112 신고 데이터를 활용한 사건 종별 단어 관계성 분석

김민정[1,2], 권은정[1], 변성원[1], 박현호[1], 이민정[1], 정의석[1]
[1]한국전자통신연구원, [2]경찰대학
20190016@police.ac.kr, {ejkwon, swbyon, hyunhopark, minjunglee, esjung}@etri.re.kr

An Analysis of Word Relationship by Using Each Category of 112 Data

Minjeong Kim[1,2], Eunjung Kwon[1], Sungwon Byon[1], Hyunho Park[1], Minjung Lee[1],Eui-suk Jung[1]
[1]ETRI, [2]Korean National Police University

요 약

본 논문에서는 과거 발생한 112 신고내용을 기반으로 새로운 신고 발생 시 해당 신고의 유형과 내용상 특징을 파악하고, 유사한 신고에 대한 상황분석을 용이하게 하고자 사건 종별 단어 관계성 분석을 진행하였다. Sentence BERT 기반으로 토픽 모델링을 수행하는 BERTopic 모델을 활용하여 112 신고 내에서 사용된 단어의 출현 빈도수에 따라 네트워크 클러스터링을 함으로써 비슷한 유형의 단어끼리 분류됨을 시각적으로 확인하였고, 112 신고의 사건 종별 단어 간 관계성 분석이 가능하다는 결과를 도출하였다. 본 논문의 사건 종별 단어 관계성 분석을 신고상황 분석에 활용한다면 더욱 정확한 112 신고 상황분석이 가능할 것으로 기대한다.

Ⅰ. 서 론

112 신고는 신고자가 처한 위기의 상황을 알리기 위한 수단이며, 112 신고를 통한 사건 상황 파악은 경찰의 핵심 업무로 꼽힌다. 112 신고를 접수하는 순간부터 짧은 시간 동안 긴박한 사건 상황을 분석하고, 분석한 사건 정보를 현장 경찰관에게 전달하는 기술이 필요하다 [1]. 사건 종별 112 신고 데이터에서 주로 쓰인 단어와 단어의 관계성을 분석한다면, 더욱 정확하게 사건 상황을 분석할 수 있을 것이다.

본 논문에서는 BERTopic을 통해 112 신고 데이터를 활용한 사건 종별 단어 관계성 분석을 수행하였다. 신고내용을 분석하기 위해서는 단어 간의 관계성을 파악하여야 하는데, 이를 위해 사회 연결망 분석(SNA: Social Network Analysis) [2, 3]을 수행할 수 있다. SNA 기법 중 하나인 클러스터링은 노드들의 속성 간 유사성을 계산하기에 용이하다. 한편 다량의 텍스트에 대한 분석에서는 텍스트 마이닝 기법이 유리하며, 그중 토픽 모델링은 텍스트 내에 존재하는 키워드를 추출하는 데 효과적인 방법이다. 토픽 모델링의 방법으로는 LSA(Latent Semantic Analysis), LDA(Latent Dirichlet Allocation), BERTopic 등이 있다. BERTopic [4]은 텍스트 데이터를 SBERT(Sentence BERT)로 임베딩하여 문서를 군집화한 후 토픽 표현을 생성함으로써 토픽 모델링을 진행하는 모델이며, BERT를 통한 112 신고 데이터 분석과 연동하여 사용할 수 있다. 따라서 본 논문에서는 BERT를 통한 112 신고 데이터 분석을 용이하게 하기 위하여, BERTopic을 통해 112 신고 데이터의 사건 종별 관계성 분석을 수행하였다. BERTopic을 활용하여 신고 텍스트에서 토픽이 되는 단어를 추출하고 신고 카테고리별 단어 출현 빈도를 활용하여 중심성 분석을 진행하였다. 그리고, 본 논문에서 신고내용 내 단어 간의 관계성 및 단어 빈도수 기반 연구의 유의미성을 연구하였다.

Ⅱ. 112 신고 데이터의 사건 종별 단어 관계성 분석

(1) 데이터 소개 및 전처리

본 논문에서는 112 신고 데이터를 사용하여 연구를 진행하였다. 활용한 112 신고 데이터는 1만 1천여 개의 신고내용을 담고 있으며, 58가지 사건 종별로 구성되어 있다. 이 중 5가지 카테고리(교통사고, 기타 경범, 기타형사범, 소음, 풍속영업)의 데이터를 추출하여 토픽 모델링에 활용하였다. 5가지 사건 종별 데이터의 수는 그림 1과 같으며, 사건 종별에 따라 신고내용에 포함된 단어의 개수나 신고 개수에서 편차가 존재하였다.

