계로 확산되고 있다[3]. 다크웹으로 파생되는 사이버 위협은 유형에 따라 위협 수준과 파급 범위가 상이하며, 단일 대응 체계만으로는 효과적인 탐지, 차단이 어렵다. 그간, 다크웹 위협 대응은 크게 네트워크 트래픽 분석, 사이버 위협 인텔리전스(CTI) 탐지, 콘텐츠 분류 등에 치중되어 있어 개별 주제에 대한 규칙을 세우지 못하는 등 대응전략을 수립하지 못하였다. 특히, 방대한 데이터 규모와 지속적인 콘텐츠 변화에 효과적으로 대응하기 어려웠고, 새로운 위협 유형에 대한 적용이 어렵다는 지적이 많았다[4]. 따라서 본 연구는 기존에 개별적으로 활용된 기법들을 동일하게 통제된 조건에서 병렬 비교하여 다크웹에서 유통되는 데이터의 주제 범주와 각 기법의 적용 가능성을 평가하고자 한다. 구체적으로는 단어 빈도(BoW) 기반의 LDA와 NMF, 임베딩 기반의 BERTopic과 Top2Vec, TopicGPT 프레임워크 기반의 Llama 3.2와 Gemma 3 등 3개 패러다임 6개 모델을 최종 분석 대상 11,952개 문서에 적용하여 정량적·정성적 비교 분석을 수행한다. 이를 위해 제2장에서는 다크웹에서 발생하는 범죄현황과 위협수준을 평가하고, 현재 대응현황에 대해 살펴본다. 제3장에서는 토픽 모델링에 사용할 데이터 셋 설계와 전처리 방법에 대해서 설명하고, 제4장에서는 선별 모델 6개의 결과를 분석한다. 은어, 약어, 다국어 혼용, 노이즈가 많아 일반 NLP 기법을 그대로 적용하면 성능이 약해지는 한계를 보완하고, 비정형화된 데이터에도 강인한 모델을 식별하여 향후, 다크웹 추적 기술 발전에 기초자료로 활용될 것으로 기대된다.

II. 다크웹 위협과 대응현황

2.1. 다크웹 범죄와 특징

다크웹(Dark Web)은 일반 검색 엔진을 통해 접근할 수 없는 인터넷 영역으로, 접속을 위해서는 Tor(The Onion Router)와 같은 특수 브라우저와 정확한 주소 정보가 요구된다. Tor 네트워크는 이용자의 신원과 접속 위치를 3개 이상의 릴레이 노드를 지나면서 암호화한다. 최근, 접속할 때 자체적인 신원을 비식별한 채 접속할 수 있도록 지원하는 VPN서비스 등이 등장하면서 다크웹에서의 활동을 표면으로부터 완전히 폐쇄시키기도 한다. 이는 다크웹에서 범죄자들이 신원을 드러내지 않은 채, 불법정보를 거래하거나 판매하는 등의 범죄 생태계를 마련해주었다[5].

일반적으로 다크웹 범죄는 마약, 아동성착취물, 살인청부 등 불법정보를 거래하거나 해킹, 악성코드 등을 서비스화하여 판매한다[6]. 특히, 서비스형 범죄(Crime-as-a-Service, CaaS)가 활성화되면서 공격기술에 대한 전문지식이 부족하더라도 수월하게 범죄에 진입할 수 있게 되었다. Tor Metrics에 따르면 2026년 6월 초 기준 Tor 네트워크에는 약 9,700개의 relay와 약 2,700개의 bridge가 운영되고 있으며, v3 onion service의 고유 .onion 주소 수는 약 78만 개 수준으로 추정된다. Boshmaf(2022)에 따르면 2022년 기준으로 다크웹에서는 70만 개 이상

의 고유 onion 도메인이 존재하는 것으로 보고되었다[7]. Pastor-Galindo(2024)의 연구에 따르면 80,049개의 onion 서비스를 분석한 결과 실제 고유 사이트는 약 6.1%에 불과하며 대부분 카드 정보거래, 암호화폐 서비스, 마켓플레이스 등이 주요 콘텐츠를 구성한다고 보고하였다[8].

서로 다른 공간에 있어도 다크웹을 통해 수월하게 범죄를 공모, 실행할 수 있고, 구매자 리뷰, 에스크로(Escrow) 결제 시스템, 판매자 신용도 등 기업형 플랫폼 구조를 갖추고 전 세계를 대상으로 국경 없이 확산되는 특징이 있다. 홍보, 자금관리, 제품관리 등 역할이 분담되어 있는 조직구조도 보인다. 개인이 활동하는 것 같지만, 연관된 계정을 연결하다 보면 관계성이 드러나 조직이라고 추정하기도 한다. SNS와 결합하면, 다른 국적의 조직원을 포섭할 수 있어 초국가적 범죄 대응을 위해 국제공조의 중요성도 부각되고 있는 추세이다[9]. Silk Road, AlphaBay, Hansa Market 등 대규모 다크웹 마켓이 국제공조 작전으로 폐쇄되었으나, 다크웹은 분산형 플랫폼, 익명화 기술, 프라이버시가 강화된 암호화폐의 활용 증가로 인해 추적 및 증거 수집에 한계가 있다. 국가별 법과 수사권한의 차이로 인해 국제공조 과정에서 어려움이 있으며, 법적 쟁점도 지속적으로 제기되고 있다[10].

2.2. 다크웹 사이버 위협수준

사이버범죄에서 다크웹은 강력한 익명성과 암호화 기술에 기반한 네트워크로 인하여 주요 범죄 활동 공간이 되었다. 단순히 불법거래를 넘어 범죄자들이 정보를 공유하고, 공격도구를 거래하며 취약점을 유통하는 핵심 인프라도 발전하면서 그 위협수준도 지속 증가하고 있는 추세이다. Europol의 보고에 따르면 다크웹에서 유통되는 사이버 범죄 도구와 서비스는 전통적인 사이버 범죄의 진입 장벽을 낮추는 주요 요인으로 지목되고 있다[11]. 또한 일반 검색엔진에 색인되지 않음에도 온라인 영역으로 운영되어 접속, 배포 등이 수월하다.

다크웹에서는 악성코드, 랜섬웨어, 익스플로잇 키트, 탈취 계정, 개인정보 등 불법정보가 조직적으로 거래되고, 실제 사이버공격 피해가 지속적으로 증가하고 있다. 특히, 서비스형 랜섬웨어는 다크웹 범죄 생태계의 주요 구성요소로 기능한다. 다크웹에서 해커를 고용하고, 랜섬웨어 도구를 상품으로 판매한다. WannaCry[12], NotPetya[13] 등 거대 랜섬웨어는 전 세계 조직에 수백만 달러의 재정적 손실을 입혔다. 최초 발생 당시 다크웹의 해커 포럼이나 익명 네트워크를 통해 유포되거나 테스트되었을 가능성이 보안 전문가들에 의해 지속적으로 제기되었다. 뿐만 아니라, 다크웹은 해커들의 소통공간으로서 부족한 기술을 공유하고, 서로의 기술을 사고 파는 형태를 이루고 있다. APT29(코지 베어), APT28(팬시 베어)과 같은 그룹들은 은밀한 작전을 수행하기 위해 다크웹 포럼에서 자원을 공유하는 것으로 보고되고 있다[14]. 개인정보나 탈취한 정보 공유도 심각하다. '스텔스모어 인텔리전스(StealthMole)'의 조사 결과에 따르면, 다크웹 마켓 및 해킹 포럼에서 무료·유료로 유통되고 있는 전 세계 데이터(약 900억 건)를 전수 분석한 결과, 한국인과 관련된 계정(ID/PW 조합), 이메일, 금융 정보 등이 약 4억 6,000만 건 추산된다고 발표하였다[15]. 다크웹 내 불법 마켓플레이스의 매출은 연간 약 26억 달러(약 3조 5,000억 원)에 달하는 것으로 추정하고 있다.

이제 다크웹 마켓에서는 오프라인 거래만 허용되던 마약, 성, 살인청부 등도 유통한다. 이는 단순히 정신적 피해를 넘어 신체적으로도 손상을 입히는 하이브리드형 범죄로 전락하고 있다. 대표적으로 실크 로드는 가장 악명 높은 다크웹 마켓 중 하나로, 마약부터 위조문서, 청부 살인 서비스까지 광범위한 불법 상품과 서비스가 거래되었다. 로스 울브리히트(Ross Ulbricht)가 설립한 사이트로 2011년에 출시되어 FBI에 의해 폐쇄(2013년 10월)될 때까지 운영되었고, 이때부터 다크웹이 범죄 거래를 가능케 하는 방식을 광범위하게 드러내고, 공론화되는 계기를 마

련하였다. 단순한 불법 거래 공간을 넘어 사이버 범죄자들이 정보를 공유하고 공격 도구를 거래하며 취약점 유통 및 공격을 준비하는 핵심 인프라로 발전하였으며 그 위협 수준 또한 지속적으로 높아지고 있는 추세이다.

2.3. 다크웹 위협 대응 현황

2.3.1. 사이버 위협 인텔리전스

사이버 위협 인텔리전스(Cyber Threat Intelligence, CTI)는 다크웹에서 유통되는 악성코드, 해킹 도구, 유출 정보 등 범죄 관련 정보를 수집·분석하여 잠재적 위협을 조기에 식별하는 분야이다. 최근, 자연어처리와 인공지능 기술을 활용하여 다크웹 위협 정보를 자동 수집·분석하는 연구가 활발히 이루어지고 있다. 대표적으로 Samtani et al.(2020)은 해커 포럼 데이터를 활용하여 신규 사이버 위협을 조기에 탐지하는 D-GEF(Diachronic Graph Embedding Framework)를 제안하였다[16]. Moraliyage et al. (2022)은 설명가능 인공지능(Explainable Artificial Intelligence, XAI) 기반의 멀티모달 딥러닝 모델을 활용하여 Onion Service의 유형을 자동 분류하는 방법을 제안하였다[17]. Shah & Madisetti (2025)은 대규모 언어모델에 기반하여 다크웹 게시글로부터 취약점, 해킹정보, 악성코드 등을 자동으로 수집, 분류하였다[18]. 그간, 다크웹에서 CTI는 발생하는 잠재적 위협을 조기에 식별·수집할 수 있다는 점에서 효과적이나 주로 게시글이나 웹페이지 등 소규모 단위의 위협 정보 탐지에만 집중되어 있어 범죄조직의 구조나 구성원 간 관계를 분석하는 데 한계가 있다.

2.3.2. 다크웹 트래픽 분석

다크웹 대응을 위해 Tor 등 익명 네트워크에서 발생하는 통신 데이터를 분석하여 이용 여부나 악성 행위를 탐지하려는 시도가 계속되고 있다. 대표적으로 Sarwar et al. (2021)은 합성곱 신경망(Convolutional Neural Network, CNN)과 장단기 메모리(Long Short-Term Memory, LSTM)를 결합하여 트래픽을 탐지하였고, 약 96%의 탐지 성능과 약 89%의 분류 성능을 달성하였다[19]. Mandela et al. (2025)은 CNN-LSTM 하이브리드 모델을 활용하여 다크웹 트래픽을 분류하고, 98.4%의 성능을 달성하였다[20]. 그간, 다크웹 트래픽을 기반으로 네트워크 수준의 이상행위를 탐지하여 사용 여부를 분류해내는데 집중하였으나 새로운 암호화 기술이나 난독화 기법이 도입되는 순간 탐지 성능이 급격히 저하될 가능성이 있다. 즉, 과거에 수집된 정형화된 다크웹 트래픽을 사후 분석하는 데는 강력하지만, 실시간으로 진화하는 다크웹 우회 트래픽을 현장에서 실시간 차단하기에는 한계가 있다. 무엇보다 트래픽 분석만으로 접속 주체를 특정하거나 행위를 귀속하는데 한계가 있다.

