Journal of KIIT. Vol. 22, No. 1, pp. 203-213, Jan. 31, 2024. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2024.22.1.203

사건 변화와 주체 감성 추이 분석을 위한 KPF-BERT 기반 뉴스 동향 시각화 시스템 개발

손효원*, 한영민**, 남경현***, 한수빈****, 유길상*****

Development of a News Trend Visualization System based on KPF-BERT for Event Changes and Entity Sentiment Analysis

Hyowon Son*, Youngmin Han**, Kyounghyun Nam***, Subin Han****, and Gilsang Yoo*****

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(RS-2023-00246191)

요 약

최근 한국에서는 뉴스 복잡성으로 인해 뉴스에 대한 무관심과 이해도의 부족이 46개국 평균보다 더 높게 나타나고 있다. 현대 뉴스는 다양한 관점과 복잡한 사회적 문맥을 포함하고 있어서 독자들은 이러한 복잡성 내에서 사건과 주체 간의 흐름을 이해하는 데 어려움을 겪고 있다. 본 연구에서는 이러한 문제점을 해결하기 위해 뉴스 동향을 마인드맵 형태로 시각화하고 주체 간의 관계를 나타내는 시각화 시스템을 제안하였다. 제안한 시스템은 다양한 기사를 KPF-BERT 기반 개체명 인식 기법으로 주체를 추출하고, 클러스터별 주요 주체의 감성 분석을 통해 사건의 흐름을 시각화로 나타나도록 하였다. 시뮬레이션 결과, 제안한 시스템을 통한 뉴스의 접근성 향상과 미디어 리터러시의 증진을 기대할 수 있다.

Abstract

In recent years, South Korea has witnessed a heightened disinterest and lack of understanding towards news due to the proliferation of various news media, with its rates surpassing the average of 46 countries. Contemporary news encompasses diverse perspectives and intricate societal contexts, making it challenging for readers to comprehend the flow and relationships between events and their main actors. This study proposes a visualization system that represents news trends in a mind-map format, highlighting relationships between these actors. The system extracts entities from multiple articles by named entity recognition based on KPF-BERT and visualizes the progression of events through sentiment analysis of principal entities in each cluster. Simulation results suggest that the proposed system can enhance accessibility to news and promote media literacy.

Keywords

news trend visualization, clustering, named entity recognition, targeted sentiment analysis, KPF-BERT

* 한국외국어대학교 ELLT학과 학사과정
- ORCID: https://orcid.org/0009-0003-8584-8325
** 홍익대학교 건설도시공학부 도시공학과 학사과정
- ORCID: https://orcid.org/0000-0003-3959-120X
*** 동국대학교 산업시스템공학과 학사과정
- ORCID: https://orcid.org/0000-0002-4300-1375
**** 고려대학교 정보대학 컴퓨터학과 학사과정
- ORCID: https://orcid.org/0009-0007-4681-9301
***** 고려대학교 정보대학 정보창의교육연구소 교수(교신저자)
- ORCID: https://orcid.org/0009-0002-1085-5355

· Received: Oct. 18, 2023, Revised: Nov. 10, 2023, Accepted: Nov. 13, 2023
· Corresponding Author: Gilsang Yoo
Creative Informatics and Computing Institute, Korea University,
145 Anam-ro, Seongbuk-gu, Seoul, Korea
Tel.: +82-2-3290-1674, Email: ksyoo@korea.ac.kr

Ⅰ. 서 론

최근 젊은 층(35세 미만 연령층)에서 이른바 '뉴스 회피' 현상이 두드러지게 나타나는 현상이 전 세계적으로 문제가 되고 있다. 이러한 경향은 한국에서도 두드러지게 나타나는데, 2022년 대한민국의 디지털 뉴스 보고서에 따르면 연령별 뉴스 무관심층 비율을 조사한 결과, 젊은 층의 뉴스 무관심층 비율이 21%로 높은 것으로 나타났다[1]. 2030세대는 주로 짧은 텍스트와 이미지를 통해 정보를 소비하기 때문에, 긴 호흡의 글에 관한 관심이 줄어든 것으로 해석할 수 있으며, 이러한 선호도 변화와 함께 정치 뉴스를 완전히 이해하려면 축적된 배경지식이 필요하다는 점이 정치 뉴스 회피 현상의 주요 요인으로 분석되었다.

또한, 뉴스 회피 이유로 한국에서 뉴스를 이해하기 어렵다고 답한 한국 독자의 비율이 46개국 평균보다 높게 나타났다. 현대의 뉴스는 복잡한 사회적 상황과 다양한 관점을 다루기 때문에 독자들은 이러한 복잡성 내에서 사건과 주체 간의 흐름을 이해하는 데 어려움을 겪고 있다.

대표적인 뉴스 기반 시각화 서비스에는 한국언론진흥재단의 BIGKinds가 있다[2]. BIGKinds는 1990년부터 수집한 뉴스 빅데이터를 기반으로 형태소와 개체명을 분석하여 네트워크 형태로 시각화 서비스를 제공하고 있다. 그림 1에서와 같이, 이 서비스는 주요 인물, 기관 및 장소에 대한 키워드 클러스터를 워드클라우드 형태로 단순히 시각화하거나, 그림 2와 같이 상위 뉴스 검색 결과에서 추출한 개체명 사이를 연결하여 단순 네트워크 형태로 표현한 것 시각화 기능을 가지고 있다.

