심층 주제, 지역, 장르를 모두 분류할 수 있는 다면적 뉴스 기사 자동 분류 모델 연구

Research on Multi-facted News Article Classification Models Classifying Subjects, Geographies and Genres

이 효 진 (Hyojin Lee)*
최 성 필 (SungPil Choi)**

목 차

초 록

본 연구는 한국어 사전학습 모델을 활용하여 뉴스 기사를 주제, 장르, 지역별로 각각 분류하는 모델을 구축하였다. 이를 위해 국내 언론사의 분류체계를 참고하여 새로운 뉴스 기사 분류체계를 설계하였다. 주제 및 장르 분류 모델은 대분류와 중분류 모델을 연결한 계층적 구조의 분류 모델로 구현하여 카테고리 통합 모델의 성능과 비교하였다. 평가 결과, 계층적 구조의 분류 모델은 모호하거나 중복된 카테고리에서 카테고리 통합 모델보다 더 명확한 분류를 수행할 수 있다는 이점이 있었다. 뉴스 기사의 지역적 분류를 위해서는 18개의 카테고리에 대하여 분류를 수행하는 모델을 구축하였으며 지역 관련 뉴스 기사의 경우, 지역적 특성이 본문에 명확히 드러나 높은 성능을 기록할 수 있었다. 본 연구는 주제, 장르, 지역의 다각적인 측면에서 뉴스 기사를 효과적으로 분류할 수 있음을 보여주었으며, 이를 통해 사용자 요구에 부합하는 다차원적 뉴스 기사 분류 서비스의 가능성을 제시한 점에서 의의가 있다.

ABSTRACT

This study developed a model to classify news articles into categories of topic, genre, and region using a Korean Pre-trained Language model. To achieve this, a new news article classification system was designed by referring to the classification systems of domestic media outlets. The topic and genre classification models were implemented as hierarchical classification models that link the main categories and subcategories, and their performance was compared with that of an integrated category model. The evaluation results showed that the hierarchical structure classification model had the advantage of providing more precise categorization in ambiguous or overlapping categories compared to the integrated category model. For regional classification of news articles, a model was built to classify into 18 categories, and for regional news articles, the regional characteristics were clearly reflected in the text, resulting in high performance. This study demonstrated the effectiveness of classifying news articles from multiple perspectives—topic, genre, and region—and emphasized the significance of suggesting the potential for a multi-dimensional news article classification service that meets user needs.

키워드: BERT 모델, 뉴스 기사 분류, 계층적 분류 모델, 다중 클래스 분류 모델, 다차원 분류
BERT Model, News Article Classification, Hierarchical Classification Model, Multi-Class Classification Model, Multidimensional classification

* 경기대학교 문헌정보학과 석사과정(caveel185@gmail.com / ISNI 0000 0005 1751 5585) (제1저자)
** 경기대학교 문헌정보학과 부교수(sungpil@gmail.com / ISNI 0000 0004 6772 9269) (교신저자)
논문접수일자: 2024년 7월 26일 최초심사일자: 2024년 8월 13일 게재확정일자: 2024년 8월 20일
한국문헌정보학회지, 58(3): 65-89, 2024. http://dx.doi.org/10.4275/KSLIS.2024.58.3.065

1. 서 론

정보 기술의 발달은 사회 전반적인 영역에 걸쳐 변화를 촉발하고 있다. 특히, 뉴스 산업은 이러한 기술적 진보에 의해 근본적인 변화를 겪고 있다.