2.3.3. 콘텐츠 분류 및 유형화

다크웹에서 수집되는 게시글, 포럼, 마켓플레이스 상품 정보 등을 분석하여 범죄유형을 식별·분류하는 시도도 다수 발견되었다. Basheer & Alkhatib (2024)은 Correlated Topic Model(CTM)과 온톨로지 학습(Ontology Learning)을 결합하여 다크웹 게시글에서 주요 토픽을 추출한 뒤, 범죄 커뮤니티의 주요 관심사와 개념 간 관계를 구조적으로 표현해내었다[21]. Prado-Sánchez et al. (2025)은 CoDA 데이터셋 10,000건을 활용하여 GPT-4o, Claude 3.5 Haiku, Gemini 2.0 Flash, DeepSeek Chat 등 상용 대규모 언어모델의 Zero-shot 분류 성능을 비교하였다[22]. 전통적인 확률적 토픽 모델링(CTM)과 지식 그래프(Ontology)를 결합한 방

식은 범죄 커뮤니티의 거시적인 관심사를 시각화하는 데 훌륭하지만, 변화하는 범죄 환경에 적용하기 어렵다. 또한 GPT-4o, Claude 3.5 Haiku, Gemini 2.0 Flash 등 고성능 LLM을 데이터 가공 없이(Zero-shot) 프롬프트만으로 다크웹 분류에 활용하는 방식은 인간 수준의 문맥 이해도를 보여주지만, 환각 현상(Hallucination)과 일관성 부족이 발생할 수 있다는 단점이 있다. 결과적으로는 수사·추적에 직접 활용하기 위해서는 추가적인 결과 검증과 관계 분석의 결합이 필요하다.

2.3.4. 구조 및 관계 분석

사용자 간 상호작용, 계정 정보, 암호화폐 지갑 주소 등의 관계를 분석하여 범죄 조직의 구조와 커뮤니티 특성을 파악하는 구조 및 관계 분석 연구도 진행되고 있다. Shin et al. (2025)은 BERTopic과 '작성자 식별' 기법을 결합하여 다크웹과 표면웹 간 유사 사용자를 식별하는 방법을 제안하였다[23]. de Marcos et al. (2025)은 이메일 주소, Telegram 계정, 암호화폐 지갑 주소 등을 기반으로 이분 그래프(Bipartite Graph)를 구축하고 연결 중심성, 매개 중심성, K-core 분석을 적용하여 다크웹 내 식별정보 간 연결 구조를 분석하였다[24]. 그 결과, Telegram 계정이 해킹 관련 커뮤니티의 핵심 연결 수단으로 Monero(XMR) 지갑 주소가 암호화폐 기반 금융 거래 네트워크의 중심으로 기능하고, 해킹 및 금융·암호화폐, 마약 관련 활동이 서로 다른 네트워크 구조를 형성하고 있음을 밝혔다. SNA(사회관계망 분석)으로 범죄자의 식별 정보(ID)와 자금 세탁 경로를 엮어 조직적 연계 가능성을 탐색하려는 시도로 효과적이다. 숙련된 범죄자들은 하나의 글에 이메일, 텔레그램, 암호화폐 지갑 주소 등 동시에 많은 정보들을 노출하지 않는다. 따라서 분산된 식별정보를 노드로 구조화하여 이들 간 연결선(Edge)을 확보하는 작업이 선행되어야 한다는 단점이 있다.

III. 토픽 모델링 기법 선정 및 데이터 셋 설계

3.1. 실험방법

이 연구는 대량의 다크웹 데이터로부터 토픽 모델링을 통해 분류하여 미래 위협 대응 전략을 수립하는 기초자료 마련에 목적이 있다. 다크웹은 비정형 데이터로서 라벨링을 통한 사전 분류 작업이 어려워 지도 학습을 직접 활용하기 어렵다. 다크웹 사이트를 분류하는데 어떤 토픽 모델링 기술이 적합한지 규명하고자 한다. 먼저, 다크웹 사이트에서 HTML 본문을 크롤링하고, 중복제거, 언어필터링 등 전처리를 거쳐 분석에 적합하도록 정제된 텍스트 데이터셋을 구축한다.(① 정제된 데이터 구축) 이후, 단어 빈도 기반, 임베딩 기반, LLM(Large Language Model) 기반 등 3개 기법으로 분류하고, 각 접근법에 적합한 모델을 2개씩 선별한다. 단어 빈도 기반 모델은 LDA와 NMF, 임베딩 기반 모델은 BERTopic과 Top2Vec, LLM 기반 모델은 TopicGPT 프레임워크를 로컬 언어모델(Llama, Gemma)로 구동하여 적용한다.(② 토픽 모델링 적용) 각 기법은 데이터로부터 잠재적인 주제 집합을 도출하고, 개별 문서에 배정하는 방식으로 구동된다. BoW·임베딩 모델은 가중치가 높은 상위 단어들로, LLM 모델은 자연어 라벨로 토픽을 표현하였고, 개별 문서는 주제 분포, 가중치가 가장 높은 단일 주제에 할당하였다. 1차 분류 결과 이질적인 하위 유형을 포함하는 것으로 판단된 범주에 대해서는 동일한 기법을 재적용하여 보다 세분화된 하위 주제를 도출하는 2단계 계층적 분류를 수행하였다. 이때, 2단계 분류 대상은 특정 범주에 4,000건 이상의 문서가 집중된 경우에 해당한다. 결과적으로 범주(대주제)에 라벨(세부주제)이 적절하게 태깅되도록 하였다. 마지막으로 모델별 토픽 구성, 세부 범주 분리 수준, 처

리 시간 및 라벨 해석 가능성을 분석한다.(③ 평가)

3.2. 데이터 수집 및 전처리

다크웹은 도메인이 빈번하게 변경되고 사이트의 생성과 소멸 주기가 짧아, 일반적인 표면 웹 크롤링 방식으로는 유효한 데이터를 안정적으로 수집하기 어렵다. 데이터 수집은 진필근 등(2024)이 제시한 비표준화된 다크웹 생태계에 안정적인 수집이 가능한 재귀적 크롤러를 활용하였고[25], 2025년 9월 22일~2026년 5월 22일(약 8개월)까지 총 165,461개의 HTML 데이터를 대상으로 하였다. 재귀적 크롤러는 원천시드를 기준으로 삼아 이 시드로부터 접근 가능한 모든 도메인을 재귀적으로 확장하여 해당 사이트도 원천시드로 수렴시키는 특징이 있다. 또한 데이터가 변경되면 감지하여 해당 데이터도 수집하므로 동일한 시드 주소가 다수 존재하게 된다. 실험 데이터는 불법 콘텐츠 및 민감정보가 포함될 가능성을 고려하여 영상과 이미지는 수집 대상에서 제외하였으며, 분석은 로컬 환경에서 수행하였다.

동일 URL에 대해 가장 최근에 수집된 데이터를 기준으로 과거에 수집된 중복 데이터를 제거하여 12,104개의 시드를 확보하였다. 토픽 모델링에 적용하기 위해서 하나의 시드에서 추출한 텍스트 데이터 집합을 문서로서 간주하고, 각 문서를 토픽 모델링의 입력 단위로 취급한다. 또한 데이터는 HTML 태그, 스크립트, 잡음 텍스트 등 노이즈가 포함되어 있어 토픽모델링에 적합한 형태로 정제하는 작업이 필요하다. 먼저, HTML 데이터에서 태그와 스크립트, 스타일 요소를 제거하고 사이트의 본문 텍스트만을 분리하였다. 다양한 언어가 혼재되어 있어 모델 입력에 정합성을 위해 언어정보, 자연어처리를 통해 “영어” 텍스트만 추출하였다. 이후 모든 텍스트를 소문자로 변환하고, URL·숫자·특수문자를 비롯하여 영문 알파벳이 아닌 문자, 즉 이모지나 비라틴 문자 등을 모두 제거하여 영문 알파벳과 공백만을 남기는 정규화를 수행하였다. 또한 본문 길이를 기준으로 10자 미만의 문서는 내용이 거의 없는 것으로 보아 분석에서 제외하였으며, 최대 5,000자까지의 본문만 사용하여 지나치게 긴 문서로 인해 분석에 과도한 영향이 미치지 않도록 하였다. 정규화된 텍스트에 대해서는 “the”, “and”, “a”, “an”과 같이 빈번하게 출현하나 주제 식별에는 기여하지 않는 불용어(stop-word)를 제거하여 잡음을 줄였으며, 텍스트를 단어 단위의 토큰으로 분할하는 토큰화를 수행하였다. 마지막으로 표제어 추출을 통해 문서 내 단어들을 사전에 등재된 기본형으로 변환해 동일한 의미를 가진 단어들이 서로 다른 형태로 분산되어 발생하는 희소성 문제를 완화했다. 결과적으로 총 11,952개 문서를 분석 대상으로 하였다<Table 1>. 참고로 수집 원본은 jsonl 형식(약 404MB, 41,389행)으로 저장되었으며, 이는 전처리 이전의 중간 산출물 규모로서 최종 분석 대상 문서 수와는 구별된다.

<Table 1> 데이터 전처리 과정

중복 제거	→	HTML 정제	→	정규화	→	길이 필터링	→	토큰화
최신 수집 기준, URL 중복 제거		html에서 사이트 본문만 추출		영문 추출, 소문자 변환		10 ~ 5,000자		불용어 제거, 표제어 변환

3.3. 토픽 모델링 모델 설정 및 동작원리

3.3.1. 단어 빈도(BoW) 기반

단어빈도 기반 모델은 문서 내 단어의 출현 빈도를 수치화하여 문서-단어 행렬을 구성하고, 문서 집합에 숨어있는 잠재 토픽을 추출하는 방식이다. 확률 기반 모델과 행렬분해 기반 모델이 있

고, 각각 LDA(Latent Dirichlet Allocation)와 NMF(Non-negative Matrix Factorization)를 선정하였다. 다크웹에서 상품 설명이나 포럼 글은 정제되어 있지 않고, 매우 짧고, 은어와 신조어가 가득하며, 맥락이 파편화되어 있다. 그럼에도 두 모델은 특정 단어가 어떤 가중치와 확률로 이 토픽에 기여했는지 토픽-단어 확률과 단어 가중치를 통해 해석 근거를 제시할 수 있다.

LDA와 NMF는 문서를 단어 출현 통계에 기반하여 문서-단어 행렬로 표현하는 공통점이 있지만, 가정에 따라 입력을 달리한다는 차이점이 있다. 먼저, LDA는 문서에 대해서는 잠재 주제들의 확률분포로, 주제에 대해서는 단어들의 확률 분포로 모형화하는 생성모형이다. 단어의 출현 빈도(term frequency)를 그대로 사용하는 Bag-of-Words 표현을 입력으로 한다[26]. tomotopy 라이브러리로 구현하였으며, 문서-주제 분포의 디리클레 사전 하이퍼파라미터 α를 0.1, 주제-단어 분포의 사전 하이퍼파라미터 β를 0.01, 난수를 42로 설정하고 1,000회 반복 학습하였다. 반면, NMF는 문서-단어 행렬을 문서-주제 행렬과 주제-단어 행렬의 곱으로 근사 분해하는 선형대수적 기법이다. 거의 모든 문서에 등장하는 도메인 공통어의 영향을 완화하기 위해 문서 빈도로 가중치를 보정한 TF-IDF 표현을 입력으로 사용하였다. 모델은 scikit-learn 라이브러리로 구현하였다. 입력 TF-IDF 행렬은 최대 어휘 20,000개, 최소 문서 빈도 5, 최대 문서 비율 0.95, 유니그램 및 바이그램(n-gram 범위 (1, 2))의 조건으로 생성하였으며, 난수는 42로 설정하였고, 손실 함수로는 Frobenius norm을 사용하여 최대 400회 반복으로 학습하였다.