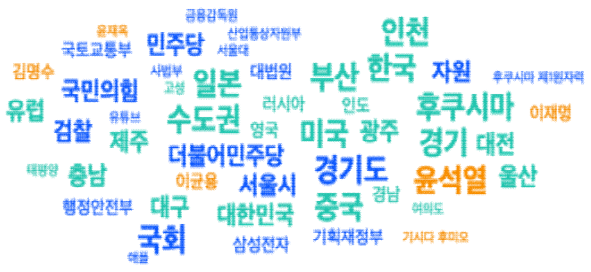

그림 1. 키워드 클러스터
Fig. 1. Keyword clusters

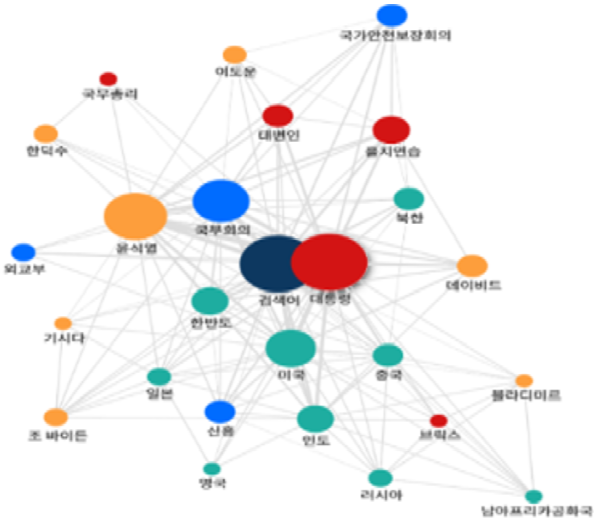

그림 2. 개체명 관계망
Fig. 2. Entity network

따라서, 본 논문은 뉴스의 흐름을 볼 수 없는 기존 서비스를 개선하기 위하여, KPF-BERT 기반 개체명 인식 기법으로 주체를 추출하고, 클러스터별 주요 주체의 감성 분석을 통해 사건의 흐름을 직관적으로 살펴볼 수 있는 시각화기법을 제안하였다. 이를 위해 매일 수집한 정치 기사에서 상위 5개의 클러스터를 정하여 각 클러스터에 대해 주요 주체를 개체명 인식으로 추출한다. 이후 주체가 가지는 극성(polarity)을 감성 분석을 통해 해석하고, 여러 시점의 클러스터들 간의 관계를 분석하여 독자들이 시간에 따른 사건의 변화를 확인할 수 있도록 하였다. 최종적으로 웹 기반 시각화를 통해 결과들을 종합하여 다양한 정치 이슈를 직관적으로 이해하고 분석할 수 있는 시스템을 구현하였다. 최종적으로, 제안한 시스템은 사건의 흐름과 주체 간의 관계를 웹 기반 시각화를 통해 직관적으로 제시함으로써 사용자들에게 정치 이슈의 다양한 측면을 이해하고 분석할 수 있는 도구로 활용될 수 있도록 하였다.

본 논문의 구성을 다음과 같다. 2장은 관련연구로서 자연어처리를 위한 기존 언어모델과 대표적 문서 클러스터링 및 키워드 추출 기법에 관해 기술하였다. 또한 개체명 인식과 감성 분석을 위한 딥러닝 기법을 기술하였다. 3장에서는 본 논문에서 제안하는 사건 변화와 주체 감성 추이 분석을 위한 KPF-BERT 기반 뉴스 동향 시각화 시스템에 관해 기술하였다.

4장에서는 제안한 모델을 기반으로 작동하는 시각화 시스템들을 웹 기반에서 구현하고 이를 기반으로 사용자 경험 평가에 대해 논하였다. 마지막으로 5장에서는 연구결과에 대한 기대효과와 향후 연구에 관하여 기술하였다.

II. 관련 연구

2.1 언어 모델

자연어 처리(NLP) 분야에서는 다양한 모델이 목적에 따라 개발되어 왔다. RNN은 시퀀스 데이터 처리에 적합하지만, 긴 시퀀스에 대해서는 그래디언트 문제가 발생한다. 이를 해결하기 위해 Attention 메커니즘과 Transformer 모델이 도입되었다. Transformer는 RNN의 순환적 특성을 배제하고, Attention을 통해 텍스트의 각 부분을 독립적으로 처리한다[3]. BERT는 Transformer를 기반으로 한 양방향 인코딩 모델로, 대규모 데이터셋으로 사전 학습되어 문맥을 파악하는 데 강점을 보인다[4]. SBERT는 BERT를 문장 수준 임베딩으로 확장하여, 문장 간 유사도 측정에 효율적이다[5].

KPF-BERT는 한국형 표준 뉴스 기사 인공지능 언어 모델이다[6]. BERT를 기반으로 한국언론진흥재단이 보유한 BigKinds 기사 데이터를 활용해 사전학습에 사용되었으며 빅카인즈 기사 약 4,000만 건을 학습해 언론사 및 뉴스 기사 활용 기술에 적합한 모델로 개발되었다. KPF-SBERT는 SBERT를 BigKinds의 언론사 기사 데이터를 학습시켜 빠르고 효율적으로 의미 비교 등의 작업에 활용되고 있다.

2.2 문서 클러스터링 및 키워드 추출

문서 클러스터링과 키워드 추출은 자연어 처리 기술 분야의 중요한 부분이다. 문서 클러스터링에는 크게 키워드 기반과 문맥 기반 방식이 있다. 키워드 기반 방식에는 BoW(Bag of Words)[7] 기반 TF-IDF[8] 혹은 KeyBERT[9]가 있다. 단어의 빈도로만 문서를 표현하는 BoW를 보완하기 위해 TF-IDF를 결합하였다. 이후 K-means를 적용하여 문서별 키워드들을 클러스터링한다. KeyBERT는 BoW와 TF-IDF 이후 자연어처리 분야의 발전에 힘입어 발표된 BERT 모델을 기반으로 한다. BERT는 단어의 빈도를 세는 것에서 더 나아가 문맥 기반으로 단어의 의미를 학습하고 문서의 의미를 파악하는 데 탁월한 성능을 보인다. KeyBERT는 BERT 모델을 이용해 문서 레벨에서의 주제를 파악하고 N-gram 단어 및 구절에 대해 임베딩을 추출한다. 이후 문서와 가장 유사한 단어 및 구절을 찾기 위해 코사인 유사도를 사용한다. 가장 유사한 단어는 전체 문서를 가장 잘 설명하는 단어로 식별할 수 있다고 가정한다.

