KSCE Journal of Civil and Environmental Engineering Research
Vol. 43, No. 6: 811-819/ December, 2023
DOI: https://doi.org/10.12652/Ksce.2023.43.6.0811

ISSN 1015-6348 (Print)
ISSN 2799-9629 (Online)
www.kscejournal.or.kr

Construction Management **시공관리**

LDA 및 BERTopic 기반 해외건설시장 뉴스 기사 토픽모델링 성능평가

백준우* · 정세환** · 지석호***
Baik, Joonwoo*, Chung, Sehwan**, Chi, Seokho***

Evaluation of Topic Modeling Performance for Overseas Construction Market Analysis Using LDA and BERTopic on News Articles

ABSTRACT

Understanding the local conditions is a crucial factor in enhancing the success potential of overseas construction projects. This can be achieved through the analysis of news articles of the target market using topic modeling techniques. In this study, the authors aimed to analyze news articles using two topic modeling methods, namely Latent Dirichlet Allocation (LDA) and BERTopic, in order to determine the optimal approach for market condition analysis. To evaluate the alignment between the generated topics and the actual themes of the news documents, the research collected 6,273 BBC news articles, created ground truth data for individual news article topics, and finally compared this ground truth with the results of the topic modeling. The F1 score for LDA was 0.011, while BERTopic achieved a score of 0.244. These results indicate that BERTopic more accurately reflected the actual topics of news articles, making it more effective for understanding the overseas construction market.

Keywords : Overseas construction, News article, Topic modeling, Ground truth, Performance evaluation

초 록

해외건설사업 시, 현지 상황을 정확하고 빠르게 파악하는 것은 프로젝트 성공을 위해 매우 중요한 요소이다. 이는 토픽모델링을 활용한 뉴스 기사 분석을 통해 실현될 수 있다. 본 연구는 Latent Dirichlet Allocation(LDA)과 BERTopic 두 토픽모델링 기법을 활용하여 뉴스 기사를 분석하고, 최적의 기법을 찾고자 하였다. 모델링 결과로 자동생성된 토픽과 실제 문서 주제와의 일치 여부를 확인하기 위해 BBC 뉴스 기사 6,273건을 수집하여 ground truth를 생성하고, 이를 모델링된 토픽과 비교하였다. 그 결과 LDA의 F1 score는 0.011, BERTopic은 0.244로 나타났다. 이를 통해 BERTopic이 실제 뉴스 기사의 주제를 잘 파악하며, 해외건설시장의 주요 이슈를 자동으로 이해하는 데 더욱 용이하다는 것을 확인할 수 있었다.

검색어 : 해외건설, 뉴스 기사, 토픽모델링, Ground truth, 성능평가

1. 서 론

해외건설사업 시 현지 시장 상황을 신속하게 파악하는 것은 사업의 성공적인 수행을 위해 매우 중요하다(Javernick-Will and Scott, 2010). 사업 발주국의 뉴스 기사에는 현지의 다양한 사건들이 서술되므로, 이를 분석하여 유용한 정보를 추출하고 시장 상황을 파악할 수 있다(Goldszmidt et al., 2011). 그러나 뉴스 기사의 양이 매우 많기 때문에 모든 기사를 일일이 분석할 수 없다는

* 서울대학교 건설환경공학부 석박통합과정 (Seoul National University · matthewbaik@snu.ac.kr)
** 종신회원 · 서울대학교 건설환경공학부 박사과정 (Seoul National University · hwani751@snu.ac.kr)
*** 종신회원 · 교신저자 · 서울대학교 건설환경공학부 교수 (**Corresponding Author** · Seoul National University · shchi@snu.ac.kr)

Received August 8, 2023/ **revised** September 12, 2023/ **accepted** September 14, 2023

문제점이 있다.

이를 해결하기 위해 많은 양의 문서 집합을 자동으로 요약하여 토픽을 추출하는 토픽모델링 방법을 사용할 수 있다(Blei, 2012). 건설 분야에서는 토픽모델링을 사용해 결함 소송 사례(Jallan et al., 2019), The World Bank 뉴스 기사(Moon et al., 2018) 등의 텍스트에서 주요 이슈를 도출하는 연구가 수행되었다. 토픽모델링의 대표적인 기법으로는 Latent Dirichlet Allocation(LDA)와 BERTopic이 있다. LDA는 2003년에 제안되었으며(Blei et al., 2003), 최근까지 가장 보편적으로 사용되는 기법으로 알려져 있다. BERTopic은 비교적 최근에 제안되었으며, 다른 토픽모델링 기법 대비 우수한 성능으로 인해 자연어처리 연구 분야에서 LDA보다 최근 더 많은 관심을 받고 있다(Abuzayed and Al-Khalifa, 2021; Grootendorst, 2022).

그러나 기존의 성능평가 방식으로는 예측된 토픽이 실제 문서의 주제와 일치하는지 알 수 없다는 한계가 있다. 이를 극복하기 위해, 본 연구에서는 뉴스 기사를 직접 확인하여 기사별 주제(ground truth)를 생성하고, 토픽모델링을 적용한 결과를 개별 기사 수준에서 비교하여 성능을 평가하는 방법론을 제안한다. 제안된 성능평가 방법론을 LDA와 BERTopic 기법에 적용함으로써, 둘 중 어떤 기법이 실제 문서의 주제를 더욱 명확하게 식별할 수 있는지 찾아내고자 한다. 제안된 방법은 해외건설사업 수행 시 현지 상황을 신속하게 파악하는 데 유용하게 활용될 것이다.