3.3.2. 임베딩 기반

임베딩 기반 모델은 문서를 의미 정보가 압축된 고차원 밀집 벡터(dense embedding)로 변환하고, 이 벡터 공간에서 의미적으로 유사한 문서들을 군집화하여 잠재 토픽을 추출하는 방식이다. 다크웹 데이터의 가장 고질적인 문제인 문맥 단절과 노이즈 문제를 완화하는데 효과적으로 알려진 BERTopic과 Top2Vec를 선정하였다. 두 모델은 문맥상 유사한 표현까지 동일 주제로 묶을 수 있고, 토픽 수를 사전 지정하는 단어 빈도 기반 기법과 달리 토픽 수가 데이터의 밀도 구조로부터 결정된다는 공통점이 있다. 다만, 문서 임베딩을 외부 사전학습 모델에서 얻는지(BERTopic), 분석 대상 코퍼스에서 직접 학습하는지(Top2Vec)에 따라 임베딩 생성 방식을 달리한다는 차이점이 있다.

BERTopic은 문서 임베딩, 차원 축소, 군집화, 주제 표현의 단계로 동작한다. 각 문서를 SBERT 기반 모델로 의미 벡터로 변환한 뒤 UMAP으로 차원을 축소하고 HDBSCAN으로 밀도 기반 군집화를 수행하여 토픽을 형성하며, 이때 토픽 수와 이상치 비율은 최소 군집 크기에 의해 결정된다. 각 군집을 하나의 문서로 간주한 군집 단위 변형 TF-IDF(c-TF-IDF)로 토픽별 대표 키워드를 추출한다. 임베딩 모델은 all-MiniLM-L6-v2(384차원), UMAP은 이웃 수 15, 축소 차원 5, 최소 거리 0.0, HDBSCAN은 최소 군집 크기 50, 최소 표본 수 기본값(None)으로 설정하였다. 반면, Top2Vec은 문서와 단어를 동일한 의미 공간에 함께 표현한 뒤 밀집 영역을 주제로 식별한다. 사전학습 모델을 쓰는 BERTopic과 달리, Doc2Vec으로 분석 대상 코퍼스에서 문서·단어 임베딩을 함께 학습한 후 동일하게 UMAP과 HDBSCAN으로 군집을 탐색하고, 각 군집의 중심을 토픽 벡터로 정의하여 가장 가까운 단어 벡터들을 대표 키워드로 추출한다. Top2Vec 또한 토픽 수를 사전 지정하지 않고 임베딩은 Doc2Vec으로 직접 학습한 300차원 벡터, UMAP은 이웃 수 15, 축소 차원 5, 최소 거리 0.0, 거리 척도 cosine, HDBSCAN은 최소 군집 크기 15, 최소 표본 수 기본값(None)으로 설정하였다.

3.3.3. LLM 기반 토픽 모델링

LLM 기반 모델은 단어의 출현 통계가 아니라 문서의 텍스트를 대규모 언어모델(LLM)에 직접 입력하여 모델의 언어 이해 능력으로부터 잠재 토픽을 추출하는 방식이다. 이 모델링을 구현하기 위해서 먼저 스캐닝을 통해 추론된 주제를 추출해내는 TopicGPT 프레임워크를 적용하였다[27]. TopicGPT는 토픽 생성과 토픽 할당의 두 단계로 동작한다. 토픽 생성 단계에서는 전체 문서 집합에서 표본을 추출하여 LLM에 제시하고, 이전에 생성된 토픽 목록을 함께 입력으로 주어 새로운 토픽을 반복적으로 생성하게 한 뒤, 중복 토픽을 병합하고 출현 빈도가 낮은 토픽을 제거하는 정제 과정을 거친다. 토픽 할당 단계에서는 정제된 토픽 목록을 바탕으로 개별 문서에 가장 적합한 토픽을 할당하며, 이때 판단의 근거가 되는 문서 내 인용문을 함께 제시한다. 단어의 출현 통계를 입력으로 하는 앞의 두 부류와 달리, 본 기법은 문서의 텍스트를 직접 입력하며 단어 묶음이 아니라 사람이 이해하기 쉬운 자연어 형태의 토픽 레이블을 자동으로 산출한다는 점에서 차별화된다. 저자가 공개한 TopicGPT 구현을 사용하였으며, LLM은 Ollama 기반의 로컬 모델로 구동하였다. 토픽 생성 단계는 TopicGPT의 절차에 따라 수행하였고, 토픽 할당(분류) 단계에는 Llama 3.2(3B)와 Gemma 3(4B)를 사용하였다.

IV. 토픽 모델링 적용 결과 및 평가

4.1. 단어 빈도(BoW) 기반 토픽 모델링 결과

4.1.1. LDA(Latent Dirichlet Allocation)

전체 11,952개 문서를 대상으로 최적의 K를 선정하기 위해서 1부터 순차적으로 실험한 결과, K가 20을 초과하면 토픽 간 키워드 중복과 소규모 토픽 분할이 증가하였다. 따라서 1~20 범위에서 평가지표들의 총 점수(mean_cv)가 0.5 이상으로 측정되었고, 그 구간은 $K \in \{7, 8, 10, 12, 13, 14, 15\}$이 해당되었다. 평가지표로는 LDA에서 널리 사용되는 C_V coherence, C_NPMI coherence를 고려하였다[28]. 또한 사이트 간 중복 텍스트에 의한 평균값 왜곡을 보완하기 위해서 mean과 median을 함께 활용하였다<Table 2>.

<Table 2> LDA K별 Coherence 결과

K	mean_cv	median_cv	mean_npmi	cv_min	cv_max
7	0.6018	0.5756	0.0934	0.3657	0.8884
8	0.6177	0.6065	0.1134	0.3459	0.8884
10	0.5633	0.5412	0.0740	0.3011	0.9493
12	0.6020	0.6389	0.1101	0.1826	0.9493
13	0.6804	0.6829	0.1588	0.3189	0.9493
14	0.6589	0.6414	0.1482	0.4182	0.9687
15	0.6530	0.6597	0.1363	0.2128	0.9493

K 값은 13으로 하였다. mean_cv가 0.6804로 가장 높고, mean과 median이 유사하여 coherence 분포의 중심 경향이 크게 다르지 않음을 확인했다. 또한 토픽 상위 단어 간 동시출현의 정규화 점별 상호정보량으로서 값이 높을수록 토픽 일관성이 높음을 나타내는 mean_npmi와, 값이 낮을수록 최저 품질 토픽의 존재를 시사하는 cv_min을 종합적으로 평가하였다. K=13 적용 결과, 전체 11,952개 문서는 Blocked 5,568건(46.59%), Financial Fraud

2,306건(19.30%), Pornography 1,100건(9.20%), Hacking 1,400건(11.71%), Marketplace 425건(3.56%), Cryptocurrency 333건(2.79%), Forum 138건(1.15%), Others 682건(5.71%)으로 분류되었으며, Blocked 단일 토픽이 전체의 절반에 가까운 비율을 차지하는 것이 특징이다<Table 3>.

<Table 3> LDA 전체 토픽별 분류 결과

토픽	범주	라벨	상위 단어	문서수	비율
0	Blocked	Seized/Closed Sites	support, information, police, bavarian, anonymous, website, btc, buy, guide, socks	5,568	46.59%
1	Hacking	Hacking-for-Hire/ Spoofing Services	as, hacking, about, services, service, they, an, contact, so, time	1,400	11.71%
2	Financial Fraud	Credit-Card Dump/ Hacked PayPal Shops	usd, united, states, eur, buy, personal, confirmed, zip, yes, premier	349	2.92%
3	Marketplace	Drug Marketplaces/ Online Pharmacies	add, cart, price, buy, transfer, product, out, options, sale, escrow	425	3.56%
4	Pornography	CP/Pedo	porn, free, video, cp, teen, child, sex, girl, videos, young	221	1.85%
5	Pornography	Teen/Boys	visit, porn, teen, boys, videos, cute, young, orgasm, best, collection	879	7.35%
6	Others	Onion Link Lists/ Pirated Media	onion, tor, links, scam, search, market, dark, web, hidden, iphone	474	3.97%
7	Cryptocurrency	Bitcoin Scams	btc, bitcoin, wallet, transfer, price, address, receive, order, sent, wallets	333	2.79%
8	Forum	Forums, News Mirrors/ Translation Tools	ago, minutes, hacking, cracking, hours, carding, tutorials, methods, discussion, database	138	1.15%
9	Others	French Activist/ Independent Media	order, de, cards, service, card, money, out, information, an, everything	208	1.74%
10	Financial Fraud	Money-Transfer/ Account Shops (PayPal/Escrow)	paypal, transfer, buy, money, cards, account, accounts, western, union, transfers	1,030	8.62%
11	Financial Fraud	Cloned-Card/ Western Union Shops	cards, card, buy, online, pin, atm, world, best, code, stores	603	5.05%
12	Financial Fraud	Counterfeit Cash/ Gift Cards	buy, card, gift, money, cash, make, amazon, sale, usd, eur	324	2.71%
Total				11,952	-

2차 분류 대상 조건에 Blocked가 해당되어, 총 5,568개 문서에 대해 Seized, Closed, Others로 2차 분류하였다. 동일한 방법으로 선정된 K 값은 8이다<Table 4>. 결과적으로 Seized에 해당하는 Sub 1과 Sub 2가 전체의 91.90%를 차지하였으며, 운영 중단 안내문으로 해석되는 Closed 문서는 존재하지 않았다. Others에 해당하는 6개 Sub 토픽은 PayPal/Escrow, Bitcoin Service, Cloned-Card 등 금융사기 관련으로 세부 분류하였다. 종합적으로 살펴보았을 때, LDA는 총 8개(Hacking, Financial Fraud, Marketplace, Pornography, Others, Cryptocurrency, Forum, Seized) 대주제와 14개의 세부주제로 분류되었다.

<Table 4> LDA Blocked 토픽별 2차 분류 결과

Sub 토픽	재분류 범주	라벨	상위 단어	문서 수	비율
0	Others	Money-Transfer/Account Shops (PayPal/Escrow)	paypal, socks, proxy, account, btc, verified, ww, garanteed, accounts, exclusive	59	1.06%
1	Seized	Seized	support, information, bavarian, police, anonymous, website, btc, buy, account, guide	3,341	60.00%
2	Seized	Seized	police, bavarian, information, redirecting, back, polivalente, verification, rg, unidad, complete	1,776	31.90%
3	Others	Cloned-Card/ Western Union Shops	guide, withdrawal, cards, cost, use, balance, atm, bal, pin, card	69	1.24%
4	Others	Bitcoin Service	min, bitcoin, buy, privacy, escort, guarantee, ready, service, email, order	187	3.36%
5	Others	Bitcoin Service	btc, available, buy, wallets, bitcoin, wallet, verified, blockchain, sold, out	27	0.48%
6	Others	Credit-Card Dump/ Hacked PayPal Shops	btc, account, email, paypal, socks, tutorial, cc, info, payment, full	62	1.11%
7	Others	Cloned-Card/Western Union Shops	guide, support, socks, technical, deals, value, cashout, balance, pdf, standard	47	0.84%
Total				5,568	-

4.1.2. NMF(Non-negative Matrix Factorization)

NMF는 LDA에 비해 TF-IDF 행렬을 입력으로 사용하여 범용 단어의 가중치를 사전에 억제함으로써, 토픽 간 변별력이 높은 단어에 집중적으로 가중치가 부여되는 특성을 가진다. NMF 모델에서는 K 값 선정 과정에서 cv_max 값이 모두 1.0으로 동일하여 이를 제외하였다. 또한, 기본으로 설정한 범주가 8개이므로, K 값 선정 과정에서 8 미만은 고려하지 않았다. 최적의 K 탐색을 동일한 방법으로 수행한 결과, $K \in \{8, 10, 12, 14, 16, 18\}$ 구간을 선별하였다. K가 8 이상인 후보 중 mean_cv와 median_cv가 가장 높아 K=8을 선정하였다<Table 5>.