문맥 기반 방식은 LDA(Latent Dirichlet Allocation)[10]와 BERTopic[11]이 있다. LDA는 문서의 단어를 토픽에 할당하고, 각 토픽의 단어 분포를 계산하여 문서의 토픽 분포를 계산하는 방법이다. 반면, BERTopic은 BERT로 문서를 임베딩한 것을 UMAP(Uniform Manifold Approximation and Projection)[12]으로 차원을 축소하여 HDBSCAN(Hierarchical Density-Based Spatial Clustering Applications with Noise)을 이용해 클러스터링을 하는 방식이다. 벡터화된 문서들에 대해 유사한 문서들끼리 묶어준다. HDBSCAN은 K-means와는 달리 중심 기반 클러스터링(Center-based) 방식이 아닌 밀도 기반(Density-based) 클러스터링 방식이다[13]. 밀도 기반 군집화는 데이터 포인트의 밀도를 기준으로 군집을 생성하여 불특정한 형태로 분포하는 데이터도 군집화가 가능하다[14].

키워드 추출은 MMR(Maximal Marginal Relevance) 알고리즘이 대표적이다[15]. MMR 알고리즘은 문서와 가장 유사한 키워드를 먼저 선택한 후, 다음 키워드를 선택할 때에는 문서와는 유사하지만 이미 선택한 키워드와는 다른 키워드를 선택하는 기법이다. 따라서 중요성과 다양성을 균형 있게 고려하여 키워드를 추출할 수 있다.

2.3 개체명 인식(Named entity recognition)

개체명 인식은 미리 정의해둔 사람, 회사, 장소, 시간, 단위 등에 해당하는 단어(개체명)를 문서에서 인식하여 추출 분류하는 기법으로, 추출된 개체명은 인명(Person), 지명(Location), 기관명(Organization), 시간(Time) 등으로 분류된다. 개체명 인식은 정보 추출을 목적으로 시작되어 자연어처리, 정보 검색 등에 사용된다.

최근에는 딥러닝 모델 기반 개체명 인식이 주로 사용되고 있다. 예를 들어, KPF-BERT 모델을 기반으로 개체명 인식을 목적으로 개발된 KPF-BERT-NER 모델이 있다[16]. 이 모델은 언론 기사를 학습하여 150개 클래스를 분류한다. 20여 개의 BIO 태그셋으로 개체명 인식을 수행하는 다른 모델과 비교하여 BIO 태그셋의 개수가 매우 많은 점은 감성 분석에 유리한 장점을 가지고 있다[17].

2.4 감성 분석

감성 분석에는 사전 기반(Lexicon-based) 통계적 감성 분석과 지도(Supervised) 학습, 반지도학습(Semi-supervised)과 같이 딥러닝을 이용한 방식이 있다[18]. 사전 기반 감성 분석은 다양한 표현을 반영하기 어렵지만 별도의 훈련이나 라벨링 없이 수행될 수 있음으로 뉴스와 같이 보수적으로 표현을 사용하는 미디어에 적합하다.

$$pmi(x,y) = \log\frac{p(x,y)}{p(x)p(y)} \\ = \log\frac{p(xverty)}{p(x)} = \log\frac{p(yvertx)}{p(y)} \quad (1)$$

식 (1)의 PMI(Pointwise Mutual Information)는 사전 기반 감성 분석에서 자주 쓰이는 통계적 방법론이다[19]. PMI는 두 단어가 함께 나타날 확률을 개별 단어들의 확률의 곱으로 나눈 값으로, 이 값은 두 단어가 함께 자주 나타날수록 커진다. PMI를 통해 두 단어의 연관성을 측정함으로써 단어 간의 의미적 관계를 파악할 수 있다. PMI 값이 음수일 경우에는 이 값을 변형하여 활용하는데, 이를 PPMI (Positive PMI)라고 한다. 식 (2)의 PPMI는 특히 감성 분석이나 문맥 분석과 관련된 작업에서 유용하게 활용될 수 있다.

$$ppmi(x;y) \equiv max(\log\frac{p(x,y)}{p(x)p(y)}, 0) \quad (2)$$

감성 분석은 어떤 요소에 집중하느냐에 따라 다양한 방법론을 사용할 수 있다. TSA(Targeted Sentiment Analysis)는 특정 대상을 중심으로 감정을 분석하며, 구체적인 대상에 초점을 맞추는 장점이 있지만, 전체 콘텍스트를 무시할 수 있는 단점이 있다[20]. ABSA(Aspect-Based Sentiment Analysis)는 텍스트의 다양한 측면(Aspect)에 대한 감정을 파악하며, 상세한 분석이 가능한 장점이 있지만, 측면의 정의와 추출이 어려운 단점이 있다[21].

III. 제안 모델

제안한 뉴스 변화와 주체 감성 추이 분석을 위한 전체 시각화 모델링 과정은 그림 3과 같다.