디지털 기술의 등장 이전, 뉴스 기사의 유통 및 소비는 주로 인쇄매체를 통해 이루어져 왔다. 그러나 디지털 기술의 등장 이후에는 인터넷 및 모바일 기기의 활용이 일반화되면서 온라인 플랫폼을 통해 실시간으로 뉴스 기사가 유통되고 소비되는 형태로 변모하고 있다. 빅데이터, 인공지능 등의 디지털 기술을 활용하여 이용자 개개인의 취향 및 관심사를 반영한 이용자 맞춤형 서비스 제공이 증가함에 따라 현대 사회의 뉴스 산업에서는 소비자의 정보 요구에 효과적으로 대응할 수 있는 정확하고 신속한 정보 제공 전략을 구축하는 것이 중요해지고 있다. 이러한 변화 속에서 뉴스 기사를 체계적으로 분류하고 조직화하는 작업은 뉴스 소비자들이 원하는 정보에 쉽게 접근할 수 있도록 도와 뉴스 소비의 효과를 크게 향상한다는 면에서 필수적이라고 할 수 있다.

이에 본 연구에서는 세분화된 뉴스 기사의 다양한 주제를 효과적으로 분류할 수 있는 분류 모델을 개발하고자 하였다. 기존의 뉴스 기사 분류 연구들은 주로 특정 주제에 국한되어 있거나 대분류 수준에서의 분류에 그치는 경우가 많아 뉴스 기사의 주제 다양성을 충분히 반영하지 못하고 있다. 본 연구는 주제뿐만 아니라 장르, 지역과 같이 서비스 측면에서의 뉴스 기사 분류를 수행하고자 한다. 또한, 분류의 범위를 대분류 수준에서 그치지 않고 중분류까지 확장하여 보다 세밀하고 정교한 분류를 수행하는 것을 목표로 하였다.

2. 관련 연구

뉴스 기사를 대상으로 카테고리 분류를 수행한 연구를 보면, 김덕기, 온병원(2023)은 KoBERT를 활용하여 한국어 뉴스 기사를 문맥 벡터로 변환한 후, 클러스터링 기법을 적용하여 정치, 기술/IT, 스포츠 카테고리에 대하여 분류를 수행하였다. DBSCAN에서 가장 낮은 성능을 보였으나 클러스터 개수를 지정하지 않은 상태에서 카테고리별 분류를 수행하였을 뿐만 아니라 이상치 감지 및 클러스터 내 작은 클러스터들을 형성하여 세부 분류를 수행했다는 점에서 주목할 만하다. 김혜영(2015)은 대규모 신문 기사 자동 분류를 위해 토픽 모델링 기법을 활용하여 복합 주제를 가진 신문의 분류 방안에 대해 논의하였다. MALLET를 적용한 토픽 모델링 기법을 사용하여 수작업 분류의 방법으로 분류된 문서들에 대해 3단계로 코드 수를 늘려 분류 양상을 살펴보았다. 그 결과, 7개 주제로 분류되었던 기사들이 각 주제 내부에서 세부 주제로 세밀하게 분류되어 토픽 모델링을 활용한 자동 분류가 신문 기사의 주제 분석에 유용할 수 있음을 확인하였다. 이재욱, 고병규, 김판구(2016)는 빈도수 기반 지도학습의 빈도수 차이에 따른 카테고리 오분류 문제를 개선하기 위해 로그 정규화와 상호 정보량을 활용한 뉴스 기사의 정치, 스포츠, 사회, 문화, 경제 카테고리 분류 방법을 제시하였다.

