기업 정보보안 사고의 분쟁 유형 도출; BERTopic, Top2Vec, LDA 기반 토픽모델링의 성능 평가를 중심으로

박민정 [1], 손영진 [2], 채상미 [3]
[1] 금오공과대학교 경영학과 교수
[2] 이화여자대학교 경영학과 박사수료
[3] 이화여자대학교 경영학과 교수

mjpark@kumoh.ac.kr, teumdal@ewhain.net, smchai@ewha.ac.kr

Identify Dispute Types of Corporate Information Security Incidents; Focusing on Performance Evaluation of BERTopic, Top2Vec, and LDA-based Topic Modeling

Minjung Park[1], Young Jin Son[2], Sangmi Chai[3]
[1]Dept. of Business Administration, Kumoh National Institute of Technology
[2]Dept. of Business Administration, Ewha Womans University
[3]Dept. of Business Administration, Ewha Womans University

요 약

최근 AI 를 비롯한 데이터 기반의 비즈니스 모델 증가에 따라, 데이터 유출 등의 기업 정보보안 사고가 빈번하게 발생하고 있다. 해당 사고들은 종종 법적 분쟁으로 이어지며, 이는 기업의 막대한 경제적 손실을 초래하며 정보보안 사고를 선제적으로 대비하기 위한 기술적, 관리적 조치 마련을 위한 기업의 관심이 증가하고 있다. 이에 본 연구에서는 최근 들어 급증한 기업의 정보보안 관련 판례를 대상으로 BERTopic, Top2Vec, LDA 를 활용하여 토픽 모델링을 수행하여 산출된 토픽 기반의 기업 정보보안 사고를 유형화하고자 한다. 전통적으로 각각 다른 법적 요소와 판결을 담고 있어, 유사 사건 간의 비교 및 분석이 어려운 판례 데이터의 특징을 반영하여 본 연구에서는 앞서 제시된 3 가지의 모델을 각각 적용한다. 이를 통하여 각 모델 수행 결과의 성능 비교를 통하여 기업의 정보보안 사건의 유형화 및 동향을 파악하는 동시에 판례 데이터를 분석하기 위한 최적의 모델을 확인한다.

1. 연구배경

최근 디지털 기술의 발전과 데이터 중심의 비즈니스 모델이 확산됨에 따라 기업들의 정보보안에 대한 관심이 급증하고 있는 추세이다. 그러나 이와 동시에 기업 정보보안 사고의 발생 빈도도 급격히 증가하고 있으며, 이로 인한 법적 분쟁 또한 늘어나고 있다 [1]. 특히, 내부 정보 유출, 개인정보보호 실패, 시스템 접근 권한 탈취 등 기업의 정보보안 사고들의 유형이 다양해지고 있다. 이에 본 연구는 기업 정보보안 사고와 관련된 실제 판례 텍스트 데이터를 대상으로 토픽 모델링을 수행하여 이를 유형화하는 동시에 각 모델별 성능을 비교하여 발생된 기업의 정보보안 사고를 유형화하기 위한 최적의 모델을 찾고자 한다. 이를 통하여 본 연구는 기업의 정보보안 관련 법적 분쟁의 유형을 체계적으로 분석하여 이를 유형화함에 따라 기업의 효율적인 선제적 대응 전략 수립을 위한 기반을 마련하는 것에 주요 목적이 있다.

2. 토픽 모델링

토픽 모델링은 대규모 문서 집합을 분석하여 각 문서에 나타나는 주요 주제를 식별하고, 해당 주제들을 바탕으로 문서들의 정리 및 요약할 수 있도록 한다 [2]. 즉, 토픽 모델링은 문서를 특정 주제에 따라 그룹화한다는 점에서 각 주제는 특정 단어들의 집합으로 표현한다. 토픽 모델링을 수행하는 각 모델은 문서 내 단어의 분포를 분석하여, 어떤 단어들이 주로

같이 나타나는지를 바탕으로 토픽을 추론한다 [2,6].

2.1.1 LDA(Latent Dirichlet Allocation)

LDA 는 문서 내의 단어들이 각각 특정 토픽에 속할 가능성 및 문서 자체가 특정 토픽에 얼마나 속할 가능성이 있는지를 계산함으로써 문서의 주제를 파악하고 비슷한 주제를 가진 문서들을 그룹화하는 데 사용된다. 또한, 단어의 출현 빈도와 단어들 간의 관계를 통해 토픽을 도출한다 [3]. LDA 는 확률적 토픽 모델로서, 문서가 토픽들의 혼합으로 구성되어 있다고 가정하고, 각 토픽이 특정 단어들의 확률 분포를 모델링한다. 그러나 LDA 는 단어의 순서나 문장의 구조를 고려하지 않고 단어들을 독립적으로 처리한다는 점에서 토픽 추정의 일관성이 보장되기 어려운 단점이 존재한다 [5]. 이는 분류 분석의 타당성을 저하시키는 주요 요인이 된다 [4]. 또한, LDA 는 여러 파라미터를 조정 해야함에 따라, 계산 복잡도가 증가하고 모델 학습에 많은 시간이 소요된다.