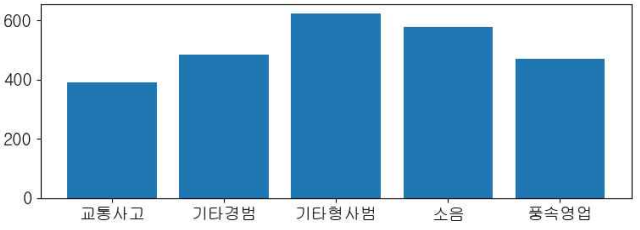

그림 1. 사건 5개 종에 대한 신고 데이터 수

신고 데이터에는 소방 공동 대응, 관할 변경 등 절차상 기록되는 단어가 있으며, 신고자가 말한 내용을 기록하면서 발생한 오타 및 불필요한 수식어가 존재한다. 그 외 경찰에서 사용하는 용어 및 은어도 존재하였다. 그렇기에 경찰에서 사용하는 용어를 포함하면서도 사건과 직접적인 관련이 없는 단어를 제거하였고, 시스템상 반복되는 단어를 합치도록 하였다.

그다음 과거 트위터 형태소 분석기였던 Okt(Open Korean Text) 모듈을 이용하여 신고내용 중 명사와 동사만을 추출하여 입력 데이터를 만들었다. 신고자의 감정이나 상태를 표시하는 형용사나 부사 등으로 인해 품사를 제한할 필요가 있었다.

(2) 전체 112 데이터에 대한 토픽 모델링

BERTopic을 사용하여 신고 데이터에 대한 토픽 모델링을 진행할 경우 모델이 설정한 토픽에 따른 단어 묶음이 도출되는 것을 확인하였다. 그러나 이러한 결과는 112 신고 데이터의 카테고리에 맞게 분류되지 않아서 적용하기 어려웠다. 그렇기에 카테고리별로 데이터를 구분하여 학습시키는 준지도 학습 이후 다시 토픽 모델링을 진행하였다. 학습 이전에는 토픽 모델링의 주제

응집도를 측정하는 값인 c_v Coherence가 0.47이었는데, 학습 이후 0.48로 올라갔다. 또한 토픽별 단어들을 살펴보았을 때 토픽 간의 상·하위 관계 및 포함·연관관계가 훨씬 뚜렷하게 나타남을 확인할 수 있었다.

(3) 카테고리별 데이터에 대한 토픽 모델링

앞서 추출한 5가지 카테고리 데이터에 대하여 BERTopic을 적용함으로써 각 카테고리에 포함된 단어를 분류하였다. 예시로 교통사고에 관한 결과를 아래 그림 2와 그림 3으로 첨부하였다. 이는 c-TF-IDF 값을 바탕으로 분류되었으며, 문서 간 단어의 중요도를 비교하는 값인 TF-IDF를 클래스 기반으로 정의한 지표다. 그림 2는 c-TF-IDF 값에 따라 토픽을 2차원 그래프에 그린 Intertopic Distance Map이고, 그림 3은 토픽 내의 단어를 순위별로 시각화한 결과다.

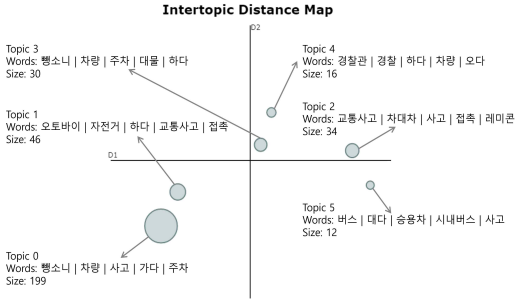

그림 2. 교통사고 관련 Intertopic Distance Map

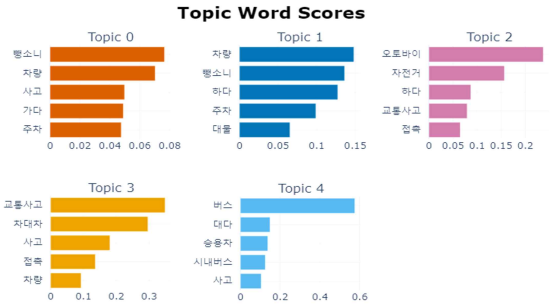

그림 3. 교통사고 관련 Topic Word Scores

그림 3에 나타난 단어를 바탕으로 각 토픽의 이름을 재정의하여 표 1과 같이 작성하였다. 여기서 토픽 및 토픽별 단어는 BERTopic에서 도출된 내용이며, 토픽명 재정의 부분은 단어의 내용을 살펴본 뒤 교통사고와 관련이 있는 세부 주제로 묶은 것이다.