2. 문헌 고찰

2.1 건설 분야에서의 토픽모델링

토픽모델링은 문서에 대한 사전 정보 없이 텍스트 데이터의 주제를 자동으로 추출한다. 이러한 장점 때문에 이미지, 소셜 네트워크, 유전자 등 다양한 응용 분야의 데이터에서 패턴을 찾는 데 활용되고 있다(Blei, 2012). 건설 분야에서도 토픽모델링을 활용한 여러 연구가 수행되었다. 국내에서는 건설 프로젝트의 Building Information Modeling(BIM) 사용 유무를 분류하는 연구가 수행되었고(Jung and Lee, 2019), The World Bank 뉴스 기사를 분석하여 해외건설사업 시 발생하는 주요 이슈를 파악하는 연구도 수행되었다(Moon et al., 2018). 해외의 경우 건설 프로젝트에서 발생한 결함 관련 소송 사례로부터 품질 문제 발생 원인을 파악하는 연구(Jallan et al., 2019)와, 중국 Three Gorges Dam 건설 프로젝트에 대한 학계의 주요 우려 사항을 도출하는 연구가 수행되었다(Jiang et al., 2016).

2.2 토픽모델링의 성능평가

토픽모델링 결과에 대한 성능평가는 다양한 방식으로 수행되었다. 초기 연구들은 정확도를 측정하고(Wei and Croft, 2006), 혼잡도를 계산하여 모델의 성능을 평가하였으며(Wallach et al., 2009), 이후 연구들은 토픽 내의 키워드들 간 의미적 유사성을 나타내는 지표인 주제 일관성(Topic coherence)으로 모델의 성능을 평가하였다(Newman et al., 2010). 더불어, 토픽의 다양성을 나타내는 주제 다양성(Topic diversity)을 활용하여 LDA와 BERTopic 모델링 결과를 평가한 연구도 여럿 수행되었다. BBC 뉴스 기사, 20 NewsGroups과 Donald Trump의 트윗 데이터(Grootendorst, 2022), 아랍어 문서 데이터(Abuzayed and Al-Khalifa, 2021)로 모델링한 연구들에서 BERTopic이 LDA보다 높은 주제의 다양성과 일관성을 보였다.

그러나 이전 연구에서 사용된 지표인 다양성과 일관성으로는 실제 문서의 토픽과 예측 결과가 일치하는지 확인할 수 없다는 한계가 있다. 주제의 다양성 지표는 토픽모델링 결과로 생성된 토픽들이 다양한 주제를 포함할수록 높은 값을 가지나, 주제의 다양성이 높다고 해서 식별된 주제가 반드시 문서의 실제 주제와 일치한다고 할 수 없다. 주제의 일관성은 토픽 내 키워드들이 의미적으로 유사할수록 높은 값을 갖는데, 모델이 잘못된 단어나 문맥에 기반하여 일관되게 잘못된 키워드를 생성할 수 있어 실제 문서 주제와의 일치 여부를 확인하기 어렵다. 본 연구에서는 이러한 한계를 극복하기 위해 직접 뉴스 기사를 확인하여 기사별 주제의 ground truth를 생성하고, 기사 단위로 예측 결과와 비교함으로써 실제 문서의 주제와 생성된 토픽의 일치 여부를 판단하는 방식으로 토픽모델링의 성능을 검증하였다.

3. 연구 방법

본 연구는 아래 Fig. 1과 같이 총 5단계로 구성된다. 우선 뉴스 기사를 수집하고, 각 기사의 제목과 본문 내용을 분석하여 토픽에 대한 ground truth 주제를 생성하였다. 그 후 LDA, BERTopic 두 기법을 사용하여 토픽모델링을 수행하였다. 이를 통해 기사별로 생성된 키워드를 확인하고, 해당 기사에 적합한 토픽을 배정하였다.

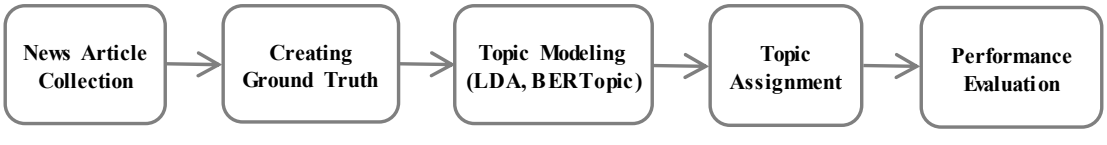

Fig. 1. Flow Diagram of Research Methodology

마지막으로, ground truth와 모델링된 토픽을 비교하여 성능평가를 실시하였다.