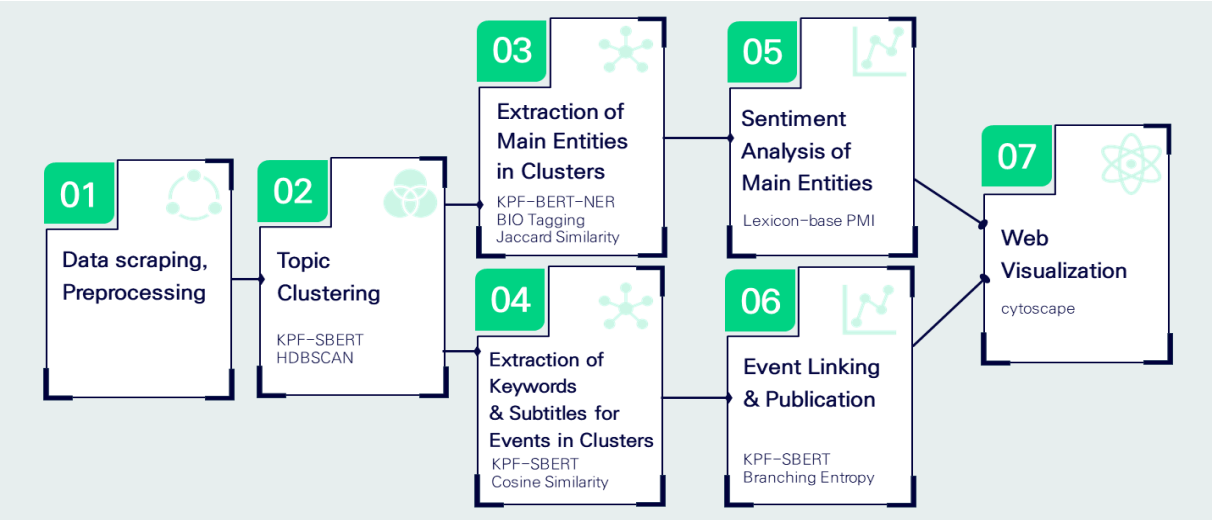

그림 3. 제안한 모델의 전체흐름도
Fig. 3. Overall flowchart of the proposed model

3.1 뉴스 데이터 스크래핑 및 전처리

네이버 뉴스에서 10개의 종합지(경향신문, 동아일보, 서울신문, 중앙일보, 한국일보, 국민일보, M news, 조선일보, 세계일보, 한겨레신문)에서 발행된 정치 기사를 기자 이름과 함께 수집하였다. 수집한 뉴스 중 사진과 사진에 대한 설명만 있는 사진 기사를 제거하고, 특수문자를 제거하거나 변환하여 전처리하였다. 추가로, 모델에서 기사 원문의 시퀀스가 길어 발생할 수 있는 문제를 예방하고자 원문에서 요약을 추출하였다. 표 1과 같이 전처리한 본문에는 마침표를 제외한 특수문자가 지워져 있고 요약은 전처리한 기사 본문 문장들 중 기사 제목과 가장 유사한 세 문장을 추출한다. 기사 제목 간의 유사도는 KPF-SBERT로 각 문장을 벡터화한 후, 제목과 각 문장 간의 코사인 유사도를 적용하였다.

표 1. 뉴스 데이터 세트
Table 1. News dataset

Original text	(...) 그는 “전북 새만금에서 개최된 (...) 못하였다”고 지적했다.
Preprocessing	(...) 그는 전북 새만금에서 개최된 (...) 못하였다 고 지적했다.
Summary	새만금서 부족했던 일정을 대한민국 문화의 (...)

3.2 뉴스 클러스터링

수집된 기사들을 문맥을 기준으로 클러스터링하고 기사 본문에서 추출한 요약은 KPF-SBERT로 임베딩하였다. 이후 UMAP 임베딩을 통해 고차원 데이터를 저차원으로 매핑하는 과정을 수행한다. HDBSCAN으로 클러스터링된 기사의 결과는 그림 4와 같다.

본문을 클러스터링의 입력으로 넣을 경우 본문에서 비교적 중요하지 않은 문장까지 임베딩되어 클러스터링의 성능이 낮아진다. 반면 기사 요약을 클러스터링 입력으로 적용할 경우 본문에서 기사 제목과 유사한 문장들이 추출되어 성능이 좋아지는 결과를 보였다.

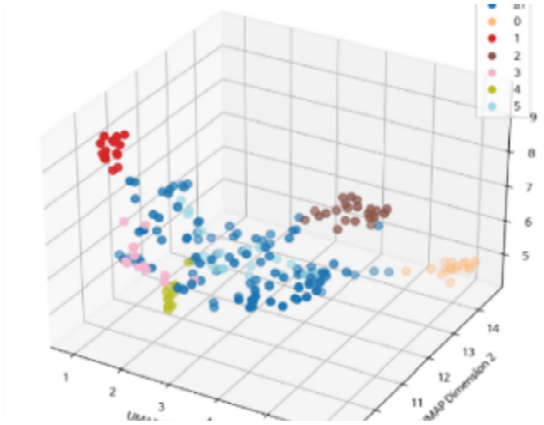

그림 4. HDBSCAN 클러스터링
Fig. 4. Clustering by HDBSCAN

비슷한 키워드 (e.g., 정당명, 정치인)가 자주 등장하는 정치 도메인 기사에서 키워드 기반 클러스터링보다는 문맥 기반 클러스터링이 더 우수한 결과를 보여줌으로 키워드 기반과 문맥 기반 클러스터링 방식 중 문맥 기반 클러스터링 방식을 선택하였다.

또한 KPF-SBERT의 학습 데이터가 뉴스 기사인 점, 문장 단위 임베딩을 통해 의미 손실을 줄인 모델이라는 특성을 고려하여 임베딩 모델은 KPF-SBERT로 선정하였고, 클러스터링 모델은 전역 밀도 임계값을 사용하는 DBSCAN보다 기사 데이터 클러스터링 특성을 반영하여 가변적인 차원에 높은 성능을 보여주는 HDBSCAN로 선정하였다[13].

3.3 클러스터별 키워드 추출

클러스터별 키워드 추출의 입력값은 기사 제목으로 지정하였다. 기사 제목으로 키워드를 추출하는 이유는 요약과 본문에서 키워드를 추출할 경우 부수적인 단어들이 키워드에 자주 추출되었기 때문이다. 기사 제목을 전처리하는 단계에서 기사 제목에 포함된 한자들을 한글로 변환하는 작업을 수행하였다. CountVectorizer를 사용하여 텍스트를 토큰화하고 불용어를 제한 후 다양한 후보 키워드를 추출하였다.

다음으로 KPF-SBERT를 이용해 기사 제목과 후보 키워드들의 임베딩을 생성한다.

마지막으로, 키워드 추출 단계에서는 MMR을 활용하여 후보 키워드와 기사 원제목 간의 임베딩의 코사인 유사도를 계산하고 다양한 키워드를 선택하

였다. 이를 통해 키워드의 중요성과 다양성을 고려하여 최종 키워드를 추출하였다.

3.4 클러스터별 주요 주체 추출

NER을 수행하는 모델인 KPF-BERT-NER 모델을 사용하여 모든 주체를 추출하였다. 이후 BIO tagging을 사용하여 연속적으로 나오는 인물과 직책을 묶어 하나의 주체로 처리하였고, 기관명 또한 주체로 추출하였다.

클러스터별로 기사에 나타난 주체들의 빈도수를 기준으로 기사에서 제일 많이 언급된 상위 5개 주체를 선정하였다. 표 2에서처럼, 개체명 인식에서 해결해야 하는 중요한 문제는 대용어 문제다[23]. 동일한 개체가 다른 형태로 재진술될 때 동일하다고 판단하기 위하여 자카드 유사도를 적용하여 대용어를 검출하였다. 추출된 주체들 사이에는 대용어가 많이 발견되었는데 자카드 유사도를 사용하여 단어 간의 유사도를 비교하고 임계치를 0.6으로 설정하여 대용어 쌍을 만들어 처리하였다.