특정 분야의 기사에 대하여 카테고리 분류를

수행한 연구를 보면, 장지형, 홍참길(2022)은 데이터가 부족한 북한 및 통일 관련 기사를 활용하여 정치, 경제, 국제, 생활문화, 스포츠 등 5개 분야에 대한 뉴스 기사 자동 주제 분류를 시도하였다. 연구 결과, 분류 모델들은 약 87-91%의 사이의 정확도를 보였다. 또한, 정치적인 소재를 다루는 기사에서 예측 오류가 높았으며, 이는 데이터 지면 간의 모호성이 분류 모델에도 반영되는 것으로 추측되었다. 김미선(2022)은 농업분야 신문기사를 대상으로 통합 분류체계를 통해 분류 기준을 선정하고, 분류 특징을 추출한 후, BERT를 활용하여 자동 분류하는 알고리즘을 제안하였다. 연구 결과, 단일 기사의 핵심 키워드 추출 방안 적용한 경우, 자동분류기의 성능이 향상됨을 보여주었다. 성나영, 구명완(2018)은 담배와 관련된 인터넷 기사를 대상으로 메모리 네트워크 모델을 활용하여 제품, 기업, CSR활동이라는 3개의 카테고리로 분류하는 연구를 수행하였다. 해당 연구에서는 WPM을 기반으로 한 MemN2N 모델의 성능이 91.9%로 가장 우수한 것으로 나타났다. 강승태, 장길진(2023)은 COVID-19의 전파 및 해결방안 도출을 위해 다국어 기사들을 자동 분류하는 방법을 제안하였다. 감염병 기사 관련 사건을 자동 분류하는 모델을 위해 확진자수, 완치자수, 사망여부, 집단감염 등의 10가지 사건 분류체계를 제안하였고, 다국어 버트를 활용하여 실험을 수행하였다.

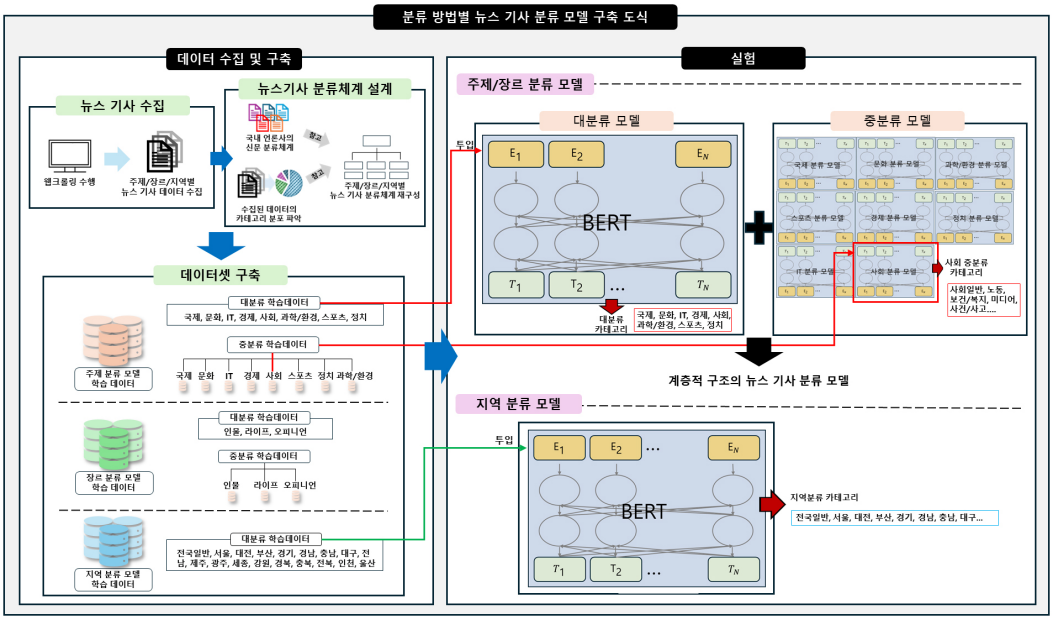
뉴스 기사 분류에 대한 연구는 지속적으로 진행되고 있지만 대다수의 연구가 특정 분야에 대한 기사 분류나 대분류 수준에서의 접근에 초점을 맞추고 있다는 점에서 한계를 보인다. 일반적으로 정치, 경제, 사회, 문화, 스포츠 등 주로 사람들이 많은 관심을 가지는 분야나 뉴스에서 자주 다루는 카테고리를 중심으로 이루어지며, 특정 주제나 카테고리에 대한 분류에 중점을 두고 있어 뉴스 기사의 세부 분류나 뉴스 기사 텍스트의 주제 다양성을 고려한 연구는 상대적으로 부족한 실정이다. 또한, 대부분의 연구가 주제 분류에만 제한되어 있고, 특정 사용자 또는 고객을 위해 개별화된 서비스 제공에 필요한 서비스 측면에서의 기사 분류 방법을 시도한 연구는 드물다. 이와 같은 연구 경향은 국내 뉴스 기사의 통일된 분류체계 부재와 연관이 있다고 생각된다.