표 1. 교통사고에 대한 토픽별 단어 및 토픽 재설정

토픽	토픽별 단어	토픽명 재정의
Topic 0	뺑소니-차량-사고-가다-주차	뺑소니 사고
Topic 1	오토바이-자전거-하다-교통사고-접촉	오토바이/자전거 사고
Topic 2	교통사고-차대차-사고-접촉-레미콘	차대차 접촉사고
Topic 3	뺑소니-차량-주차-대물-하다	주차 뺑소니
Topic 4	경찰관-경찰-하다-차량-오다	경찰 대응
Topic 5	버스-대다-승용차-시내버스-사고	버스 사고

(4) 네트워크 분석 및 시각화

교통사고 신고내용에서 나타난 단어 간의 관계를 데이터 분석 도구인 Gephi를 이용하여 그림 4와 같이 시각화하였다. Modularity 값을 기반으로 단어를 클러스터링하였으며, 10가지의 클래스로 구분된 것을 그림 4의 색상을 통해 확인하였다.

그림 5는 그림 4에서의 Modularity 클래스 0번에 속한 단어들을 보여주는데, 뺑소니와 차량, 도망, 추격 등이 주요 단어로 나타났다. 이는 이웃한 노드들의 중요도를 통해 해당 노드의 중심성을 계산하는 고유벡터 중심성(Eigenvector Centrality) 값에 비례하여 글자 크기를 나타내도록 하였다.

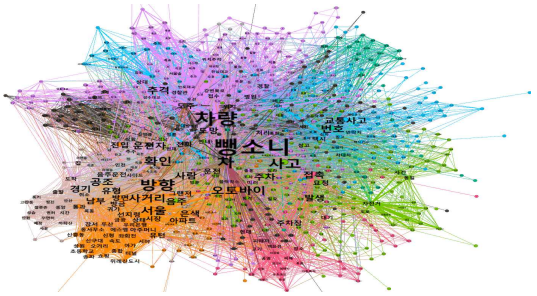

그림 4. 교통사고 신고내용 내의 단어 간 네트워크 분석

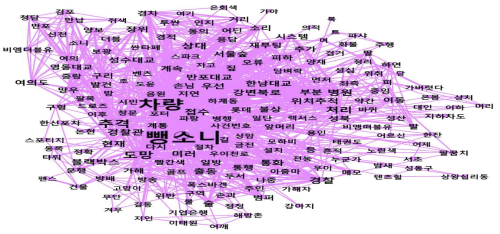

그림 5. 교통사고 신고내용 Modulity Class 0번

Ⅲ. 결 론

본 논문에서는 BERTopic을 활용한 토픽 모델링 및 단어 간 네트워크 클러스터링을 진행하였다. BERTopic으로 토픽 모델링을 진행함으로써 112 신고 전체 데이터보다는 신고내용 카테고리별로 학습 및 분류를 진행하는 것이 더 일관성 있는 결과를 도출함을 알 수 있었다. 그리고 112 신고에 대하여 단어의 출현 빈도를 기반으로 클러스터링한 결과 비슷한 유형의 단어가 같은 클러스터로 묶인다는 점을 확인하였다. 112 신고에 대한 사건 종별 단어 출현 빈도 분석 기술이 향후 112 신고 데이터를 이용한 상황 분석 연구에 도움이 되리라 기대한다.

ACKNOWLEDGMENT

이 논문은 2023년도 정부(경찰청)의 재원으로 지원받아 수행된 연구 결과임 [내역사업명: 112 긴급출동 의사결정 지원 시스템 / 연구개발과제번호: PR08-03-000-21]

참 고 문 헌

[1] 홍세은, 백명선, 변성원, 권은정, 박현호, 정의석, 이용태. "긴급신고 사건유형 및 긴급성 추론기술과 현장대응 정보제공을 통한 의사결정 지원 기술 연구." 한국통신학회 학술대회논문집, pp. 1866-1868.

[2] Linton C. Freeman. “Centrality in Social Networks Conceptual Clarification. Social Networks”, *Social Networks*, vol. 1 no. 3, pp. 215-239, 1978.

[3] IBM Corporation. “IBM i2 Analyst’s Notebook Social Network Analysis”, IBM Corporation, Somers: NY, Technical Report ZZW 03174-USEN-01, 2012.

[4] Grootendorst, M.. “BERTopic: Neural topic modeling with a class-based TF-IDF procedure”. arXiv:2203.05794, 2022.