<Table 5> NMF K별 Coherence 결과

K	mean_cv	median_cv	mean_npmi	cv_min
8	0.7304	0.7319	0.1099	0.3047
10	0.7220	0.7055	0.1394	0.3047
12	0.6578	0.6225	0.0960	0.3047
14	0.6900	0.6790	0.1337	0.3047
16	0.6960	0.6937	0.1359	0.3047
18	0.6332	0.6390	0.0982	0.2339

분류 결과, 전체 11,952개 문서는 Blocked 5,124건(42.87%), Cryptocurrency 3,943건(32.99%), Financial Fraud 1,607건(13.45%), Pornography 1,278건(10.69%)의 4개 범주로 분류되었으며, Marketplace, Hacking, Forum은 도출되지 않았다<Table 6>. 이중, Blocked가 기준치인 4,000개 이상에 해당되어 2차 분류를 하였다.

<Table 6> NMF 전체 토픽별 분류 결과

토픽	범주	라벨	상위 단어	문서 수	비율
0	Blocked	Seized/Closed Sites	bavarian, information bavarian, op alice, authorities information, onion op, alice site, bavarian police, bavarian authorities, seized bavarian, site seized	2,139	17.90%
1	Pornography	Teen/Boys	visit, porn, teen, orgasm, cute boys, cute, boys, young, orgasm fairly, teen amateurs	744	6.22%
2	Blocked	Seized/Closed Sites	support, police anonymous, support website, anonymous information, website support, information support, support support, anonymous, website, information	2,985	24.97%
3	Financial Fraud	Cloned-Card/ Western Union Shops	cards, card, union, western union, western, visa, pin, stores, paypal, prepaid	847	7.09%
4	Financial Fraud	Counterfeit Cash/ Gift Cards	gift, gift card, card, buy, card buy, amazon, buy sale, sale, buy amazon, amazon gift	349	2.92%
5	Pornography	CP/Pedo	porn, visit, cp, child, free, video, videos, sex, teen, girl	534	4.47%
6	Cryptocurrency	Bitcoin Service	transfer, bitcoin, buy, money, paypal, btc, price, transfer price, order, cash	3,943	32.99%
7	Financial Fraud	Credit-Card Dump/ Hacked PayPal Shops	united, buy, united states, states, usd, personal buy, bazaar plastic, bazaar, counterfeits, plastic	411	3.44%
Total				11,952	-

Blocked에 해당하는 5,124개 문서에 대해 최적의 K를 탐색한 결과, 2가 선정되었다. 라벨링을 Seized, Closed, Others로 분류하였다<Table 7>. 결과적으로 Blocked 범주 내 Seized (2,139건, 41.74%)와 운영 중단 안내문 패턴이 중심인 Closed(2,985건, 58.26%)로 분류되었으며, Others에 해당하는 문서는 없었다. 종합하면 총 5개의 대주제(Pornography, Financial Fraud, Cryptocurrency, Seized, Closed)와 세부주제는 8개가 분류되었다.

<Table 7> NMF Blocked 토픽별 2차 분류 결과

Sub 토픽	재분류 범주	라벨	상위 단어	문서 수	비율
0	Seized	Seized	bavarian, information bavarian, op alice, authorities information, onion op, alice site, bavarian police, bavarian authorities, seized bavarian, site seized	2,139	41.74%
1	Closed	Closed	support, police anonymous, support website, anonymous information, website support, information support, support support, anonymous, website, information	2,985	58.26%
Total				5,124	–

4.1.3. 평가

LDA는 '인간 분석가의 사후 개입'이 불가피한 구조적 제약이 있었다. 수학적 연산 결과로 도출된 단순 단어 뭉치를 유의미하게 해석하기 위해 인간의 도메인 지식을 주입해야 했다. 단일 사이트가 무기와 마약을 동시에 취급하는 사례가 빈번하나, 각 사이트를 하나의 토픽에만 배정하는 방식은 실제 사이트의 복합적 성격을 단순화할 우려가 있다. 결과적으로 LDA를 통해 본

다크웹의 실태는 압수된 상태의 다크웹(46.59%)도 많지만, 나머지 절반은 금융사기(19.30%)와 해킹 서비스(11.71%)가 실질적인 불법 거래를 주도하고 있는 생태계라고 할 수 있다. 반면, NMF는 전체 재구성 오차를 최소화하는 방식으로 작동하기 때문에, 오차 기여가 작은 소수 어휘군은 독립 토픽으로 수렴되기 어려운 구조적 특성을 가진다. 전체 사이트에서 소수 비율을 차지하는 어휘군이 독립 토픽으로 분류되지 못하고 인접한 대형 토픽에 흡수되는 경향이 나타난다. 상위 단어와 실제 사이트 내용이 불일치하는 사례도 일부 발견되어 수동 분류하는 공수도 요구되었다. 그러나 NMF 모델은 다크웹 내 'Blocked'을 Seized 및 Closed로 해석되는 토픽을 구분하여 형성하였다. 결과적으로 NMF를 통한 다크웹은 사법 통제나 자발적 중단으로 접근 차단영역(42.87%) 외에, 가상자산 서비스(32.99%)와 금융사기(13.45%)가 지배하고 있으며, 마약이나 해킹 등 소수 위협 주제가 금융 키워드에 흡수·매몰되었다고 볼 수 있다. 즉, 다크웹이 사실상 거대한 자금 유통 인프라로 작동하고 있음을 시사한다. 사법기관의 대응에 대해서도 LDA는 차단된 사이트 대부분이 수사기관이 '압수(Seized, 42.82%)'한 성과로 해석한 반면, NMF는 압수(Seized, 17.90%)보다 운영자의 자발적 잠적이나 호스팅 만료로 '단순 폐쇄(Closed, 24.97%)'의 비중이 더 높게 나타났다<Table 8, 9>.

<Table 8> LDA 범주별 사이트 비율

범주		건수(재분류 수)	비율
Marketplace		425	3.56%
Hacking		1,400	11.71%
Financial Fraud		2,306	19.30%
Pornography		1,100	9.20%
Forum		138	1.15%
Cryptocurrency		333	2.79%
Others		682(451)	5.71%
Blocked	Seized	5,117	42.82%
	Others	451	3.77%
Total		11,952	-

<Table 9> NMF 범주별 사이트 비율

범주		건수(재분류 수)	비율
Financial Fraud		1,607	13.45%
Pornography		1,278	10.69%
Cryptocurrency		3,943	32.99%
Blocked	Seized	2,139	17.90%
	Closed	2,985	24.97%
Total		11,952	-

4.2. 임베딩 기반 토픽 모델링 결과

4.2.1. BERTopic[29]

총 11,952개 문서를 대상으로 모델에 입력한 결과, K = 20이 최적임을 확인하였고, 5씩 증가하여 중복된 주제가 등장하는 시점인 K = 60일 때, 종료하였다. 이후, 최적의 토픽 개수 선정을 위해 다양성(Diversity), 아웃라이어 비율(Outlier), 최대 토픽 수(Largest), 평균 코사인 유사

도(Cos_Avg) 등 지표를 종합하여 최종 점수(Score)를 산출하였다<Table 10>.

<Table 10> BERTopic K별 지표 산정 결과

K	Diversity	Outlier	Largest	Cos_Avg	Score
20	0.8789	0.2342	0.5008	0.3302	0.7110
25	0.8542	0.2757	0.4972	0.3577	0.7028
30	0.8414	0.2693	0.4946	0.3911	0.6836
35	0.8088	0.2882	0.4500	0.4220	0.6771
40	0.7872	0.2087	0.4826	0.4195	0.6761
45	0.7545	0.2447	0.4514	0.4413	0.6635
50	0.6490	0.2549	0.2711	0.3895	0.6864
55	0.6796	0.2380	0.2658	0.3846	0.7028
60	0.5559	0.2820	0.2347	0.3746	0.6660

최종 토픽의 개수(K)는 20으로 지정하였다. 토픽의 다양성 점수(Diversity)가 가장 높으며 생성된 토픽들이 서로 중복되지 않고 독립적인 의미가 부여된 것을 의미한다. 또한 같은 토픽 집합 중에서 “서로 얼마나 비슷한 내용인가”를 나타내는 점수인 코사인 유사도 평균(Cos_Avg)이 가장 낮아 주제별 식별이 가능하다. 그 결과, Outlier를 제외하고 Blocked, Marketplace, Financial Fraud, Pornography, Cryptocurrency, Hacking 및 Others의 7개 주요 범주가 도출되었다. 전체 11,952개 문서 기준으로 Blocked 5,249건(43.92%), Marketplaces 2,499건(20.91%), Financial Fraud 1,717건(14.37%), Pornography 1,112건(9.30%), Outlier 715건(5.98%), Cryptocurrency 282건(2.36%), Hacking 197건(1.65%), Others 181건(1.51%) 순으로 나타났다<Table 11>.

<Table 11> BERTopic 전체 토픽별 분류 결과

토픽	범주	라벨	상위 단어	문서 수	비율
-1	Outlier	Generic Money/Shopping Pages	buy, money, iphone, usd, make, cash, rich, order	715	5.98%
0	Blocked	Seized Sites	bavarian, support, information, this	5,191	43.43%
1	Marketplace	Drug Marketplaces/ Online Pharmacies	onion, add, hacking, cart	1,866	15.61%
2	Financial Fraud	Cloned-Card/Western Union Shops	cards, card, them, visa	960	8.03%
3	Pornography	Teen/Boys	visit, porn, orgasm, teen	724	6.06%
4	Financial Fraud	Money-Transfer/Account Shops (PayPal/Escrow)	paypal, transfer, accounts, account	590	4.94%
5	Pornography	CP/Pedo	porn, free, video, cp	388	3.25%
6	Marketplaces	Gift Card Resale Shops	giftcards, escort, min, amazon	275	2.30%
7	Marketplaces	Counterfeit Cash/ Gift Cards	gift, card, buy, amazon	163	1.36%
8	Cryptocurrency	Bitcoin Scams	btc, wallet, wallets, address	150	1.26%
9	Hacking	Hacking-for-Hire/ Spoofing Services	account, paypal, socks, proxy	138	1.15%
10	Marketplaces	Credit-Card Dump/ Hacked PayPal Shops	bazaar, plastic, counterfeits, paypals	138	1.15%

11	Cryptocurrency	Bitcoin Mixers/Tumblers	bitcoin, btc, bitcoins, double	132	1.10%
12	Financial Fraud	Card Scam Sites	easy, fast, safely, code	96	0.80%
13	Others	Anonymous Tor Hosting	anon, hosting, service, hidden	91	0.76%
14	Others	Onion Link Lists/ Pirated Media	coredir, libraries, trusted, explore	90	0.75%
15	Financial Fraud	Bitcoin Multiplier Scams	bitcoin, wallet, blockchain, exploit	71	0.59%
16	Hacking	Stolen Credit-Card Data Shops	order, data, information, stolen	59	0.49%
17	Blocked	Closed Sites (Error/Block Pages)	refresh, disabled, wait, please	58	0.49%
18	Marketplaces	Firearms & Weapons Marketplaces	caliber, mm, price, guns	57	0.48%
Total				11,952	-

2차 분류 대상에 해당되는 Blocked 총 5,191개 문서에 대해 세부 토픽 분류를 수행하였다. 이전과 동일한 방법으로 평가한 결과, 토픽 개수는 K=3으로 확인되었다. 분석결과, 사법당국의 압수 안내문 형태인 'Seized' 패턴이 Blocked 범주의 대부분을 차지하고 있음을 확인하였다 <Table 12>. 종합적으로 6개 (Marketplaces, Financial Fraud, Pornography, Cryptocurrency, Hacking, Blocked)의 대주제와 세부주제 20개로 구분되었다.