디지털 기술의 발전과 함께 정보의 생산과 소비 방식이 급격하게 변화하고 있는 현대 사회에서 뉴스 기사 분류 연구는 다양한 주제와 카테고리를 포괄하고, 사용자들의 다양한 관심사를 반영할 수 있도록 세부 분류 및 서비스 측면에서의 분류에 대한 연구가 확대되어야 할 필요가 있다. 이에 본 연구는 뉴스 기사의 다양한 주제를 효과적으로 분류할 수 있는 분류 모델을 개발하고자 뉴스 기사 텍스트의 주제 다양성을 고려한 뉴스 기사의 세부 분류체계를 제안하여 기존 뉴스 기사 분류 연구의 한계를 극복하고자 한다.

3. 실험 설계

본 연구에서는 뉴스 기사에 대하여 심층 주제, 장르, 지역 분류를 효과적으로 수행할 수 있는 분류 모델 개발을 위하여 〈그림 1〉과 같이 실험을 진행하였다.

본 연구에서 구축한 뉴스 기사 분류 모델은



〈그림 1〉 분류 방법별 뉴스 기사 분류 모델 구축 도식

주제 분류 모델, 장르 분류 모델, 지역 분류 모델 총 3개이다.

먼저, 뉴스 기사를 수집한 후, 주제별, 지역별, 장르별 분류체계를 설계하였고, 이를 기반으로 실험데이터를 구축하였다. 뉴스 기사의 주제와 장르 분류 모델은 〈그림 1〉과 같이 뉴스 기사의 대분류를 담당하는 "대분류 모델"과 각 대분류 내에서 세부 카테고리를 분류하는 "중분류 모델"을 각각 학습시킨 후, 대분류 모델과 중분류 모델을 연결하여 "계층적 구조의 뉴스 기사 분류 모델"을 구현하였다. 지역 분류 모델은 총 18개의 지역 카테고리에 대해 분류를 수행하는 대분류 모델 하나로 구성하였다.

3.1 데이터 수집

본 연구에서는 데이터 수집을 위해 국내 언론사를 대상으로 뉴스 기사 크롤링을 수행하였다. 뉴스 기사 카테고리의 다양성을 위해 경향신문, 문화일보, 뉴시스 등 뉴스 기사를 중분류까지 분류한 언론사들의 뉴스 기사를 대상으로 제목, 본문, 카테고리명을 수집하였다. 지역 분류체계의 경우, 일부 언론사는 지역 분류를 수행하지 않기 때문에, 지역별 뉴스 기사는 지역 분류체계가 존재하는 언론사를 대상으로 뉴스 기사를 수집하였다.

크롤링을 통해 수집된 뉴스 기사 데이터는 약 192,000건이다. 주제별 뉴스 기사가 약 150,000건, 장르별 뉴스 기사가 약 17,000건, 지역별 뉴스 기사가 약 25,000건 수집되었다.