표 2. 대용어 예시
Table 2. Anaphor examples

홍길동 OO그룹 회장	홍 회장, 그
한국은행	한은
전년 대비	(수치를 대신)

기사 특성상 명칭의 전체가 언급된 후 축약어가 등장하기 때문에 더 긴 단어를 대용어의 중심 단어(Main word)로 판단하였다. 인물 동의어가 기사에는 유독 많기 때문에 BIO태깅을 통해 인물 직위 · 직책을 묶어서 한 번에 처리하고 주체 간의 음절 유사도를 판단하기 위하여 가장 적합한 자카드 유사도를 사용하였다.

자카드 유사도와 코사인 유사도는 대표적인 유사인 측정 방법이다. 자카드 유사도는 두 벡터의 교집합의 크기를 두 벡터의 합집합의 크기로 나눈 값이다. 값은 0과 1 사이에서 변하며, 1에 가까울수록 두 벡터가 유사하다고 판단한다. 자카드 유사도는 키워드 간의 유사도를 판단하는 데에 적합한 것으로 알려져 있다[22].

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

표 3. 대용어 처리 예시
Table 3. Mapping anaphora to a main word

word	label	description	main word
강 수석대변인	PS_NAME	인물	강민국 수석대변인

NER(Named Entity Recognition)과정은 텍스트에서 명명된 실체(예: 사람 이름, 장소, 기관, 날짜 등)를 식별하고 분류하는 작업으로 많은 런타임이 소비된다. 이를 최소화하기 위해 본문을 input으로 개체명 추출하지 않고, 대신 요약을 input으로 하여 런타임을 최소화하였다. NER의 결과가 무작위 태깅된 개체명을 반환 (e.g. 명더불어민주당)할 경우 본문과 비교하여 그 개체명이 등장했는지 확인하고 일반적인 유사도 비교 방법인 코사인 유사도로는 대용어 문제를 해결할 수 없으므로 자카드 유사도가 더 적절하다.

3.5 주요 주체 감성 분석

클러스터별 주요 주체를 대상으로 감성 분석 과정을 거친다. 먼저, 기존의 기사 감성 분석 연구는 주가 예측을 목적으로 한 금융 기사에 한정해서 활발하게 이루어져 왔기 때문에, 기존 모델로 주체의 감성을 분석하기에는 적절하지 않다.

기존의 감성 사전은 용언 위주로 감성 사전이 구성되어 있어서, 정치 기사에서 '의혹'과 같이 빈번히 등장하는 부정적인 극성을 지닌 체언들을 제대로 다루지 못하는 문제점이 있다. 따라서, 정치 기사에 적합한 사전을 구축하기 위해 n-gram 분석을 통해 빈번하게 등장하고 극성을 지닌 체언을 식별하였다. 긍정적인 체언와 부정적인 체언 각각 20개로 사전을 구성하였다. 표 4의 정치기사 감성 사전은 기사 제목 및 내용에서 자주 등장하는 단어들을 기반으로 구성되어있다.

이후 주체와 극성을 가진 체언 간의 동시 출현 확률을 나타내는 PMI 행렬을 구성하기 위해 데이터 전처리 단계에서는 해당 날짜까지의 문장들을 형태소 분석하였다.

표 4. 정치 기사 감성 사전
Table 4. Sentiment dictionary for political news

Positive	Negative
극찬, 화합, 지지, 만회, 신뢰, 복구, 협력, 지원, 영웅, 평화, 번영, 환영, 찬사, 용기, 찬성, 개선, 청신호, 존경, 도약, 날개	고발, 위협, 하자, 논란, 갈등, 폭발, 충돌, 불안, 설전, 막말, 신경전, 위법, 파행, 피해, 저격, 심각, 탄핵, 비겁, 은폐, 비하

이때 '국민의힘'과 같이 여러 형태소가 합쳐진 고유명사의 경우 '국민, 의, 힘'과 같이 형태소가 분리되지 않도록, 데이터베이스에 등록된 모든 주체들을 사용자 사전에 추가하였다.

다음으로, 필요한 확률들을 계산하여 PMI 행렬을 생성하였다. 이 과정에서 불용어는 PMI 행렬에 반영하지 않았으며, 바이그램과는 다르게 순서를 고려하지 않고 한 문장에서 등장하면 동시 출현으로 간주하였다.

마지막으로, 극성값을 계산하는 단계에서는 긍정 단어와의 PMI의 합과 부정 단어와의 PMI의 합의 차이를 계산하였다. PMI 결과가 양수인 경우, 주체가 긍정 단어와 더 자주 함께 등장한 것이고 음수인 경우, 부정 단어와 더 자주 등장한 것으로 판단하였다. 특정 주체와 긍·부정 단어 간의 관련성을 파악하는 PMI 값을 도출하였고, 이를 통해 태스크에 적합한 TSA를 수행하였다.

표 5. 검찰의 PMI 행렬
Table 5. PMI matrix for prosecutor

	...	accusation	peace
prosecutor	...	3.2637	0.9284

예를 들어, 검찰(Prosecutor)의 경우 PMI 점수가 -4.3로 부정적인 극성값을 가진다. 이는 '개선', '협력' 같이 긍정적인 단어보다는 '고발'(Accusation), '피해'와 같은 부정적인 단어와 더 자주 등장했음을 의미한다.

3.6 사건 연결 및 발행, 사건 소제목 선정

다른 시점에 생성된 클러스터를 하나의 사건으로 연결하는 최종 단계로, 시간에 따른 사건의 흐름을 살펴보기 위한 가장 핵심적인 과정이다.

단순히 자카드 유사도만을 사용하여 클러스터 간의 유사성을 평가하여 하나의 사건으로 판단한다면, 문제점이 발생할 수 있다. 가령, 공통된 음절의 개수만을 고려하는 경우, 단어의 의미나 상황의 변화를 반영하지 못한다. 또한 당명이나 주요 정치 인명 등은 관련이 없는 여러 클러스터에서 자주 등장하기 때문에, 단순히 자카드 유사도만으로 판단하기 어렵다.