<Table 12> BERTopic Blocked 토픽별 분류 결과

Sub 토픽	재분류 범주	라벨	상위 단어	문서 수	비율
Sub 0	Seized	Bavarian Police Seizure	bavarian, by, authorities, alice support, website, anonymous,information, datenschutz, impressum, und, officebavarian, this, site, has	5,116	97.47%
Sub 1	Closed	Database Error/ Wait Page	please, wait, refresh, page	75	1.43%
Sub 2	Others	PAYPAL Store Remnants	buy, personal, united, premier	58	1.10%
Total				5,249	-

4.2.2. Top2Vec[30]

총 11,952개 문서를 대상으로 모델에 입력한 결과, K = 20이 최적임을 확인하였고, 5씩 증가하여 중복된 주제가 등장하는 시점인 K = 70일 때 종료하였다. 이후, 최적의 토픽 개수 선정을 위해 다양성(Diversity), 최대 토픽 수(Largest), 평균 코사인 유사도(Cos_Avg) 등 지표를 종합하여 최종 점수(Score)를 산출하였다<Table 13>. Top2Vec은 BERTopic과 다르게 모든 문서를 대상으로 탐색하면서 가장 가까운 주제로 강제 할당하지만, 자체적인 임베딩 공간 학습 과정에서 대량의 중복 문서나 안내문 등이 지배적인 군집을 형성하여 평가지표가 비정상적으로 왜곡되는 현상이 발생한다. 즉, BERTopic과 동등한 조건에서 모델 고유의 주제 분류 성능을 객관적으로 평가하고 비교하기 위해, 지표 산출 과정에서 다크웹 반복 데이터 및 노이즈 항목을 의도적으로 분리하여 최적의 K를 선정하는데 Outlier는 제외하였다.

<Table 13> Top2Vec K별 지표 산정 결과

K	Diversity	Largest	Cos_Avg	Score
20	1	0.4215	0.114	0.8297
25	0.9800	0.4214	0.1085	0.8218
30	0.9800	0.4213	0.1113	0.8213
35	0.9740	0.4213	0.1045	0.8199
40	0.9767	0.4214	0.1117	0.8197
45	0.9750	0.4213	0.1126	0.8188
50	0.9711	0.4213	0.1090	0.8177
55	0.9650	0.4213	0.1009	0.8174
60	0.9650	0.4213	0.0972	0.8174
65	0.9569	0.4213	0.0931	0.8145
70	0.9571	0.4213	0.0150	0.8149

K=20은 전체 후보군 중 종합 평가 점수(score)가 0.8297로 가장 높아 서로 중복되거나 간섭하지 않고 독립적인 의미 체계를 가장 잘 보존하고 있음을 의미하며, 최종 토픽 수로 선정하였다. 그 결과, Blocked는 5,104건(42.70%), Financial Fraud는 1,835건(15.35%), Others는 1,709건(14.30%), Marketplace는 1,622건(13.57%), Pornography는 1,125건(9.41%), Hacking은 304건(2.54%), Cryptocurrency는 253건(2.12%)으로 나타났다<Table 14>.

<Table 14> Top2Vec 전체 토픽별 분류 결과

토픽	범주	라벨	상위 단어	문서 수	비율
0	Blocked	Seized Sites	bavarian, seized, site, police, support, alice, authorities, information, website, anonymous,	5,104	42.70%
1	Others	Independent Tech Blogs/ News Mirrors	censorship, blog, project, projects, github, install, browser, articles, tech, servers, subscribe, internet, organization, development, encrypted	865	7.24%
2	Pornography	Adult Porn Link Aggregators	amateurs, orgasm, bitches,raped, xonions, fairly, tons, cams, assholes, rape, collection	723	6.05%
3	Marketplace	Gift Card Resale Shops	affiliated, giftcardsmarket, privacy, confirmation, cashout, mar, what, email, distributing, provide, carries, thank, deals, replace, leave	672	5.62%
4	Financial Fraud	Money-Transfer/Account Shops (PayPal/Escrow)	moneygram, after, send, need, clean, lordpay, coinlib, own, chargeback, responsiblity, fasttrans, paypal, fastransfers, perfectmoney, harassed	428	3.58%
5	Pornography	CP/Pedo	cp, naked, anal, child, preteen, hebe, nude, boy, porno, amateur, pthc, mom, pics, creampie, gay	402	3.36%
6	Others	Onion Link Lists/ Pirated Media	directory, links, search, tor, markets, engine, wiki, torch, oniondir, coredir, hidden, ahmia, web, engines, urls	391	3.27%
7	Marketplace	Multi-Country Gift Card Shops	netherlands, cyprus, nl, fr, austria, zealand, denmark, spain, kindom, paypals, bazaar, glovo, starbucks, au, giftcardxpress	342	2.86%

8	Financial Fraud	Carding with US Payment Apps	carded, ccv, linkable, cashapp, banknotes, rewarded, zelle, revolut, brands, rating, anonymo, venmo, affiliate, popularity, app	321	2.69%
9	Marketplace	Drug Marketplaces/ Online Pharmacies	cocaine, pills, crystal, mdma, xanax, lsd, meth, oxycodone, heroin, drugs, ecstasy, powder, mg, benzos, amphetamine	304	2.54%
10	Marketplace	Counterfeit Cash/Banknote Production	chf, buyrealmoney, shred, gbp, rub, klyde, timezones, legally, levels, pts, rights, reserved, sjyva, lzxs, those	304	2.54%
11	Hacking	Hacking-for-Hire/ Spoofing Services	hire, hacker, media, recovery, grades, hacking, facebook, recover, instagram, twitter, remotely, vulnerabilities, professional, snapchat, device	304	2.54%
12	Financial Fraud	Cloned-Card/ Western Union Shops	express, replacement, deepnet, atms, master, nucleus, american, evolution, regular, pricing, worlwide, untill, occasionally, including, failure	288	2.41%
13	Others	Anonymous Tor Hosting	congratulations, anon, hosting, hosted, server, by, servers, username, storage, manage, site, sign, provider, host, vps	258	2.16%
14	Cryptocurrency	Bitcoin Scams	allbtcshop, allbtc, xrp, blockchain, ethereum, wallet, litecoin, ripple, generated, negotiation, addresses, decrease, eth, enhance, miners	253	2.12%
15	Financial Fraud	Bulk Stolen Account Shops	stremcloud, ww, locks, sold, incorrect, caused, netauth, otherwise, operation, include, garanteed, provider, bitdefender, vault, buycheck	221	1.85%
16	Financial Fraud	Money/Card Scam Sites	withdrawals, world, work, digit, proofs, yoe, cooocklin, safely, hhd, specify, cftead, ifdeipdttf, wide, blocking, shopping	208	1.74%
17	Financial Fraud	PayPal Account Resale Shops	united, accouts, states, italy, kingdom, db, ee, ed, argentina, marcos, proxies, dc, bigger, ebd, cb	197	1.65%
18	Others	French Activist/ Independent Media	des, mai, pr, que, un, te, du, en, pas, les, der, sur, von, et, dans	195	1.63%
19	Financial Fraud	Stolen Credit-Card Data Dumps	storecard, refresh, estimated, plasticsharks, costly, mean, mistakes, restaurants, skepticism, check, become, glad, massive, few, experienced	172	1.44%
Total				11,952	-

2차 분류 대상에 해당되는 Blocked 총 5,104개 문서에 대해 세부 토픽 분류를 수행하였다. 이전과 동일한 방법으로 평가한 결과, 토픽 개수를 K=3으로 선정하였다. 독일어 기반의 법적 고지 형태인 'German Legal Notice' 패턴이 과반수(51.20%)를 차지하였다. 종합적으로 총 7개의 대주제(Blocked, Financial Fraud, Others, Marketplace, Pornography, Hacking, Cryptocurrency)와 21개의 세부 주제(라벨)로 분류되었다<Table 15>.

<Table 15> Top2Vec Blocked 서브 토픽별 분류 결과

서브토픽	재분류 범주	라벨	상위 단어	문서 수	비율
0	Blocked	German Legal Notice	und, impressum, datenschutz, support, site, op, seized, by, website, alice	2,613	51.20%
1	Blocked	Standard English	anonymous, authorities, information, been, police, this, bavarian, has, alice, website	1,766	34.60%
2	Blocked	Mixed Bilingual	authorities, by, website, bavarian, been, anonymous, datenschutz, police, seized, site	725	14.20%
Total				5,104	-

4.2.3. 평가

임베딩 기반 접근법(BERTopic, Top2Vec)은 고차원 벡터 공간의 의미론적 밀도 구조를 활용하여 단어 빈도 모델의 한계를 극복하고자 하였으나, 군집화 및 차원 축소 알고리즘의 메커니즘 차이에 따른 구조적 제약이 관찰되었다. BERTopic은 노이즈 문서나 혼합 주제를 아웃라이어(Outlier, -1번 토픽)로 별도 유보하였다. 메인 분류에서 Blocked 관련 문서가 대형 토픽을 형성하였으나, 2차 분류를 적용했을 때 사법당국의 압수 고지 형태인 Seized 패턴이 42.80%(5,116건)으로 구분되었다. BERTopic 결과에서는 단어 빈도 모델들이 '금융사기'라는 단일 거대 범주로 포섭했던 마약, 아동성착취물 등 세부주제를 식별해내면서 마약류 마켓플레이스(Marketplace, 20.91%)와 금융사기(Financial Fraud, 14.37%)가 골고루 배치되었다.

반면, Top2Vec은 모든 문서를 공간 상에서 가장 가까운 주제로 강제 할당하므로 아웃라이어가 발생하지 않는 특성이 있다. 코퍼스 내부에서 학습한 Top2Vec 공간에 크게 의존하기 때문에 압수페이지 문구가 반복될 경우, 반복 안내문 패턴이 대형 토픽 형성에 영향을 줄 수 있다. 결과적으로 활성화된 불법 서비스 내에서 금융사기(Financial Fraud, 13.91%)와 마켓플레이스(Marketplace, 13.57%)가 상대적으로 큰 비중의 토픽으로 나타났다<Table 16, 17>.

<Table 16> BERTopic 범주별 사이트 비율

범주		건수(재분류 수)	비율
Marketplace		2,499	20.91%
Financial Fraud		1,717(75)	14.37%
Pornography		1,112	9.30%
Outlier		715	5.98%
Cryptocurrency		282	2.36%
Hacking		197	1.65%
Others		181	1.51%
Blocked	Seized	5,116	42.80%
	Closed	75	0.63%
	Others	58	0.49%
Total		11,952	-

<Table 17> Top2Vec 범주별 사이트 비율

범주		건수	비율
Financial Fraud		1,835	15.35%
Others		1,709	14.30%
Marketplace		1,622	13.57%
Pornography		1,125	9.41%
Hacking		304	2.54%
Cryptocurrency		253	2.12%
Blocked	German Legal Notice	2,613	21.86%
	Standard English	1,766	14.78%
	Mixed Bilingual	725	6.07%
Total		11,952	-

4.3. LLM 기반 토픽 모델링 결과

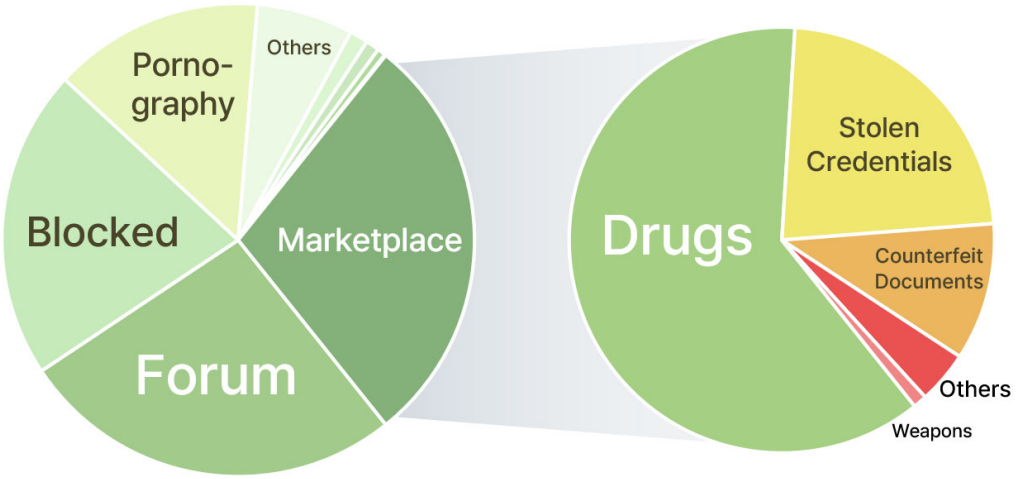
4.3.1. Llama 3.2:3b

TopicGPT에 문서 데이터 셋을 입력하여 스캐닝한 결과, 총 10개의 범주가 추정되었다. 그 중, Whistleblowing, Search, Communication, Personal 등 4개는 의미 중복이 있고 대표성이 부족하여 Others로 통합하였고, 압수/폐쇄 또는 운영을 중단한 경우는 Blocked로 추가 정의하였다<Table 18>.