3.2 뉴스 기사 분류체계 설계

본 연구에서는 다양한 언론사의 뉴스 기사를

수집하면서 각 언론사가 사용하는 뉴스 분류체계에 차이가 있음을 확인하였다. 예를 들어, 국방과 국방/외교, 보건/질병과 보건/복지와 같이 카테고리의 포괄 범위, 구성, 명칭 등에서 차이가 존재하였다. 이는 국내 신문사들이 통일된 분류체계를 갖고 있지 않아 언론사별로 특정 주제에 대한 중요성이나 분류체계에 대한 접근 방식이 상이한 것으로 판단된다. 이러한 분류체계의 비일관성은 모델이 학습하는 과정에서 혼동을 일으킬 수 있어 모델의 성능에 부정적인 영향을 미칠 수 있다. 따라서 모델의 성능 향상을 위하여 체계적이고 일관된 뉴스 기사 분류체계를 작성할 필요가 있었다. 이에 본 연구에서는 2024년을 7월을 기준으로 국내 주요 언론사들의 홈페이지에서 제공하는 뉴스 카테고리 항목(이하 분류체계)을 분석하여 주제별, 장르별, 지역별 새로운 뉴스 기사 분류체계를 작성하였다.

3.2.1 뉴스 기사 주제 분류체계

먼저, 뉴스 기사의 주제 카테고리는 기사 내용의 중심 소재나 분야를 나타내며, 구체적인 영역에서 다루는 정보나 사건을 포함하는 카테고리로 정의하였다. 뉴스 기사 주제 분류체계 구성을 위하여 경향신문(경향신문, 2024), 문화일보(문화일보, 2024), 한국일보(한국일보, 2024), 중앙일보(중앙일보, 2024), 뉴시스(뉴시스, 2024)의 분류체계를 기반으로 새로운 분류체계를 제시하였다. 다만, 경향신문의 경우, 경향신문, 스포츠경향, 주간경향, 레이디경향의 4개 매체가 존재하여 이들의 분류체계를 모두 가져오되 일부 카테고리에 대하여 통합 및 제거를 수행하였다. 언론사별 주제 분류체계는 〈표 1〉, 〈표 2〉, 〈표 3〉, 〈표 4〉, 〈표 5〉와 같다.

5개 언론사의 주제 분류체계를 살펴본 결과, 정치, 경제, 사회, 국제, 스포츠, 문화 카테고리가 공통으로 존재함을 확인하였다.

정치 카테고리의 경우, 중분류에 국방, 외교 카테고리가 분리되거나 통합되어 존재하였고, 북한과 관련된 카테고리의 경우, 북한/한반도, 북한/통일, 북한 등 다양한 명칭으로 존재하였다. 이에 국방과 외교 카테고리는 통합하여 국방/외교로, 북한 관련 카테고리는 모두 북한/한반도로 통일하였다.

경제의 중분류 카테고리의 경우, 언론사별로 세분화한 정도가 달랐다. 이에 언론사의 분류체계와 수집 데이터의 카테고리 분포를 참고하여 중분류 카테고리를 경제일반, 산업/통상, 금융/

〈표 1〉 문화일보 주제 분류체계

대분류	중분류
정치	정치일반, 국회/정당, 대통령실, 선거, 여론조사, 행정, 외교, 국방, 북한/통일
경제	경제일반, 재정/재무, 금융, 증권/주식, 부동산, 무역/통상, 산업/기업, 유통/생활, 취업/창업, IT, 과학, 모빌리티, 국제경제
사회	사회일반, 사건/사고, 교육/청소년, 환경, 법원/검찰, 보건/의료/식품, 노동/복지
국제	국제일반, 아시아, 유럽, 북미/중남미, 중동, 오세아니아, 아프리카
스포츠	스포츠일반, 골프, 야구, 축구, 농구, 배구, 올림픽, 경마, 기타종목, e스포츠
문화	문화일반, 도서/출판/학술, 문학, 공연, 전시, 문화재, 종교, 여행, 스타일/패션, 생활/요리, 의학/건강
연예	연예일반, 가요/팝, 영화, 방송/연예, 만화/게임