3.1 뉴스 기사 수집 및 ground truth 생성

본 연구는 웹에서 뉴스 기사를 수집하기 위해 특정 웹 페이지에서 원하는 정보를 추출하는 web scraping 기법을 사용하였다(Glez-Peña et al., 2014). 웹 사이트는 Hypertext Markup Language(HTML) 언어로 작성되어 있으며, HTML 언어는 글꼴, 사운드, 그래픽 콘텐츠 등 문서와 관련된 정보를 나타내는 태그로 구성된다(예. 뉴스 제목). 따라서, 원하는 정보를 담고 있는 태그(예.)만을 선별함으로써 원하는 정보(예. 뉴스 제목)를 추출할 수 있다. 본 연구에서는 뉴스를 제공하는 웹 사이트의 모든 웹 페이지 Uniform Resource Locator(URL)을 얻은 뒤, 각각의 URL에 접속하여 각 뉴스 기사의 제목, 본문, 날짜를 자동으로 수집하였다. 이후, 수집한 뉴스 기사 데이터를 직접 확인하면서 뉴스 토픽 리스트를 생성하고 각각의 기사에 토픽을 배정하여 토픽모델링 성능평가를 위한 ground truth를 생성하였다.

Ground truth 생성은 분류체계 초안 작성, 토픽 리스트 작성, 토픽 리스트 기반 나머지 기사 분류, 최종 검토 및 재분류의 과정으로 수행하였다. 먼저 1~500번 기사의 제목과 본문을 읽고 각각의 토픽을 지정하여 토픽 분류체계의 초안을 작성하였다. 예를 들어, 1번 기사(Table 1)는 도로 폐쇄 및 재개방과 관련된 내용을 담고 있어 "도로/교통" 카테고리로 분류하였다. 다음으로, 분류체계 초안에 따라 토픽 리스트를 작성하였다. 개별 기사의 토픽은 소분류로 지정하고, 유사한 소분류를 대분류로 묶어 표기하였다. 예를 들어, "정치" 대분류 안에는 "국회/정당", "법안", "선거" 등의 소분류가 존재한다. 이어서 토픽 리스트에 따라 나머지 기사들을 분류하였다. 중간에 새로운 토픽이 필요할 경우 토픽 리스트에 추가하였고, 앞서 이미 분류했던 기사 중 새롭게 추가된 토픽에 해당하는 기사가 존재하는지 검토한 후 업데이트하는 과정을 거쳤다. 이렇게 전체 기사의 1차 토픽 배정을 완료한 후, 최종 검토 과정을 통해 토픽 이름과 분류를 더 명확하게 수정하거나, 모호한 토픽은 삭제하고 해당 기사들을 다시 분류하였다.

3.2 토픽모델링

3.2.1 LDA

LDA는 문서 내에 숨어있는 주제를 찾아내기 위해 문서와 키워드를 결합하는 비지도학습 기법이다(Blei et al., 2003). LDA는 두 가지 가정에 기반을 두고 있다. 첫 번째 가정은 토픽이 단어의 순서를 고려하지 않고, 단어의 출현 빈도에만 초점을 둔 Bag of Words(BoW) 모델로 표현된다는 것이다. 이에 따라 토픽은 단어의 확률 분포로 표현된다. 두 번째 가정은 하나의 문서에는 여러 토픽이 존재하며, 문서는 토픽의 확률 분포로 표현된다는 것이다. 이 두 확률 분포를 결합하여 문서 내에 숨어있는 주제를 찾아내는 기법이 LDA이다. Fig. 2는 LDA의 간략한 과정을 보여 준다.

Table 1. Number 1 Article of the Collected BBC News

Title	Suffolk: A14 reopens after seven hour closure
Date	2022-10-01
Contents	The A14 in Suffolk has reopened after two crashes closed the road for seven hours. Highways England said police attended a collision at about 09:00 BST between junction 50, for Stowmarket, and junction 51, ...

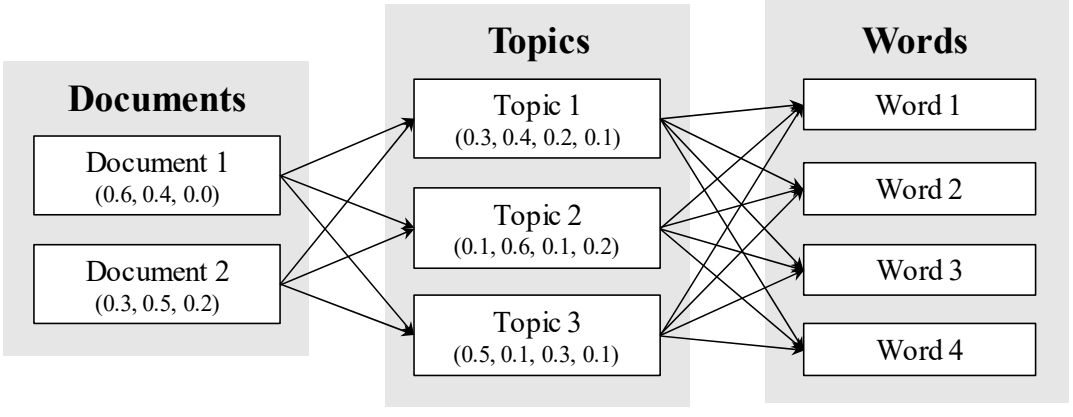

Fig. 2. LDA Process

LDA 토픽모델링은 전처리, 토픽 개수 산정, 모델링의 순서로 수행된다. 전처리 단계에서는 우선 전체 텍스트를 소문자로 변환하고, 띄어쓰기 단위로 토큰화를 수행하였다. 이후 자연어 처리를 위한 파이썬 패키지인 Natural Language Toolkit(NLTK)에 구축되어 있는 영어 불용어 사전을 사용해 불용어를 제거하고, 뉴스 기사 특성상 끝부분에 "Follow BBC News"와 같은 광고성 문구가 반복적으로 등장하는 것을 확인하여 해당 문구를 제거하였다.