<Table 18> 분류 범주 목록

구분	TopicGPT 기반 범주 목록		구분	최종 범주 목록
1	Marketplace		1	Marketplace
2	Hacking		2	Hacking
3	Financial Fraud		3	Financial Fraud
4	Pornography		4	Pornography
5	Forum	→	5	Forum
6	Cryptocurrency		6	Cryptocurrency
7	Whistleblowing		7	Others
8	Search			
9	Communication			
10	Personal			
11	-		8	Blocked

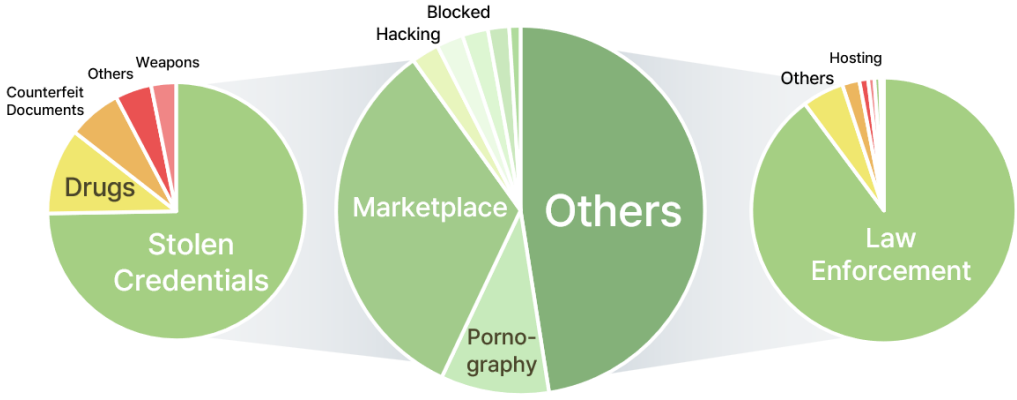
최종 범주를 Llama 모델에 입력하여 수행한 결과, Marketplace가 전체의 40.1%로 가장 높은 비중을 차지하였고, Forum(28.7%), Blocked(19.7%), Pornography(8.2%) 순으로 확인되었다. 이 중에서 Marketplace(총 4,790건)를 대상으로 TopicGPT를 다시 수행한 결과, Drugs, Weapons, Counterfeit Documents, Stolen Credentials 등 4개의 범주가 분류되었다. 추출된 범주가 중복 없이 분류되어 그대로 사용하였다. 결과적으로 Marketplace를 구성하던 사이트의 상당수가 Drugs(61.73%)에 해당하였으며, 그 외 Stolen Credentials(22.78%), Counterfeit Documents(10.46%), Others(3.99%), Weapons(1.04%)의 비율로 구성되어 있음을 확인하였다. 최종적으로 보았을 때, Forum(28.7%), Drugs(24.7%), Blocked(19.7%), Stolen Credentials(9.1%)가 다크웹의 주요 분포를 이루고 있음을 확인하였다<Figure 1>.



<Figure 1> Llama 모델 기반 범주별 사이트 비율

4.3.2. Gemma 3:4b

TopicGPT를 통해 도출된 10개 범주에서 모델을 수행하였다. <Table 18>의 최종 범주를 Gemma 모델에 입력하여 수행한 결과, Others가 전체의 47.5%로 가장 높은 비중을 차지하였고, Marketplace(33.1%), Pornography(9.5%), Hacking(2.4%) 순으로 확인되었다. 이 중에서 Others(총 5,680건)와 Marketplace(총 3,953건)를 대상으로 TopicGPT를 다시 수행한 결과, Law Enforcement, Hosting, Scam, Search Engines, Privacy, Trade, Communication Service, Drugs, Weapons, Counterfeit Documents, Stolen Credentials 등 11개의 범주가 분류되었다. 최종 도출된 19개의 범주 목록을 동일하게 적용하여 모델 기반의 분류를 수행하였다. 분석 결과, Law Enforcement 범주가 전체의 42.7%로 가장 높은 비중을 차지하는 편중이 관찰되었으며, 이는 본 실험 조건에서 Gemma 모델이 압수 안내문 계열 문서를 광범위하게 해당 범주로 수렴시킨 데 따른 것으로 해석된다. 그 뒤를 이어 Stolen Credentials(24.7%), Pornography(9.5%), Drugs(3.6%), Hacking(2.4%) 순으로 분포가 확인되었다<Figure 2>.



<Figure 2> Gemma 모델 기반 Others, Marketplace 세분화 후 범주별 사이트 비율

4.3.3. 평가

LLM 기반 접근법(TopicGPT 프레임워크)은 앞선 통계 모델들이 해결하지 못한 자연어 형태의 후보 레이블 생성을 자동화하여 인간의 주관적 개입을 줄일 수 있었다. 텍스트 내 노출된 개별 어휘의 통계적 빈도에 기계적으로 의존하는 전통적 기법들과 달리, Llama 모델은 문서 전체의 맥락을 바탕으로 사이트의 실질적인 의도와 기능을 반영한 범주를 생성하였다. 그 결과 메인 분류에서 거대 범주로 군집화된 Marketplace를 대상으로 수행한 2차 세분화 연산에서 Drugs, Stolen Credentials, Counterfeit Documents 등의 하위 범주를 중복 없이 정밀하게 식별해 내었다. Llama 모델 결과에서는 Forum, Drugs, Blocked, Stolen Credentials 관련 범주가 상대적으로 큰 비중으로 나타났다. 다만 이 범주들은 서비스 형태, 페이지 상태, 거래 품목이 혼재되어 있어 이를 다크웹 전체 생태계의 구성비로 일반화하기에는 한계가 있다<Table 19>.

<Table 19> Llama 모델 기반 Marketplace 세분화 후 범주별 문서 비율

범주	건수	비율	범주	건수	비율
Forum	3,427	28.7%	Others	248	2.1%
[Marketplace]Drugs	2,957	24.7%	[Marketplace]Others	191	1.6%
Blocked	2,359	19.7%	Hacking	101	0.8%
[Marketplace]Stolen Credentials	1,091	9.1%	[Marketplace]Weapons	50	0.4%
Pornography	981	8.2%	Financial Fraud	23	0.2%
[Marketplace]Counterfeit Documents	501	4.2%	Cryptocurrency	23	0.2%
Total				11,952	-

반면, Gemma 모델의 결과에서는 전체 데이터의 47.5%(5,680건)가 Others 범주로 과도하게 집중되는 양상이 나타났다. Forum과 Cryptocurrency는 각각 1.9%, 1.0%로 상대적으로 낮은 비중으로 분류되었으며, 본 실험 조건에서는 일부 범주 간 분리 수준이 제한적이었다<Table 20>.

<Table 20> Gemma 모델 기반 Others, Marketplace 세분화 후 범주별 문서비율

범주	건수	비율	범주	건수	비율
[Others]Law Enforcement	5,098	42.7%	[Marketplace]Others	181	1.5%
[Marketplace]Stolen Credentials	2,954	24.7%	[Marketplace]Weapons	124	1.0%
Pornography	1,137	9.5%	[Others]Hosting	119	1.0%
[Marketplace]Drugs	427	3.6%	Cryptocurrency	118	1.0%
Hacking	290	2.4%	[Others]Scam	61	0.5%
[Others]Others	289	2.4%	[Others]Search Engines	43	0.4%
Blocked	281	2.4%	[Others]Privacy	40	0.3%
Financial Fraud	270	2.3%	[Others]Trade	13	0.1%
[Marketplace] Counterfeit Documents	267	2.2%	[Others]Communication Service	12	0.1%
Forum	223	1.9%	-		
Total				11,952	-

4.4. 종합분석

데이터 셋 11,952건의 문서를 대상으로 6개의 모델로 주제 분류를 수행한 결과, 각 모델별로

특징을 확인할 수 있었다. 단어 빈도 기반 모델과 임베딩 기반 모델은 모든 문서의 어휘 행렬 및 차원 축소 연산을 대수적으로 일괄 처리하기 때문에, 해당 실험 환경에서 각각 5분 미만 및 20분 내외라는 압도적으로 빠른 연산 속도를 기록하였다. 반대로 LLM 기반 접근법은 개별 문서를 로컬 언어모델의 추론 프로세스에 순차적으로 하나씩 통과시켜야 하는 구조적 한계로 인해, 동일한 데이터셋 처리에 약 8시간이라는 극심한 지연이 발생하였다. 한편, 세 계열에 서로 다른 평가 지표를 적용한 것은 각 계열의 표현 공간에 정의된 표준 관행을 따른 것으로, coherence (C_V/NPMI)는 단어 분포 기반 확률 토픽에, diversity와 outlier 비율은 밀도 기반 클러스터에 각각 정의되는 지표이다. 이들 지표는 모델 간 우열 판단이 아니라 각 모델의 토픽 수(K) 선정 근거로만 활용하였다. 다크웹은 공인된 분류 체계나 라벨이 존재하지 않는 비라벨 코퍼스이므로, 계열 간 비교는 동일 코퍼스에 대한 토픽 세분화 정도, 처리 시간, 라벨 자동화 수준, 상위 단어와 대표 문서의 정성 교차검토라는 계열 횡단 공통 기준으로 수행하였으며, 이는 비라벨 도메인에서 통용되는 평가 방식이다<Table 21>.