〈표 2〉 중앙일보 주제 분류체계

대분류	중분류
정치	정치일반, 국회/정당, 대통령실, 외교, 국방, 북한
경제	경제정책, 경제일반, 산업, 금융증권, 부동산, IT/과학, 고용노동, 글로벌경제
사회	사건/사고, 검찰/법원, 교육, 복지, 보건/질병, 환경, 교통, 전국, 사회일반
국제	국제일반, 미국, 중국, 일본, 유럽, 기타
스포츠	스포츠일반, 야구, 축구, 골프, 농구/배구
문화	문화일반, 책, 미술/전시/문화재, 클래식/공연, 가요, 영화, 방송

〈표 3〉 경향신문 주제 분류체계

대분류	중분류
정치	대통령실, 국회/정당, 국방/외교, 북한/한반도, 선거, 정치일반
경제	금융/재테크, 산업, IT/가전, 부동산, 자동차, 생활경제, 취업/창업, 경제일반, 경제인, 기업소식
사회	사건/사고, 법원/검찰, 교육/입시, 노동, 보건/복지, 미디어, 젠더, 사회일반
국제	미국/중남미, 일본, 중국/대만, 유럽, 아시아/호주, 중동/아프리카, 국제일반
과학/환경	우주/항공, 기후/날씨, 환경/생태, 과학/환경일반
스포츠	야구, 축구, 골프, 농구/배구, 스포츠 종합, 격투기, 종합스포츠, 생활체육, 월드컵, 올림픽/아시안게임, 경주
문화	책, 연극/클래식, 학술/문화재, 종교, 미술/건축, 대중음악, 영화, 방송, 문화일반
연예	스타, 방송, 음악, 영화

〈표 4〉 한국일보 주제 분류체계

대분류	중분류
정치	정치일반, 국회/정당, 대통령실, 외교, 국방/북한
경제	경제정책, 산업, 금융/증권, 부동산, IT/과학, 경제일반
사회	사건/사고, 법원/경찰, 교육, 노동, 보건/복지, 날씨/환경, 사회일반
국제	아시아/호주, 미국/중남미, 유럽, 중동/아프리카, 글로벌이슈, 국제일반
스포츠	야구, 축구, 농구, 배구, 해외스포츠, 골프, 바둑, 스포츠일반
문화	책, 공연/전시, 문화일반
연예	연예일반, 영화, 방송, 음악, 실시간 연예, 포토 뉴스

〈표 5〉 뉴시스 주제 분류체계

대분류	중분류
정치	정치최신, 대통령실, 국회/정당, 국방/외교, 북한, 행정, 지방정가
경제	경제최신, 경제정책, 국제경제, 건설부동산
사회	사회최신, 사건/사고, 법원/검찰, 의료/보건, 복지, 교육, 노동, 환경, 날씨, 수능
국제	국제최신, 국제기구, 미주, 유럽, 중국, 일본, 아시아/오세아니아
스포츠	야구, 해외야구, 축구, 해외축구, 농구, 배구, 골프, 스포츠최신
문화	문화최신, 공연, 전시, 문화재, 책, 여행/레저, 종교
금융	금융최신, 증권, 금융정책/통화, 은행, 보험/카드, 시황/환율, 종목분석, 재테크, 블록체인, 은행은 지금
산업	산업최신, 산업/ESG, 유통/생활경제, 패션/뷰티, 음식/맛집, 창업/취업, 자동차/항공, 전기/전자, 중기/벤처, 해양수산
IT/바이오	모바일, 인터넷/SNS, 통신, 보안, 컴퓨터/SW, 게임, 과학, 제약/바이오, 의료기기/헬스케어, 병의원, 건강/질환, 출산/육아
연예	연예최신, 방송/TV, 영화, 가요, 드라마, 예능, 해외연예, 스타

재테크, 자동차, 생활경제/소비자, 부동산, 기업 소식, 경제인, 창업/취업 9가지로 분류하였다. 산업/통상 카테고리에는 무역, 통상, 산업 등 산업 및 통상 활동과 관련된 주제를 포함하였다. 금융/재테크 카테고리에는 증권, 금융, 은행, 재테크, 주식, 재정 등 개인 및 기업의 자산 관리와 관련된 주제를 포함하였다. 생활경제/소비자 카테고리에는 소비 트렌드 변화와 관련된 주제를 포함하였다.