그 후 전처리가 완료된 텍스트를 활용하여 최적의 토픽 개수를 결정하기 위해 coherence score를 계산하였다. Coherence score는 토픽모델링 결과로 생성된 토픽 내의 키워드들이 얼마나 의미론적으로 유사한지 측정하는 지표로(Newman et al., 2010), 아래의 Eq. (1)로 계산할 수 있다. Eq. (1)에서 $C(w_i,w_j)$는 두 단어 w_i와 w_j 사이의 유사도를 나타내는 coherence score이다. $D(w_i)$는 단어 w_i를 포함하는 문서의 수를 나타내고, $D(w_i,w_j)$는 단어 w_i와 w_j를 포함하는 문서의 수를 나타낸다(Stevens et al., 2012).

$$C(w_i,w_j) = \log\frac{D(w_i,w_j)+1}{D(w_i) \cdot D(w_j)} \quad (1)$$

3.2.2 BERTopic

BERTopic은 문서 임베딩, 차원 축소, 클러스터링 및 키워드 추출의 과정을 통해 토픽을 분류한다(Grootendorst, 2022). 먼저, 사전 훈련된 언어 모델인 Bidirectional Encoder Representations from Transformer(BERT)를 사용하여 개별 뉴스 기사를 벡터로 표현하는 임베딩 작업을 수행한다. 그 후, Uniform Manifold Approximation and Projection(UMAP)이라는 차원 축소 기법을 사용하여 임베딩된 벡터의 차원을 축소한다. 그리고 밀도 기반 클러스터링 알고리즘인 Hierarchical Density-Based Spatial Clustering of Applications with Noise(HDBSCAN)를 사용하여 유사한 주제를 가진 문서끼리 군집을 생성한다. 이때, 어떤 군집에도 속하지 않는 문서는 Outlier로 처리된다. HDBSCAN 알고리즘 적용 시, 파라미터를 조정함으로써 토픽 개수를 특정 개수로 설정하거나, 알고리즘에 의해 최적의 토픽 개수를 자동으로 찾아낼 수 있다. 본 연구에서는 BERTopic 알고리즘이 최적의 성능을 보이는 토픽 개수를 자동으로 찾도록 설정하였다. 생성된 문서 군집에 class-based Term Frequency-Inverse Document Frequency (c-TF-IDF)를 적용하여 토픽별 키워드를 추출하였다. BERTopic에 적용되는 c-TF-IDF는 각각의 군집을 하나의 문서로 간주하여, 군집별 주요 단어의 TF-IDF 값을 계산함으로써 각 토픽의 주제를 설명할 수 있는 주요 키워드를 도출하였다(Fig. 3).

3.3 토픽 이름 배정 및 성능평가

LDA와 BERTopic 토픽모델링 기법은 토픽 리스트와 각 토픽의 키워드를 출력한다. 연구진은 이를 기반으로 각 토픽의 이름을 할당하였다. 예를 들어, 키워드가 "Rescue", "Service", "Blaze", "Firefighters"와 같이 초기에 배정된 토픽의 경우, "Fire"라는 이름으로 토픽을 다시 할당하였다. LDA 모델의 결과에서는 하나의 문서에 여러 개의 토픽이 분포하므로, ground truth와의 일대일 비교를 위해 각 뉴스 기사에 대해 가장 높은 확률을 가지는 토픽을 배정하였다. BERTopic의 경우, outlier에는 이름을 할당하지 않고 성능평가 시 모두 오답으로 분류하였다.

LDA와 BERTopic 모델링 결과에 배정한 토픽과 ground truth의 소분류를 일대일로 비교하여 Accuracy, Precision, Recall, F1 score를 계산하였다(Table 2). Fig. 4는 100개의 문서에 대해 토픽 개수가 4개일 때의 multi-class confusion matrix를 생성한 예시이다. 이 경우, 전체 토픽에 대한 Accuracy 및 토픽 a에 대한