이를 해결하기 위해 KPF-SBERT를 이용하여 각 클러스터의 키워드를 임베딩하고 클러스터 임베딩 간의 유사도를 계산하였다. 이로써 의미를 고려하여 다른 시점의 클러스터를 하나의 사건으로 정교하게 연결할 수 있다.

우선, 과거 클러스터 20개의 키워드와 오늘 클러스터의 키워드를 SBERT로 임베딩한다. 이렇게 얻은 임베딩 값을 이용하여 클러스터 간 코사인 유사도를 계산한다. 유사도가 0.5 이상일 경우, 두 클러스터는 동일한 사건으로 연결되는 것으로 간주한다. 즉, 오늘의 클러스터가 과거의 클러스터의 후속 주제가 된다. 반면에 유사도가 0.5 미만인 경우에는 두 클러스터가 별개의 사건으로 간주되어, 오늘의 클러스터를 새로운 사건으로 발행한다. 예를 들어 세로축은 오늘의 클러스터이고 가로축은 오늘의 클러스터와 유사도를 계산할 과거의 클러스터 20개이다. 그림 5에서와 같이, 주제가 '잼버리 극복'인 클러스터는 '잼버리 외교', '잼버리 정상화'(Nomalization Jamboree) 클러스터와 높은 유사도를 보여준다. '잼버리 아수라장'(Chaotic Jamboree)을 주제로 하는 클러스터는 '잼버리 외교'(Jamboree diplomacy)와 '잼버리 공방'(Blame game about Jamboree) 클러스터와 높은 유사도를 나타내고 있다.

가장 최근에 연결된 클러스터의 기사 제목들 중 클러스터의 키워드를 가장 많이 포함한 기사 제목을 사건의 소제목으로 보여줌으로써 사건의 최근 동향을 소제목으로 파악할 수 있다.

새로운 사건을 발행할 때 사건의 이름을 짓기 위해 Branching Entropy의 개념을 적용하였다. Branching Entropy는 문자열에서 다음에 나올 글자의 불확실성을 의미하는 지표로, 이 값이 낮으면 다음에 나올 글자의 경우의 수가 적다는 것을 의미한다[24].

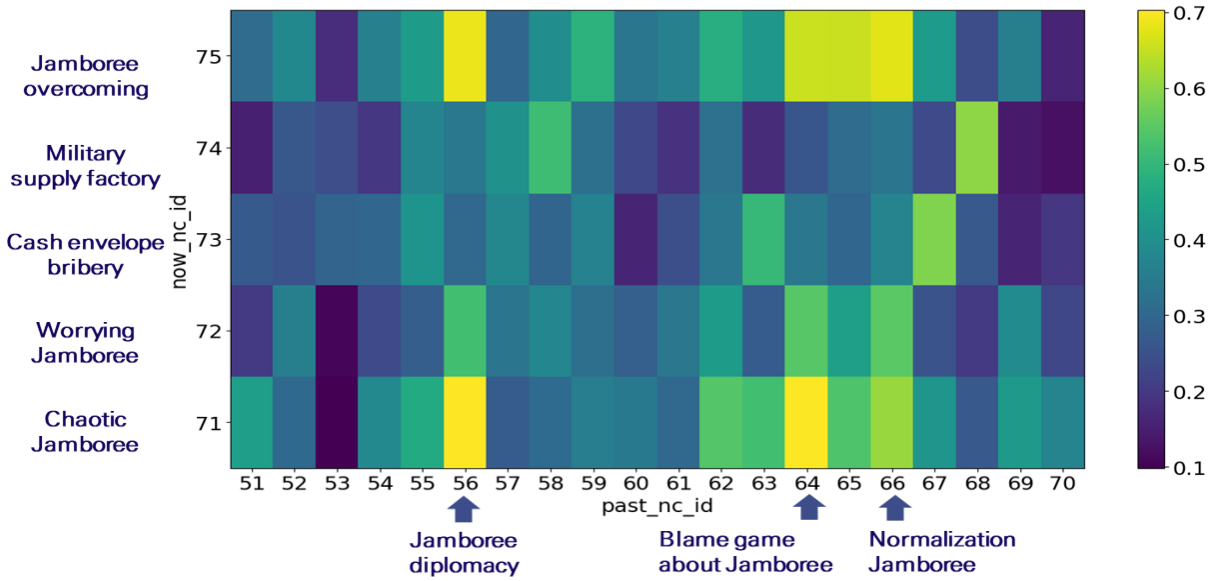

그림 5. 클러스터 간 연결
Fig. 5. Connecting between clusters

$$H(X|X_n) = -\sum_{x \in X} P(x|x_n) \times log(P(x|x_n)) \quad (4)$$

식 (4)의 H 수치가 낮을수록 한정적으로만 사용되는 단어로, 사건의 주요 키워드일 확률이 높은 단어라고 가정하였다. 이후 클러스터로 묶인 기사별 요약에서 명사만 추출하여 그 중 등장 빈도가 높은 순, Branching Entropy 점수가 낮은 순, 단어의 글자 수가 긴 순의 3가지 기준을 적용하여 상위 3개의 단어를 정하였다. 이러한 기준을 적용하여, '준비, 행사, 새만금'의 세 단어를 사건 이름으로 정한다.

Ⅳ. 웹 기반 시각화 구현 결과

웹 기반 시각화 시스템의 동작 과정은 다음과 같다. 그림 6은 제안한 기법으로 구현된 서비스 웹페이지 초기화면으로 최근 일어난 정치 사건 중 관심 있는 사건을 확인할 수 있다.

그림 7에서와 같이, 사건과 주요 주체, 그리고 기사들 간의 관계를 마인드맵으로 확인할 수 있다. 그림 8에서와 같이 사용자는 슬라이드 바를 통해 기간을 조절하여 사건의 시작부터 최근까지의 변화를 파악할 수 있다.