2.1.2 Top2Vec

Top2Vec 은 문서 자체에서 자동으로 토픽을 학습하는 방법으로, 문서와 단어의 벡터를 함께 임베딩하여 토픽을 도출한다. Top2Vec 은 문서와 단어를 고차원 벡터로 변환하고, 이 벡터들을 사용하여 의미적으로 유사한 문서를 클러스터링한다 [7,8]. 해당 과정을 통해 각 클러스터는 하나의 주제를 나타내게 되며, 각 토픽에 대한 핵심 단어들이 자동으로 추출된다 [7].

2.1.3 BERTopic

BERTopic 은 딥러닝 기반의 토픽 모델로, BERT 와 같은 사전 훈련된 언어 모델을 사용하여 문서들 사이 의미론적 유사성을 기반으로 토픽을 생성한다 [2,9].

BERTopic 은 사전 훈련된 Transformer 기반 언어 모델로 문서 임베딩 생성, 임베딩의 클러스터링, 클래스 기반 c-TF-IDF 의 순서에 따라 토픽을 생성한다 [2]. 이와 같은 방식을 갖춘 BERTopic 은 차원축소와 클러스터링을 바탕으로 복잡한 언어 패턴과 맥락을 이해하고, 이를 바탕으로 고차원적인 토픽 구조를 추출할 수 있다는 점에서 기존의 토픽모델링과 차이를 갖는다 [10]. 또한 문서 내부에 노이즈를 일으키는 HTML Tag 등이 포함되지 않는 동시에 문서 전체의 General Topic 을 파악하기 위해 전처리를 수반하지 않는다는 장점이 있어 최근 다양한 연구에서 활용되는 추세이다 [2].

3. 연구 방법

본 연구는 미국의 법률 정보 데이터베이스인 Westlaw 를 활용하여 'cyber security', 'information breach', 'data theft', 'privacy violation' 등의 키워드를 사용하여 최근 5 년 간의 기업 정보보안 관련 판례를 수집한다. 수집된 텍스트 데이터를 바탕으로 각 BERTopic, LDA, Top2Vec 의 모델을 적용하여 발생된 기업 정보보안 사고를 유형화하는 동시에 주요 토픽을 확인한다. 나아가, 본 연구는 세 모델을 대상으로 복잡도(perplexity)와 토픽 일관성 척도(topic coherence)의 지표를 적용하여 토픽 모델의 성능 평가를 수행한다. 복잡도는 각 모델을 통하여 도출된 토픽이 얼마나 전체 문서를 잘 설명하는지를 측정함에 따라, 해당 값이 낮을수록 모델의 우수성을 의미한다. 또한, 토픽 일관성 척도는 토픽 내의 단어들이 얼마나 의미론적으로 일관성 있는지를 평가한다. 이를 통하여 본 연구의 결과는 실제 기업의 정보보안 사고를 유형화하고 관련 동향을 확인하는데 가장 적합한 토픽 모델링 모델을 확인하는 동시에 향후, 기업의 정보보안 사고를 예방하기 위한 전략 수립을 위한 기반을 제공할 수 있을 것으로 예상된다.

참고문헌

[1] Patterson, Clare M., Jason RC Nurse, and Virginia NL Franqueira. "Learning from cyber security incidents: A systematic review and future research agenda." Computers & Security (2023): 103309.

[2] Grootendorst, Maarten. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure." arXiv preprint arXiv:2203.05794 (2022).

[3] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.

[4] Egger, R., & Yu, J. (2021). Identifying hidden semantic structures in Instagram data: a topic modelling comparison. Tourism Review, 77(4), 1234-1246.

[5] Yu, Dejian, and Bo Xiang. "Discovering topics and trends in the field of Artificial Intelligence: Using LDA topic modeling." Expert Systems with Applications (2023): 120114.

[6] Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., & Hassan, A. (2023). Topic modeling algorithms and applications: A survey. Information Systems, 112, 102131.

[7] Gan, Lin, et al. "Experimental Comparison of Three Topic Modeling Methods with LDA, Top2Vec and BERTopic." International Symposium on Artificial Intelligence and Robotics. Singapore: Springer Nature Singapore, 2023.

[8] Zengul, Ferhat, et al. "A practical and empirical comparison of three topic modeling methods using a COVID-19 corpus: LSA, LDA, and Top2Vec." (2023).

[9] Borčin, Martin, and Joemon M. Jose. "Optimizing BERTopic: Analysis and Reproducibility Study of Parameter Influences on Topic Modeling." European Conference on Information Retrieval. Cham: Springer Nature Switzerland, 2024.
[10] An, Yusung, Hayoung Oh, and Joosik Lee. "Marketing insights from reviews using topic modeling with BERTopic and deep clustering network." Applied Sciences 13.16 (2023): 9443.