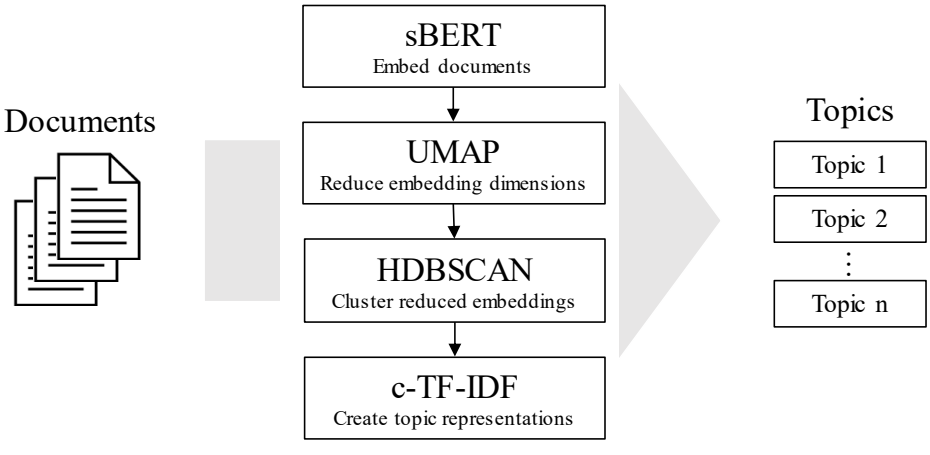

Fig. 3. BERTopic Process

Table 2. Performance Evaluation Metrics

Metric	Definition
Accuracy	Ratio of correct predictions to the total predictions
Precision	Ratio of true positives to the total predicted positives
Recall	Ratio of true positives to the total actual positives
F1 score	Harmonic mean of precision and recall

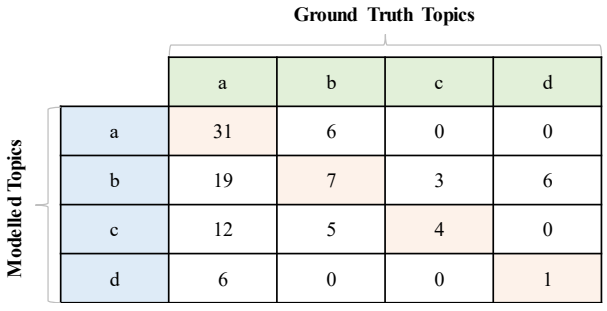

Fig. 4. Example of Multi-Class Confusion Matrix

Precision, Recall, F1 score는 Eqs. (2)~(5)와 같이 계산된다. 본 연구에서는 기사의 개수가 토픽별로 불균형하게 분포하므로, 불균형한 클래스 분포에서도 모델의 성능을 효과적으로 평가할 수 있는 F1 score를 성능평가 지표로 사용하였다.

$$Accuracy_{total} = \frac{43}{100} = 0.430 \tag{2}$$

$$Precision_{topic(a)} = \frac{31}{37} = 0.838 \tag{3}$$

$$Recall_{topic(a)} = \frac{31}{68} = 0.456 \tag{4}$$

$$F1\,score_{topic(a)} = \frac{2 \times 0.838 \times 0.456}{0.838 + 0.456} = 0.591 \tag{5}$$

4. 결과 및 논의

4.1 뉴스 기사 수집 및 ground truth 생성

본 연구는 BBC 뉴스 웹사이트에서 총 1,305,517건의 뉴스 기사를 수집하였으며, 이 중 2022년 10월에 발간된 뉴스 기사 총 6,273건을 활용하여 ground truth를 생성하였다. 토픽 리스트는 총 13개의 대분류, 128개의 소분류로 구성되었다(Table 3).

4.2 LDA 모델링

4.2.1 토픽 이름 배정 결과

LDA 모델링 결과, 토픽 개수가 10개일 때 가장 높은 coherence

Table 3. Ground Truth of BBC News

Category	Sub-category	No. of News
Politics	Defense, Congress/Parties, Bill, Election, Royal household, Diplomacy, Global politics	210
Economy	Finance, Company, Real estate, Debt, Living economy, Tax, Energy policy	395
Society	Education, Traffic accident, Rescue, Labor, Homeless, Road/Traffic, Drugs, Immigration, Water supply, Crime, Poverty, Incident, Accident, Death case, LGBT, Protest, Missing case, Kid/Teenager, Remains, Medical treatment, Human right, Racism, Charity, Disability, Trial, Gender, Religion, Hate crime, Region, Development, Opening, Closing, Birth/Abortion, Terror, Strike, Fire, Environment	3,418
Industry	IT/Science, Game, Tourism, Agriculture, Service, Energy/Resource, Space, Automobile, Ship-building, Railway, Fashion, Airline, Marine	256
Art/Culture	Concert, Competition, Museum, Broadcast, Movie, Music, Piece of work, Exhibition, Sculpture, Festival, Celebrity	377
Sports	Marathon, Sports	100
Weather	Weather	14
Natural Disaster	Natural Disaster, Flood	37
Science	Science, Fauna and flora	112
Specific Issues	Bird flu, Burkina Faso coup, COP27, Cough syrup, COVID-19, Creeslough explosion, Diwali 2022, Donald Trump, Elon Musk, Eurovision, Food bank, Kanye West, Liz Truss, Lucy Letby, Nicola Sturgeon, Northern Ireland Protocol, Pelosi, Rishi Sunak, Suella Braverman, Nobel prize, London Marathon, Ethiopia War, Ukraine War, Protest in Iran, Disaster in Itaewon, Qatar World Cup, Halloween	922
World	China	31
Information	Health, Auction, History, Stamp, People, Book, Memorial, Profile	248
Miscellaneous	Graffiti, Advertisements, Lottery, Photo, Survey, Awards, Eclipse, Quiz, Miscellaneous	153

Table 4. LDA Keywords & Topic Assignment-Topic Num 10

Topic	Keywords	Topic name	Ratio(%)
1	Police, Road, Officers, Arrested	Crime	58.1
2	Sunak, Government, Minister, Braverman	Congress/Parties	26.1
3	Protocol, Ireland, Dup, Northern	Northern Ireland Protocol	2.7
4	Pelosi, Trump, Iran, Pup	Pelosi	2.4
5	Britishvolt, Graffiti, Xi, Owl	Graffiti	2.1

score를 보였으며, 상위 5개 토픽에 대한 키워드 및 토픽 이름 배정 결과는 Table 4와 같이 나타났다. 토픽 개수의 다양성을 고려하여 20개의 토픽 개수로 설정한 모델에 대해 추가적으로 분석을 수행하였다.