또한, 경제 중분류 카테고리에 IT, 과학, IT/과학, IT/가전 등 IT와 과학 카테고리가 속해 있는 것을 볼 수 있다. 언론사별로 경제 카테고리의 IT 및 과학 관련 카테고리의 포함 여부가 달랐고, 독립적으로 분류해 놓았을 때의 세분화 정도도 상이하였다. 이는 IT와 과학 카테고리가 경제와 사회 전반에 종합적으로 큰 영향을 미치는 주제임에 나타난 것으로 생각된다. 이에 수집 데이터의 IT 및 과학 카테고리의 분포를 살펴본 결과, 각각의 카테고리에 대한 중분류 수행이 가능하다고 판단하였고, 카테고리의 다양성을 위해 IT와 과학 카테고리를 독립적으로 분류하였다. 따라서 수집 데이터의 카테고리 분포를 기준으로 하여 과학 카테고리는 경향신문의 과학/환경 카테고리와 동일하게 설정하였고, IT 카테고리는 카테고리 통합 및 제거를 통해 중분류 카테고리를 IT비즈/정책, 모바일, PC/기기가전, 게임으로 구분하였다. IT비즈/정책 카테고리에는 IT 산업의 비즈니스 정책과 관련된 모든 주제가 포함되며, 모바일 카테고리는 스마트폰 및 모바일 기기, 모바일 앱, 이동통신 등의 주제가 포함된다. PC/기기가전 카테고리는 컴퓨터, 하드웨어, 소프트웨어, 가전제품 관련 주제들을 포함하고 있으며, 게임 카테고리는 게임 및 e스포츠 관련 주제를 포함하고 있다.

사회의 중분류 카테고리는 대부분 카테고리명이나 분류의 범위가 다른 경우가 많았다. 따라서 수집 데이터의 카테고리 분포를 참고하여 유사한 주제의 카테고리는 통합하고, 불필요한 카테고리는 제거하였다. 날씨 카테고리의 경우, 과학/환경의 기후/날씨 카테고리에 포함되는 주제이므로 과학/환경〉기후/날씨 카테고리로 수정하였다. 최종적으로 사회의 중분류 카테고리는 사회일반, 법원/검찰, 보건/복지, 미디어, 사건/사고, 노동, 교육/입시, 젠더 8가지로 구성되었다.

국제 중분류 카테고리는 미국/중남미, 미주, 북미/중남미 등과 같이 동일한 주제의 카테고리명이 다른 경우에 대하여 카테고리명을 통일하였다. 또한, 중동, 아프리카, 중동/아프리카와 같이 분류 범위가 다른 경우에는 보다 넓은 범주에 해당하는 카테고리로 통합하였다.

스포츠 카테고리의 e스포츠의 경우, 게임 카테고리로 분류를 수정하였으며, 경주 카테고리를 추가하여 경마, 경륜, 경정 등의 경주 스포츠에 대해 분류가 가능하게 하였다.

본 연구에서는 문화 카테고리를 IPTC의 Media Topic NewsCodes의 arts, culture, entertainment and media 정의에 기반하여 인간의 정신, 흥미, 기술, 기호, 감정의 진보와 관계되는 분야로 모든 형태의 예술, 엔터테인먼트, 문화유산 및 미디어를 포함하는 주제로 정의하였다(IPTC, 2024). 이에 스타, 방송, 음악, 영화 등의 방송매체 관련 카테고리인 연예를 방송/연예 카테고리로 명칭을 바꾸고 문화의 중분류 카테고리에 포함하였다. 또한, 문화 카테고리에서 여행,

스타일, 레저, 연극/클래식 등 라이프 카테고리의 주제와 유사한 카테고리들은 라이프 카테고리로 변경하였다.