<Table 21> 토픽 모델링 세 접근법 비교

구분	단어 빈도(BoW) 기반	임베딩 기반	LLM 기반
사용 모델	LDA, NMF	BERTopic, Top2Vec	TopicGPT + 로컬 LLM(llama, gemma)
토픽 세분화 정도	낮음	높음	중간
토픽 표현	단어 목록	단어 목록	자연어 라벨
라벨 생성	수동	수동	자동
처리 시간	5분 미만	20분 내외	약 8시간

3개의 접근법을 종합적으로 분류하였을 때, 단어 빈도 기반 접근법(LDA, NMF)은 대규모 데이터셋의 대분류(Category) 체계를 빠르게 구조화하는 '거시적 필터링'에 적합하다. 특히, NMF는 수사기관의 압수, 폐쇄 화면을 "bavarian, police, seized" 등 대표단어로 표현하고, 정형화하여 'Seized'와 'Closed'라는 독립적인 인텔리전스 요소로 가르는 데 탁월했다. 그러나 대주제 금융사기(Financial Fraud) 내 세부주제(카드 복제, 페이팔 계정 도용 등)를 독립된 토픽으로 분리하지 못하고 하나의 거대 군집으로 수렴시켰다. 즉, 속도는 빠르나 미세 위협을 포착하는 해상도가 낮고, 인간 분석가가 사후에 라벨을 수동으로 조정해야 하는 공수가 요구된다. 임베딩 기반 접근법(BERTopic, Top2Vec)은 단어 빈도 모델의 한계를 극복하고, 1차 분류에서도 세부주제를 상세하게 분류해내었다. 다크웹 범죄자들이 사용하는 교묘한 은어 맥락을 고차원 벡터 공간 상에서 정교하게 인지함으로써, Bitcoin Scams, CP/Pedo, Cloned-Card 등 수사에 즉시 활용 가능한 수준의 정밀한 라벨링 후보군(총 13~14개 라벨)을 도출해냈다. 오간섭 및 혼합 주제로 인해 아웃라이어(Outlier)로 격리되는 현상이 발생하였으나 이를 참조하여 생성된 유효 토픽들이 서로 중복되지 않고 독립적인 의미 체계를 유지하도록 방어하는 상호작용을 하였다. LLM 기반 접근법(TopicGPT)은 앞선 두 접근법이 해결하지 못한 '라벨 생성의 자동화'를 달성하며 인간의 개입을 최소화하는데 성공하였다. 텍스트 내에 숨겨진 범죄의 '의도(Intent)'와 에스크로(Escrow) 거래 조건 등의 맥락을 인간 수준으로 독해하여 자연어 형태의 범주명을 스스로 부여하는 특징을 보이기도 하였다<Table 22>.

<Table 22> 3개 접근법별 특징 및 차이

비교 항목	단어 빈도 기반 접근법	임베딩 기반 접근법	LLM 기반 접근법
주요 핵심 모델	LDA, NMF	BERTopic, Top2Vec	Llama-3, Gemma-3 (오픈소스 sLLM)
수학적/기술적 메커니즘	베이지안 확률 분포 및 선형대수 행렬 분해	고차원 벡터 공간 정렬 및 밀도 기반 클러스터링 (HDBSCAN)	심층 트랜스포머 기반 문맥 독해 및 프롬프트 엔지니어링
다크웹 은어·노이즈 대응	민감 (단어 형태가 다르면 별개 어휘로 인지)	상대적으로 견고 (은어 및 오탈자의 문맥적 유사성 인지)	매우 견고 (범죄자의 우회 표현 및 다중 맥락 이해)
토픽 분류 해상도 (Granularity)	낮음 (금융사기, 포르노 등 거대 범주로 뭉뚱그림)	높음 (비트코인 스캠, 카드 복제 등 미세 위협 조기 포착)	중간 (마켓, 포럼 등 플랫폼 기능 단위로 우선 수렴)
최종 결과물 표현 방식	단어 가중치 및 상위 단어 목록 (Top-N Words)	단어 공간 거리에 따른 상위 단어 목록 (Top-N Words)	사람이 읽고 즉시 이해 가능한 자연어 주제 구문
범주 라벨링 방식	수동 생성 (인간 분석가의 도메인 지식 개입 필수)	수동 생성 (인간 분석가의 도메인 지식 개입 필수)	후보 라벨 자동 생성 (최종 범주 검토 필요)
1.1만 건 처리 속도	5분 미만 (실시간 대규모 처리에 가장 우수)	20분 내외 (속도와 정밀도의 균형 양호)	8시간 내외 (순차 추론 구조로 인한 극심한 지연)
치명적인 기술적 장벽	짧은 마켓 상품 페이지에서 빈도 계산력 급감	아웃라이어 격리로 인한 누락 또는 중복 문서 과도 집중	윤리 가이드라인(Safety) 충돌로 인한 분석 거부 현상
다크웹 위협 대응 실무 역할	대규모 크롤링 데이터의 고속 1차 필터링 및 대분류 구조화	신종 범죄 및 미세 위협 패턴의 실시간 탐지	선별된 핵심 고위험 위협 데이터의 인텔리전스 보고서 자동화
생태계 해석	사법 압수로 인해 마비된 인프라 위에서, 자금 유통과 가시적인 금융 범죄가 주도	사법 통제의 흔적과 신종 변종 범죄 거래가 팽팽하게 대치	범죄 커뮤니티(Forum)와 마약 유통망(Drugs)을 양대 축으로 구동되는 사회공학적 생태계

다크웹 위협 인텔리전스(CTI) 정보 분석 측면에서 수사 목적과 기술적 자원의 한계에 따라 가이드를 제시할 수 있다. 먼저, 실시간 전수 모니터링 및 대규모 데이터의 초동 조치 단계에서는 단어 빈도 기반(LDA, NMF) 및 임베딩 기반 접근법(BERTopic, Top2Vec)을 채택해야 한다. 대규모 다크웹 크롤링 코퍼스를 5분에서 20분 내외로 신속하게 구조화하고, 사법 인프라성 문서(Blocked)나 대형 위협 범주를 고속 필터링하는 데 기술적 타당성이 가장 높다. 신종 은어 탐지와 위협 탐지의 식별력이 요구된다면 Top2Vec이나 BERTopic을 실시간 엔진으로 배치하는 것이 효율적이다. 인간 분석가의 개입 공수를 극적으로 절감하고 심층적인 위협 프로파일링이 필요하다면, 단독 연산 지연(Latency) 한계를 감수하더라도 성능이 검증된 대형 매개변수 기반의 LLM 접근법을 적용할 필요가 있다. 즉, 토픽 모델링 분석 결과를 종합하면, 현재 다크웹 생태계 약 50%는 사법당국의 압수(Seized)나 자발적 운영 중단(Closed)으로 인해 무력화된 '차단 인프라의 잔해'로 파악된다. 반면, 실제 활성화되어 구동 중인 나머지 50%의 불법 서비스 중에서는 마약류 거래(Marketplace/Drugs)와 금융 사기(Financial Fraud/Cryptocurrency)가 각각 약 15% 내외의 비중을 차지하며 실질적인 지하 경제의 주류 축을 형성하고, 남은 약 20%의 영역은 아동 및 불법 성착취물(Pornography), 계정 탈취 및 개인정보 유통(Hacking/Stolen Credentials), 범죄 모의 포럼(Forum), 다양한 정보 공유 및 기타 쇼핑몰(Others/Other) 등이 파편화된 형태로 생태계의 하부 구조를 구성하고 있다고 할 수 있다.

V. 토픽 모델링 기반 다크웹 위협 대응방법

5.1. 신·변종 위협 조기 탐지 방안 마련

BERTopic과 Top2Vec은 사전 정의된 규칙이나 범주에 귀속되지 않고, 고차원 벡터 공간 내의 밀도 구조를 바탕으로 무기 거래, 비트코인 믹서 등 단어 빈도 기반 모델이 포착하지 못한 미세한 세부 토픽(Niche Threat)을 상대적으로 다수 식별하는 특성을 보였다. 고정된 분류 체계를 우회하여 끊임없이 진화하는 신종 범죄 유형이나 틈새 불법 시장을 포착하는 데 매우 유용한 기술적 자산이다. 실제 다크웹 생태계에서는 추적을 피하기 위한 신종 합성 마약류, 고도화된 악성 툴킷, 서비스형 범죄(Cybercrime-as-a-Service) 등이 지속적으로 출현하므로, 정형화된 위협 범주만으로는 이처럼 급격한 환경 변화를 적시에 반영하기 어렵다. 임베딩 기반 접근법은 문맥적 유사성을 토대로 잠재적 주제를 스스로 발굴하므로, 수사관이 인지하지 못한 변종 위협의 징후를 식별하는 데 최적의 성능을 발휘하였다. 일정 주기마다 수집 데이터셋을 바탕으로 토픽 모델을 연속 학습(Continuous Learning)하여 신규 토픽 클러스터를 자동 갱신하여 활용하면 신종 위협의 확산을 초기 단계에서 조기 차단하고 선제적인 사법 대응법이 될 것으로 예상된다.

5.2. 다중 범죄 생태계 분석을 위한 다계층적 분류 체계 수립

단어 빈도, 임베딩, LLM 기반 기법 모두 단층적 대분류 수준에 그치지 않고, 거대 범주를 대상으로 추가적인 2차 세부 분류(Sub-topic Modeling)를 연계 수행해야만 실질적인 범죄 속성이 명확하게 고해상도로 식별되는 현상을 공통적으로 확인하였다. Llama 모델에서는 전체의 40.1%를 차지하던 Marketplace 범주가 2차 세분화를 통해 Drugs(61.73%), Stolen Credentials(22.78%) 등으로 분리되었고, Gemma 모델에서는 전체의 47.5%가 집중되었던 Others 범주가 세분화 후 Law Enforcement 등 구체적 범주로 재구성되었다. 다크웹 범죄는 개별 웹사이트 단위가 아니라, 배후의 동일 조직이나 상호 연계된 거대 생태계 차원에서 유기적으로 운영되는 경향이 강하므로, 향후의 위협 분석 체계는 단층 분류를 넘어 계층적 분류와 관계성 분석을 결합하는 방향으로 고도화되어야 한다. 1차 분류 모델을 통해 대규모 인프라성 문서와 불법 서비스 대범주를 신속히 구획한 후, 핵심 범죄 그룹에 대해 2차 모델을 실행하여 세부 범죄유형을 계층적으로 명시하는 다단계 분석 체계를 도입할 필요가 있다. 마약 거래·금융 범죄·해킹 조직 간의 지하 연계 생태계를 입체적으로 시각화하는데 지원하여 결과적으로 조직을 와해, 해제시키는데 역할을 할 것으로 기대된다.

5.3. 범죄 유형별 다크웹 모니터링 절차 구축

단어 빈도 기반(LDA·NMF), 임베딩 기반(BERTopic·Top2Vec), LLM 기반 기법 모두 마약, 무기, 금융범죄, 해킹 서비스 등 다크웹 내 핵심 범죄 유형을 유의미하게 구별할 수 있음을 실증하였다. 다만, 대규모 코퍼스로부터 범죄 유형별 통계적 규모를 실시간 집계하는 데에는 연산 효율성이 높은 단어 빈도 및 임베딩 기반 기법이 적합하였으며, 사이트의 궁극적인 운영 형태와 범죄 의도를 독해하는 데에는 LLM 기반 기법이 탁월한 효과를 나타내었다. 향후 다크웹 모니터링 시스템 설계 시 단일 알고리즘에 의존하기보다, 분석 목적과 가용 자원에 따라 적합한 방법론을 결합하는 하이브리드형 접근법이 요구된다는 의미이다. 본 실험에서 실측된 처리시간(단어 빈도 기반 5분 미만, 임베딩 기반 20분 내외, LLM 기반 약 8시간)을 고려하면, 일일 수집되는

대규모 데이터의 일차적인 규모·추이 집계에는 가볍고 신속한 단어 빈도 및 임베딩 모델을 배치하고, 선별된 고위험 사이트의 시장형·포럼형·단일 판매형 등 운영 메커니즘 파악에는 LLM 기법을 적용하는 방식을 사용하는 등 전략적인 모니터링 절차를 구축할 필요가 있다. 특히 본 연구에서 전체 문서의 약 43~47%가 압수·폐쇄된 Blocked 범주로 확인된 만큼, 차단 인프라를 신속히 걸러내고 활성 위협에 치안 자원을 집중하는 우선순위 설정의 정량적 근거로 활용될 수 있다. 정기적 수집, 알고리즘 기반 자동 분류, 대시보드 시각화로 이어지는 통합 연계 분석을 통해 신규 도메인이 실시간으로 생성·소멸되는 휘발성 높은 다크웹 환경에서도 위협 동향을 선제적으로 파악하고 치안 자원 배분 및 단속 우선순위 설정의 객관적 근거 자료로 축적하는 역할을 할 것으로 예상된다.