두 경우 모두 소수의 토픽이 대부분의 비중을 차지하는 것을 확인할 수 있었다. 토픽 개수가 10개일 때 1번 토픽이 58.1%, 2번 토픽이 26.1%를 차지했고, 나머지 8개의 토픽이 15.8%의 비율로 존재하였다. 토픽 개수가 20개인 경우에도 1, 2, 3번 토픽이 80.7%로 높은 부분을 차지하였다. 토픽 개수를 20개 이상으로 설정한 경우, 나머지 토픽이 더욱 세분화되어 키워드만 보고 분류하기 어려운 경우가 늘어났기에 토픽 개수를 20개로 한정하였다.

개별 토픽 검토 결과, 하나의 토픽에 여러 주제의 키워드가 포함된 것을 확인할 수 있었다. 토픽 개수가 10개일 때, 1번 토픽 “범죄”는 “Police”, “Road”, “Officers”, “Arrested”, “Hospital”, “Council”, “Old”, “Injuries”와 같은 키워드를 포함하였다. 이러한 키워드들은 “Crime”, “Road/Traffic”, “Traffic accident” 주제의 내용을 모두 포함하고 있어, 하나의 토픽으로 한정한 것이 추후 성능평가에서 정확도가 낮아지는 원인이 되었다.

4.2.2 성능평가 결과

성능평가 결과는 Table 5와 같이 나타났다. 토픽이 10개일 때의 F1 score은 0.011, 20개일 때의 F1 score은 0.0003으로 두 경우 모두 매우 낮은 점수를 보였고, 20개로 토픽을 늘렸을 때 성능이 오히려 감소하였다. 낮은 F1 score가 얻어진 주요한 원인은 성능평가를 위해 한 문서당 하나의 토픽만을 할당하였기 때문인 것으로 파악된다. LDA 토픽 모델에는 하나의 문서에 여러 토픽이 확률적으로 분포하지만, 본 연구에서는 각 기사에 하나의 토픽만을 할당하여 성능을 평가하였기 때문에 낮은 F1 score를 얻은 것으로 보인다.

Table 5. LDA Performance Evaluation

No. of Topics	Accuracy	Precision	Recall	F1 score
10	0.165	0.013	0.022	0.011
20	0.0104	0.0002	0.0078	0.0003

또한, LDA 모델에서 설정한 총 토픽 개수는 10개와 20개이지만, 일대일 비교를 시행한 ground truth의 소분류 개수는 128개로 매우 큰 차이가 있었으며, 이러한 차이가 LDA 모델의 낮은 성능에 영향을 미친 것으로 추정된다.

4.3 BERTopic 모델링

4.3.1 토픽 이름 배정 결과

BERTopic의 경우, 토픽 개수를 자동으로 설정하고 모델링을 수행한 결과 총 87개의 토픽이 생성되었다. Table 6은 87개 토픽 중 상위 10개 토픽의 키워드와, 그에 대한 토픽 이름을 배정한 결과이다. Topic -1은 outlier로 분류된 기사들이다.

BERTopic은 LDA와 비교했을 때 토픽별 키워드가 명확하게 나타나, 토픽 이름을 할당하는 것이 더 쉽다는 이점이 있었다. 예를 들어, 1번 토픽에서 생성된 “truss”, “tax”, “ms truss”, “prime”과 같은 키워드는 영국의 전 총리인 “Liz Truss”와 관련된 단어들로 이루어져 있음을 알 수 있었다.

4.3.2 성능평가 결과

토픽 개수가 87개일 때의 성능평가 결과, 정확도는 0.394, Precision은 0.242, Recall은 0.288, F1 score은 0.244로 나타났다. 비교적 낮은 F1 score를 얻은 이유 중 하나는 outlier를 모두 오답으로 분류하였기 때문이다. Outlier는 총 6,273개의 기사 중 상당한 부분(1,785개, 28.5%)을 차지하므로 모델의 성능에 영향을 주었을 것으로 생각된다. 또한, BERTopic이 자동으로 도출한 토픽 개수(87개)가 수작업으로 분류한 토픽 개수(128개)와 차이가 나는 점도 낮은 F1 score에 영향을 미친 것으로 보인다.