그림 6. 최근 주요 정치 사건
Fig. 6. Recent major political events

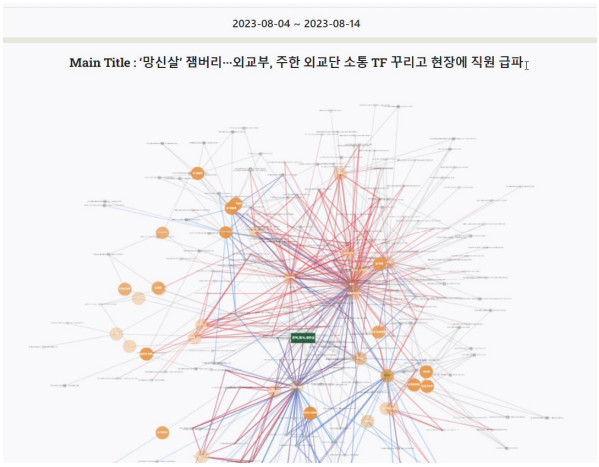

그림 7. 사건 마인드맵
Fig. 7. Event mindmap

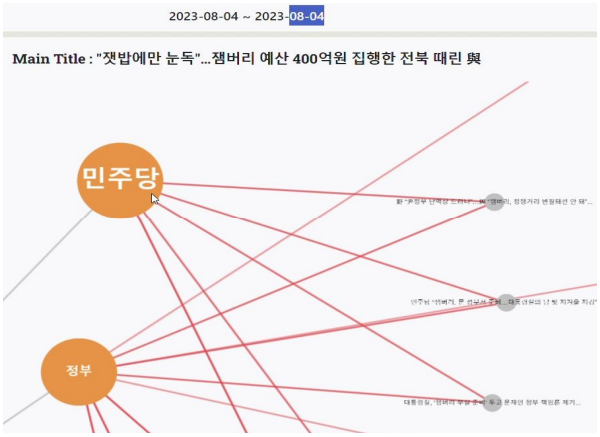

그림 8 사건 마인드맵
Fig. 8. Event mindmap

그림 9에서와 같이 마인드맵에서 노드와 기사 사이를 연결한 엣지의 색상을 통해 주요 주체가 기사에서 긍정적 혹은 부정적으로 서술되는지를 알 수 있다.

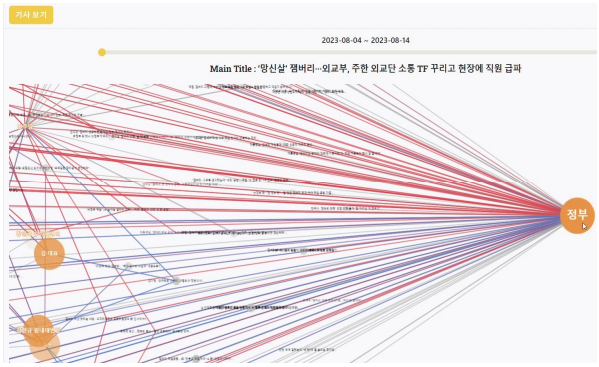

그림 9. 노드를 기사와 연결해주는 엣지
Fig. 9. Edge connecting the node to news

그림 10는 사이드 바를 클릭하여 뉴스 원문을 읽을 수 있으며 N사의 포털 사이트와 연결된다.

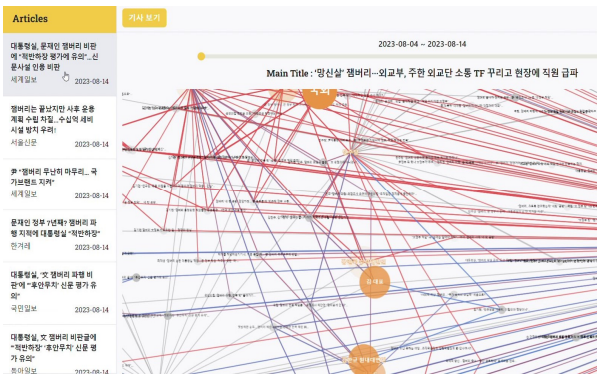

그림 10. 기사 목록 사이드바
Fig. 10. Article list sidebar

V. 결론 및 향후 과제

본 논문에서는 다양한 뉴스미디어의 출현으로 자극적인 뉴스와 낚시성 뉴스로 인하여 출처가 불분명한 뉴스에 접근하거나 이해하는 데 어려움을 느끼는 사용자를 위하여, 복잡한 정치 뉴스 내용을 간결하고 직관적으로 시각적으로 확인하는 방법을 제시하고 구현하였다. 기존 뉴스 서비스의 한계점을 극복하기 위한 방안으로 뉴스 클러스터의 주요 주체 분석과 시간에 따른 사건 변화의 시각화 방법을 제시하였다. 이를 통해 독자들이 뉴스의 복잡한 내용을 쉽게 이해하고 흐름을 파악할 수 있도록 도움을 제공하였다.

시뮬레이션 결과, 신규 독자에게 직관적인 형태의 뉴스를 제공하여 뉴스에 대한 접근성을 향상과 미디어 리터러시 증진을 기대할 수 있고, 기존 독자는 단순한 사실 파악을 넘어서 사회문제를 다각도로 바라볼 수 있는 분석 도구로 활용될 수 있다.

향후 연구과제로는, 뉴스 시각화 시스템에 다양한 언론사와 주제 영역을 포함시켜 결과의 다양성을 향상시킬 계획이다. 또한 사용자 경험을 향상시키기 위해 검색 기능의 도입과 클러스터 간 연결 기준의 다양화, 특정 주체나 장소와 관련된 사건의 아카이빙 시스템 구축에 대한 연구가 진행될 예정이다.

Acknowledgments

본 논문의 최종 투고에 이르기까지, 귀한 조언과 지도를 아끼지 않으신 유길상 교수님께 진심으로 감사드립니다. 고맙습니다.

References

[1] J. Choi and Y. Park, "Digital News Report 2022 Korea", Korea Press Foundation, pp. 12-16, 2022.
[2] News Bigdata & Analytics BIGKinds, https://www.bigkinds.or.kr/ [accessed: Sep. 19, 2023]
[3] A. Vaswani, et al., "Attention is all you need", Advances in neural information processing systems, Jun. 2017.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", in Proc. ACL, Vol. 1, pp. 4171-4186, Jun. 2019. https://doi.org/10.18653/V1/N19-1423.