Ⅵ. 결론

본 연구는 고도의 익명성과 파편화 특성을 지닌 다크웹의 위협 대응을 위해 코퍼스(총 11,952건)를 대상으로, 단어 빈도 기반(LDA, NMF), 임베딩 기반(BERTopic, Top2Vec), LLM 기반(TopicGPT 프레임워크 하의 Llama 3.2, Gemma 3)의 3대 접근법(6개 모델)을 동일 코퍼스에 적용하고 분류 결과와 모델별 특성을 비교·평가하였다. 실험 결과, 단일한 절대 우위의 모델이 존재하기보다 각 접근법의 수학적·아키텍처적 특성에 따라 뚜렷한 기술적 트레이드오프(Trade-off)와 생태계 해석의 다양성이 관찰되었다. 단어 빈도 기반 모델은 5분 미만의 압도적인 연산 속도를 보였으나 세부 범죄 포착 해상도가 낮았고, 임베딩 기반 모델은 고차원 벡터 밀도 공간을 활용하여 수사 직결형 미세 틈새 위협에 강력하였다. LLM 기반 접근법은 텍스트 내부의 의도(Intent)와 조건들을 기반으로 자연어 레이블 후보를 자동 생성하였는데, Llama는 범죄 커뮤니티(Forum)와 마약 유통(Drugs)을 두 축으로 하는 생태계 구조로 범주를 수렴시킨 반면, Gemma는 본 실험 조건에서 전체의 47.5%가 Others 범주로 집중되어 범주 간 분리 수준이 제한적이었다. 다만 이는 프롬프트 및 모델 설정에 따른 분류 경향의 차이로, 두 모델의 일반적 성능 우열로 단정하기는 어렵다. 단어 빈도 및 임베딩 기반 모델들은 사법당국에 의해 '차단 및 폐쇄된 인프라 영역(Blocked, 최대 46.59%)'과 '금융 사기'를 최상위 범주로 식별하였고, LLM 기법(Llama 3.2)은 다크웹 생태계가 사실상 '견고한 범죄 커뮤니티 포럼(Forum, 28.7%)'과 '마약류 거래 마켓플레이스(Drugs, 24.7%)'로 운영되고 있다고 추론하였다. 즉, 거시적인 관점에서는 현재 다크웹 생태계의 약 50%는 사법 통제나 자발적 운영 중단으로 마비된 잔해 영역이며, 활성화된 나머지 50% 중 마약과 금융사기가 각 15% 내외를 차지하고, 남은 20%의 영역을 성착취물, 해킹, 포럼 등이 파편화된 형태로 구성하고 있음을 시사한다. 이를 통해, 사이버위협에 대응하기 위해서 신·변종 위협 조기 탐지 방안을 마련하고, 다중 범죄 생태계 분석을 위한 다계층적 분류 체계 수립, 범죄 유형별 다크웹 모니터링 절차를 구축해야 함을 제시하였다.

다만 본 연구의 해석은 수집된 코퍼스에 한정된다는 한계가 있다. 분석에 사용한 데이터는 특정 원천시드를 기점으로 재귀적 크롤링을 통해 약 8개월간 수집된 것으로, 표본의 성격상 다크웹 생태계 전체를 대표한다고 보기 어렵다. 또한 전처리 과정에서 영어 텍스트만을 추출하고 본문 길이를 기준으로 일부 문서를 제외하였으며, 영상·이미지를 수집 대상에서 배제하였기 때문에 비영어권 사이트, 짧은 게시물, 멀티미디어 중심 콘텐츠 등은 분석에서 누락되었을 가능성이 있다. 따라서 본 연구에서 도출된 범주별 분포와 생태계 해석은 분석 대상 코퍼스의 특성을 반영한 결과이며, 이를 다크웹 전체의 구조나 규모로 일반화하는 데에는 신중할 필요가 있다. 또한, 본 연구는 정답 라벨이 부재한 다크웹 코퍼스의 특성과 불법 콘텐츠에 대한 인간 평가자의

열람이 수반하는 연구윤리적 제약으로 인해, 수작업 라벨 기반의 공통 정량 평가(순도, 정확도 등)는 수행하지 않았다. 합법적으로 검토 가능한 안전 범주에 한정한 소표본 인간 검증을 결합하는 방안은 향후 연구과제로 남긴다. 그럼에도 본 연구는 비정형 다크웹 데이터에 복수의 토픽 모델링 기법을 체계적으로 적용·비교하여 위협 유형의 분포와 모델별 특성을 실증하였다는 점에서 의의가 있으며, 향후 다크웹 배후의 초국가적 범죄 네트워크 전체를 입체적으로 추적·와해하는 실무적 기술 고도화로 확장해 나가는 첫 걸음이 되기를 기대한다.

참고문헌 (References)

[1] Gil MK. 2026. Instagram user data of 17.5 million accounts allegedly leaked and circulated on the dark web [인스타그램 사용자 정보 1,750만 건 유출 다크웹 유통 의혹…국내 사용자 주의]. Daily Secu. Available at: https://www.dailysecu.com/news/articleView.html?idxno=204175 accessed on 2026. 6. 7.

[2] Choi MO. 2025. Mercedes-Benz suffers another massive data breach, including legal and customer confidential information [메르세데스-벤츠, 또 대규모 정보 유출… "법률·고객 기밀 포함"]. SecurityFact. Available at: https://securityfact.co.kr/news/view.php?no=6810 accessed on 2026. 6. 7.

[3] Europol. 2026. IOCTA 2026 – The evolving threat landscape: How encryption, proxies and AI are expanding cybercrime. Publications Office of the European Union.

[4] Freitas S, Gharib A. 2024. Web scale graph mining for cyber threat intelligence. arXiv:2411.06239. https://doi.org/10.48550/arXiv.2411.06239

[5] Lee GB. 2018. A study on Tor-hidden service server forensic case and tracking technique [토르 히든서비스 서버 분석 사례 및 추적기법에 관한 연구]. Journal of Digital Forensics, 12(1), 37-47. https://doi.org/10.22798/kdfs.2018.12.1.37

[6] Byun MS. 2021. Strategies and international coordination for the eradication of digital sex crimes [디지털 성범죄 근절을 위한 대응전략 및 국제공조방안]. Master's thesis. Korea University, Seoul, Korea. https://doi.org/10.23186/korea.000000234562.11009.0001162

[7] Boshmaf Y, Perera I, Kumarasinghe U, et al. 2022. Dizzy: Large-scale crawling and analysis of onion services. arXiv:2209.07202. https://doi.org/10.48550/arXiv.2209.07202

[8] Pastor-Galindo J, Sandlin HÂ, Gómez Mármol F, et al. 2024. A big data architecture for early identification and categorization of dark web sites. Future Generation Computer Systems, 157, 67-81. https://doi.org/10.1016/j.future.2024.03.025

[9] Park WS. 2019. Review of new investigative techniques for regulating child pornography: Focusing on the investigation of child pornography on the dark web [아동음란물 규제를 위한 신종 수사기법 검토: 다크웹상 아동음란물 수사를 중심으로]. Law Review, 19(3), 51-83.

[10] Kim HG. 2022. Comprehensive policy for developing scientific criminal investigation and forensic science (V) [첨단과학수사 정책 및 포렌식기법 종합발전방안연구(V)], pp. 507-508. Korean Institute of Criminology and Justice, Seoul, Korea. Research Report 22-CB-01.

[11] Europol. 2021. The cyber blue line. Europol Spotlight Report. Publications Office of the European Union.

[12] Trautman LJ, Ormerod PC. 2019. Wannacry, ransomware, and the emerging threat to corporations. Tennessee Law Review, 86(2), 6.

[13] Krasznay C. 2020. Case study: The NotPetya campaign. In: Török B, editor. Információ- és kiberbiztonság, pp. 485-499. Ludovika Egyetemi Kiadó.

[14] CISA (Cybersecurity and Infrastructure Security Agency), FBI (Federal Bureau of Investigation), NSA (National Security Agency), et al. 2022. Russian state-sponsored and criminal cyber threats to critical infrastructure. Joint Cybersecurity Advisory AA22-110A. Available at: https://www.cyber.gov.au/about-us/advisories/russian-state-sponsored-and-criminal-cyber-threats-critical-infrastructure accessed on 2026. 6. 7.

[15] Jo JH. 2025. Personal data of 400 million Koreans circulating on the dark web [한국인 개인정보 4억건 다크웹 떠돈다]. Electronic Times (etnews). Available at: https://www.etnews.com/20250306000296 accessed on 2026. 6. 7.

[16] Samtani S, Zhu H, Chen H. 2020. Proactively identifying emerging hacker threats from the dark web: A diachronic graph embedding framework (D-GEF). ACM Transactions on Privacy and Security, 23(4), 21. https://doi.org/10.1145/3409289

[17] Moraliyage H, Sumanasena V, De Silva D, et al. 2022. Multimodal classification of onion services for proactive cyber threat intelligence using explainable deep learning. IEEE Access, 10,

56044-56056. https://doi.org/10.1109/ACCESS.2022.3176965

[18] Shah S, Madisetti VK. 2025. MAD-CTI: Cyber threat intelligence analysis of the dark web using a multi-agent framework. IEEE Access, 13, 40158-40168. https://doi.org/10.1109/ACCESS.2025.3547172

[19] Sarwar MB, Hanif MK, Talib R, et al. 2021. DarkDetect: Darknet traffic detection and categorization using modified convolution-long short-term memory. IEEE Access, 9, 113705-113713. https://doi.org/10.1109/ACCESS.2021.3105000

[20] Mandela N, Sonia, Mistry N, et al. 2025. Efficient dark web traffic classification using a hybrid CNN-LSTM model. International Journal of Information Technology. https://doi.org/10.1007/s41870-025-02427-x

[21] Basheer R, Alkhatib B. 2024. DarkOnto: An ontology construction approach for dark web community discussions through topic modeling and ontology learning. Human Behavior and Emerging Technologies, 2024, 7914028. https://doi.org/10.1155/2024/7914028

[22] Prado-Sánchez VP, Domínguez-Díaz A, De-Marcos L, et al. 2025. Zero-shot classification of illicit dark web content with commercial LLMs: A comparative study on accuracy, human consistency, and inter-model agreement. Electronics, 14(20), 4101. https://doi.org/10.3390/electronics14204101

[23] Shin GY, Kim DW, Park SJ, et al. 2025. Identifying similar users between dark web and surface web using BERTopic and authorship attribution. Electronics, 14(1), 148. https://doi.org/10.3390/electronics14010148

[24] de-Marcos L, Domínguez-Díaz A, Junquera-Sánchez J, et al. 2025. Unveiling dark web identity patterns: A network-based analysis of identification types and communication channels in illicit activities. Information, 16(11), 924. https://doi.org/10.3390/info16110924

[25] Jin P, Kim N, Lee S, et al. 2024. Forensic investigation of the dark web on the Tor network: Pathway toward the surface web. International Journal of Information Security, 23, 331-346. https://doi.org/10.1007/s10207-023-00745-4

[26] Blei DM, Ng AY, Jordan MI. 2003. Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993-1022.

[27] Pham CM, Hoyle A, Sun S, et al. 2023. TopicGPT: A prompt-based topic modeling framework. arXiv:2311.01449. https://doi.org/10.48550/arXiv.2311.01449

[28] Röder M, Both A, Hinneburg A. 2015. Exploring the space of topic coherence measures. Proceedings of the 8th ACM International Conference on Web Search and Data Mining (WSDM), Shanghai, China, pp. 399-408. https://doi.org/10.1145/2684822.2685324

[29] Grootendorst M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794. https://doi.org/10.48550/arXiv.2203.05794

[30] Angelov D. 2020. Top2Vec: Distributed representations of topics. arXiv:2008.09470. https://doi.org/10.48550/arXiv.2008.09470