구체적으로, 가장 높은 F1 score인 0.857을 얻은 토픽은 “Cough Syrup”으로, 인도네시아에서 100명 가까이 되는 아이들이 감기약을 먹고 사망한 사건을 다루는 토픽이다. 그 외 F1 score가 높게 나온 토픽으로는 인도의 힌두교 축제인 “Diwali 2022”(0.842), “Ukraine War”(0.833), 영국 Creeslough 지역의 주유소 폭발 사고를 다루는 “Creeslough explosion”(0.826) 등이 있다(Table 7). 이 토픽들은 모두 대분류 중 “Specific Issues”에 속하는 것으로 나타났다. 즉, BERTopic 모델은 특정 이슈를 다루는 뉴스 기사를

Table 6. BERTopic Topic Assignment

Topic	Keywords	Topic name	Ratio(%)
-1	Said, People, Mr, BBC	Outlier	28.5
0	Police, Man, Police said, Officers	Crime	12.1
1	Truss, Tax, Ms Truss, Prime	Liz Truss	3.3
2	Rescue, Service, Blaze, Firefighters	Fire	3.2
3	Russian, Ukraine, Russia, Ukrainian	Ukraine War	3.1
4	Council, Homes, Plans, Sites	Development	2.1
5	Radio, Tv, Content, Media	Broadcast	2.0
6	Patients, Hospital, Nhs, Care	Medical treatment	1.9
7	Northern, Ireland, Northern Ireland, Protocol	Northern Ireland Protocol	1.9
8	Road, Bridge, Traffic, Highways	Road/Traffic	1.7

정확하게 분류한다는 것을 알 수 있다.

F1 score가 0이 나온 토픽으로는 “Congress/Parties”와 “Tax”가 있다. “Congress/Parties” 토픽의 경우 총 23개 기사 중 5개가 “Liz Truss”로, 8개가 “Rishi Sunak”로 분류되었다. Liz Truss는 영국 78대 총리로, 금융 위기를 촉발한 경제 정책 실책으로 2022년 10월 25일 사임하였다. Rishi Sunak은 Liz Truss의 뒤를 이은 영국 79대 총리이다. 이 두 정치인에 관한 내용은 실제 뉴스 기사에는 없지만, 일반적인 정치 기사에서 나타나는 단어와 문맥이 유사하므로 BERTopic 모델에서 “Congress/Parties”로 분류되었을 것으로 추정된다. “Tax” 토픽 역시 22개 중 17개의 기사가 “Liz Truss”로 오분류되었다(Table 8).

4.3.3 비교 및 논의

BERTopic 모델링 시 87개의 토픽에 대해 0.244의 F1 score를 얻었다. 이는 LDA의 0.011보다 훨씬 높은 수치로, BERTopic 기법이 LDA보다 뉴스 기사의 주제를 실제와 더 유사하게 분류한다는 것을 확인할 수 있다. 두 경우 모두 일반적인 예측 모델에 비해 매우 낮은 F1 score를 보였으나, 뉴스 기사의 실제 주제를

Table 7. News Articles with High F1 Score

Topic	Title	Text
Cough syrup	Indonesia bans all syrup medicines after death of 99 children	The deaths of nearly 100 children in Indonesia have prompted the country to suspend sales…
Diwali 2022	Diwali celebrations planned at Leicester waterways	Activities to celebrate Diwali are being organised around Leicester's waterways…
Ukraine War	Ukraine war: Tortured for refusing to teach in Russian	Ukrainian forces say they have taken back 6,000 sq km of territory, liberating communities...
Creeslough explosion	Creeslough spirit shines amid darkness of tragedy	The names of those killed in the Creeslough petrol station explosion were made known…

Table 8. News Articles with Low F1 Score

Topic	Title	Text
Congress/ Parties	Young Tory chairman apologises for 'Birmingham is a dump' tweet	The chairman of a group of young Conservative Party members has apologised...
	Graham Brady: The man who sees off Tory prime ministers	Born in Salford in 1967, Sir Graham first became active in the Conservative Party...
Tax	Mini-budget: Keeping 45p tax rate in Wales 'would raise £45m'	Keeping the 45p rate of income tax in Wales would raise about £45m, according to Wales' finance minister...
	Community helpdesk launched to support Jersey taxpayers	A community helpdesk is being launched to support taxpayers in Jersey. Revenue Jersey...

고려한 성능평가를 시행했다는 점에서 의의가 있다. 이전 연구 중에는 ground truth와 모델링 결과로 얻은 기사별 토픽을 직접 비교한 사례가 없기 때문에, 본 연구에서 수행한 ground truth 기반의 성능평가는 토픽모델링 기법의 성능에 대한 새로운 방식을 제안할 수 있다.

또한, BERTopic에서 F1 score를 높게 기록한 토픽의 기사를 살펴본 결과, 특정 이슈를 다루는 기사가 정확하게 분류되었다. 이러한 특징은 BERTopic 모델을 해외건설사업 시 현지 상황을 파악하는 데 유용하게 쓸 수 있다는 것을 보여 준다. 특정 이슈에 해당하는 기사를 정확하게 분류함으로써 해당 지역의 이슈 및 동향을 신속하게 파악하고, 사업의 주요 의사 결정에 도움을 줄 수 있다.

5. 결 론

본 연구는 2022년 10월 한 달간의 BBC 뉴스 기사를 대상으로 실제 주제에 대한 ground truth를 생성하였다. 이를 기반으로 LDA, BERTopic 두 토픽모델링 기법으로 얻어진 토픽에 ground truth와 동일한 이름을 할당하였다. 그 후 기사 단위로 일대일 비교를 통해 성능평가를 시행한 결과, BERTopic 기법의 성능이 더 우수함을 확인하였다.