[5] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", in Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP), pp. 3980-3990, Aug. 2019. https://doi.org/10.18653/v1%2FD19-1410.

[6] BIGKinds LAB, https://lab.bigkinds.or.kr/kpfBertIntro.do/ [accessed: Sep. 22, 2023]

[7] H. P. Luhn, "A Statistical Approach to Mechanized Encoding and Searching of Literary Information", in IBM Journal of Research and Development, Vol. 1, No. 4, pp. 309-317, Oct. 1957. https://doi.org/10.1147/rd.14.0309.

[8] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval", Information processing & management, Vol. 24, No 5, pp. 513-523, 1988. https://doi.org/10.1016/0306-4573(88)90021-0.

[9] KeyBERT, https://github.com/MaartenGr/KeyBERT/ [accessed: Sep. 20, 2023]

[10] D. Blei, A. Ng, and M. Jordan. "Latent dirichlet allocation", Journal of machine Learning research, pp. 993-1022, Jan. 2003.

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

[12] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform Manifold Approximation and Projection", Journal of Open Source Software, Vol. 3, No. 29, Sep. 2018. https://doi.org/10.21105/joss.00861.

[13] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates", Pacific-Asia conference on knowledge discovery and data mining, Gold Coast, QLD, Australia, Vol. 7819, pp. 160-172, Apr. 2013. https://doi.org/10.1007/978-3-642-37456-2_14.

[14] R. J. G. B. Campello, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. 10, No. 2, pp. e1343, Mar. 2020. https://doi.org/10.1002/widm.1343.

[15] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries", Proc. of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, New York, United States, pp. 335-336, Aug. 1998. https://doi.org/10.1145/290941.291025.

[16] BIGKinds LAB KPF-SBERT-NER, https://github.com/KPF-bigkinds/BIGKINDS-LAB/blob/main/KPF-BERT-NER/ [accessed: Sep. 22, 2023]

[17] W. Kim, S. Lee, and J. Lee, "Improving the Accuracy of Extracting Sentiment in Korean Text through the BIO Tagging and Triplet Methods", Journal of Foreign Studies, Vol. 57, pp. 345-366, Sep. 2021. http://doi.org/10.15755/jfs.2021..57.345.

[18] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin, "A survey of sentiment analysis in social media", Knowledge and Information Systems, Vol. 60, pp. 617-663, Jul. 2018. https://doi.org/10.1007/s10115-018-1236-4.

[19] F. H. Khan, U. Qamar, and S. Bashir, "SentiMI: Introducing point-wise mutual information with SentiWordNet to improve sentiment polarity detection", Applied Soft Computing, Vol. 39, pp. 140-153, Feb. 2016. https://doi.org/10.1016/j.asoc.2015.11.016.

[20] Z. Gao, A. Feng, X. Song, and X. Wu, "Target-Dependent Sentiment Classification With BERT", IEEE Access, Vol. 7, pp. 154290-154299,

Oct. 2019. https://doi.org/10.1109/ACCESS.2019.2946594.

[21] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, "A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges", IEEE Transactions on Knowledge and Data Engineering, Vol. 35, No. 11, pp. 11019-11038, Nov. 2023. https://doi.org/10.1109/TKDE.2022.3230975.

[22] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, "Using of Jaccard coefficient for keywords similarity", Proc. of the international multiconference of engineers and computer scientists, Hong Kong, Vol. 1, No. 6, pp. 380-384, Mar. 2013.

[23] D. Park, "Natural Language Processing of News Articles: A Case of NewsSource beta", Communication Theories, Vol. 12, No. 1, pp. 4-52, Mar. 2016.

[24] V. Zhikov, H. Takamura, and M. Okumura, "An efficient algorithm for unsupervised word segmentation with branching entropy and MDL", Information and Media Technologies, Vol. 8, No. 2, pp. 514-527, Aug. 2013. https://doi.org/10.11185/imt.8.514.

저자소개

손 효 원 (Hyowon Son)

2023년 8월 : 고려대학교 데이터청년캠퍼스(빅데이터 기반의 지능 정보 시스템 개발 과정, 350H) 수료
2020년 3월 ~ 현재 : 한국외국어대학교 ELLT학과 학사과정
관심분야 : 데이터사이언스, 머신러닝, 자연어 처리, 딥러닝

한 영 민 (Youngmin Han)

2023년 8월 : 고려대학교 데이터청년캠퍼스(빅데이터 기반의 지능 정보 시스템 개발 과정, 350H) 수료
2015년 3월 ~ 현재 : 홍익대학교 건설도시공학부 도시공학과 학사과정
관심분야 : 데이터사이언스, 빅데이터 분석, 자연어 처리, 머신러닝

남 경 현 (Kyounghyun Nam)

2023년 8월 : 고려대학교 데이터청년캠퍼스(빅데이터 기반의 지능 정보 시스템 개발 과정, 350H) 수료
2019년 3월 ~ 현재 : 동국대학교 산업시스템공학과 학사과정
관심분야 : 데이터사이언스, 빅데이터 분석, 머신러닝, 딥러닝, 자연어 처리

한 수 빈 (Subin Han)

2023년 8월 : 고려대학교 데이터청년캠퍼스(빅데이터 기반의 지능 정보 시스템 개발 과정, 350H) 수료
2020년 3월 ~ 현재 : 고려대학교 컴퓨터학과 학사과정
관심분야 : 딥러닝, 머신러닝, 데이터사이언스, 빅데이터 분석, 자연어 처리

유 길 상 (Gilsang Yoo)

2010년 3월 ~ 현재 : (사)한국컴퓨터게임학회 이사
2021년 3월 ~ 현재 : 고려대학교 정보창의교육연구소, 지능정보 SW아카데미 교수
2023년 3월 ~ 현재 : (사)한국미디어아트산업협회 수석부회장
관심분야 : 데이터사이언스, 데이터 시각화, 빅데이터 분석, 3D영상 콘텐츠, 머신러닝, 딥러닝, 컴퓨터교육