본 연구에서 제안한 성능평가 방법론을 활용하면, 해외건설시장의 현지 상황을 파악할 때 뉴스 기사를 요약하는 데에 최적의 토픽모델링 기법을 찾을 수 있다. LDA나 BERTopic 외의 다른 토픽모델링 기법에도 이 방법론을 적용하여 가장 성능이 좋은 기법을 찾아 활용한다면, 현지 상황을 효율적으로 파악할 수 있을 것이다.

연구의 한계는 토픽모델링 결과로 나타난 키워드를 보고 연구자가 직접 토픽 이름을 설정해야 한다는 점이다. 이를 고려하여 향후에는 토픽별 키워드를 입력하면 자동으로 토픽 명칭을 할당하는 알고리즘을 적용하여, 일정한 기준에 따라 토픽을 배정하는 방식으로 연구를 수행할 계획이다.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00241758). This research was conducted with the support of the “National R&D Project for Smart Construction Technology (No.23SMIP-A158708-04)” funded by the Korea Agency for Infrastructure Technology Advancement under the Ministry of Land, Infrastructure and Transport.

ORCID

Sehwan Chung https://orcid.org/0000-0002-9256-5548
Seokho Chi https://orcid.org/0000-0002-0409-5268

References

Abuzayed, A. and Al-Khalifa, H. (2021). “BERT for Arabic topic modeling: An experimental study on BERTopic technique.” *Procedia Computer Science*, Elsevier, Vol. 189, pp. 191-194, https://doi.org/10.1016/j.procs.2021.05.096.

Blei, D. M. (2012). “Probabilistic topic models.” *Communications of the ACM*, ACM, Vol. 55, No. 4, pp. 77-84, https://doi.org/10.1145/2133806.2133826.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). “Latent dirichlet allocation.” *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022.

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M. and Fdez-Riverola, F. (2014). “Web scraping technologies in an API world.” *Briefings in Bioinformatics*, Oxford Unversity, Vol. 15, No. 5, pp. 788-797, https://doi.org/10.1093/bib/bbt026.

Goldszmidt, R. G. B., Brito, L. A. L. and de Vasconcelos, F. C. (2011). “Country effect on firm performance: A multilevel approach.” *Journal of Business Research*, Elsevier, Vol. 64, No. 3, pp. 273-279, https://doi.org/10.1016/j.jbusres.2009.11.012.

Grootendorst, M. (2022). “BERTopic: Neural topic modeling with a class-based TF-IDF procedure.” *arXiv Preprint*, arXiv: 2203.05794 [cs.CL], https://doi.org/10.48550/arXiv.2203.05794.

Jallan, Y., Brogan, E., Ashuri, B. and Clevenger, C. M. (2019). “Application of natural language processing and text mining to identify patterns in construction-defect litigation cases.” *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, ASCE, Vol. 11, No. 4, 04519024, https://doi.org/10.1061/(ASCE)LA.1943-4170.0000308.

Javernick-Will, A. N. and Scott, W. R. (2010). “Who needs to know what? Institutional knowledge and global projects.” *Journal of Construction Engineering and Management*, ASCE, Vol. 136, No. 5, pp. 546-557, https://doi.org/10.1061/(ASCE)CO.1943-7862.0000035.

Jiang, H. C., Qiang, M. S. and Lin, P. (2016). “Finding academic concerns of the Three Gorges Project based on a topic modeling approach.” *Ecological Indicators*, Elsevier, Vol. 60, pp. 693-701, https://doi.org/10.1016/j.ecolind.2015.08.007.

Jung, N. and Lee, G. (2019). “Automated classification of building information modeling (BIM) case studies by BIM use based on natural language processing (NLP) and unsupervised learning.” *Advanced Engineering Informatics*, Elsevier, Vol. 41, 100917, https://doi.org/10.1016/j.aei.2019.04.007.

Moon, S., Chung, S. and Chi, S. (2018). “Topic modeling of news article about international construction market using

latent dirichlet allocation." *KSCE Journal of Civil and Environmental Engineering Research*, KSCE, Vol. 38, No. 4, pp. 595-599, https://doi.org/10.12652/Ksce.2018.38.4.0595 (in Korean).

Newman, D., Lau, J. H., Grieser, K. and Baldwin, T. (2010). "Automatic evaluation of topic coherence." *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for ACL*, ACL, Los Angeles, USA, pp. 100-108.

Stevens, K., Kegelmeyer, P., Andrzejewski, D. and Buttler, D. (2012). "Exploring topic coherence over many models and many topics." *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning,* ACL, Jeju Island, Korea, pp. 952-961.

Wallach, H. M., Murray, I., Salakhutdinov, R. and Mimno, D. (2009). "Evaluation methods for topic models." *Proceedings of the 26th Annual International Conference on Machine Learning, Association for Computing Machinery*, ACM, New York, USA, pp. 1105-1112, https://doi.org/10.1145/1553374.1553515.

Wei, X. and Croft, W. B. (2006). "LDA-based document models for ad-hoc retrieval." *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Association for Computing Machinery, New York, USA, pp. 178-185, https://doi.org/10.1145/1148170.1148204.