언론사의 분류체계들을 바탕으로 새롭게 정리한 분류체계를 적용하여 수집 데이터의 카테고리명을 통합, 제거 및 수정하고, 데이터가 부족한 카테고리를 제거한 결과, 〈표 6〉의 새로운 기사 주제 분류체계가 구성되었다.

3.2.2 뉴스 기사 장르 분류체계

뉴스 기사의 장르 카테고리는 기사 내용의 성격, 독자와의 상호작용 방식, 정보 전달의 목적 등에 따라 다양한 형식으로 분류되는 카테고리로 정의하였다. 뉴스 기사의 장르 분류체계를 재구성하기 위해 수집된 기사의 장르 카테고리 분포를 살펴보았다. 수집된 기사는 주로 라이프, 사람(피플), 오피니언, 기획/연재, 이슈 등의 카테고리에 해당되었다. 더불어 경향신문(경향신문, 2024), 중앙일보(중앙일보, 2024), 한국일보(한국일보, 2024), 문화일보(문화일보, 2024) 4개 언론사의 분류체계도 분석하였다. 언론사별 장르 분류체계는 〈표 7〉, 〈표 8〉, 〈표 9〉, 〈표 10〉과 같다. 그리고 주제 분류체계 중 문화의 중분류 카테고리에 해당하는 일부 기사를 장르 분류체계의 라이프 카테고리로 변경하였다. 이는 〈표 9〉와 〈표 10〉에 표기하였다.

〈표 6〉 본 연구에서 구축한 뉴스 기사 주제 분류체계

대분류	중분류
국제	국제일반, 아시아/호주, 일본, 미국/중남미, 유럽/러시아, 중동/아프리카, 중국/대만
문화	문화일반, 미술/건축, 방송/연예, 책, 종교, 학술/문화재
IT	IT비즈/정책, 모바일, PC/기기가전, 게임
경제	경제일반, 산업/통상, 금융/재테크, 자동차, 생활경제/소비자, 부동산, 기업소식, 경제인, 창업/취업
사회	사회일반, 법원/검찰, 보건/복지, 미디어, 사건/사고, 노동, 교육/입시, 젠더
과학/환경	과학/환경일반, 기후/날씨, 우주/항공, 환경/생태
스포츠	배구, 농구, 경주(경마・경정・경륜), 골프, 생활체육, 올림픽/아시안게임, 축구, 야구, 스포츠종합일반, 월드컵, 격투기
정치	정치일반, 국회/정당, 선거, 북한/한반도, 대통령실, 국방/외교

〈표 7〉 경향신문 장르 분류체계

대분류	중분류
라이프	생활, 시사, 문화, 비즈라이프, 오늘의 운세, 나침반, 멍냥
오피니언	사설, 여적, 기자메모, 칼럼, 만평, 독자마당
매거진L	여행, 건강/의학, 스타일, 생활, 헬스경향
사람	인사, 부고, 동정, 인물일반
기획/연재	-
특집/이슈	-
문화/과학	-
교육/키즈	-
별자리운세	-

〈표 8〉 중앙일보 장르 분류체계

대분류	중분류
오피니언	사설칼럼, 만평, hot poll, 리셋코리아, 영상
라이프	패션, 맛, 뷰티, 리빙, 건강
피플	사랑방, 인사, 부음
스페셜	hello! Parents, 팩플, 비크닉, 머니랩, 부동상, COOKING, 디지털 스페셜, 여행레저, 더 북한, 더 차이나, 더 마음, 더 하이엔드, 더 오래, 더 헬스

〈표 9〉 한국일보 장르 분류체계

대분류	중분류
사람	인터뷰, 인사/동정, 부고, 사람일반
라이프	음식, 여행, 건강, 블론디, 백운산 오늘의 운세, 라이프일반, *공연/전시
오피니언	사설, 칼럼, 만평, 사고알림

〈표 10〉 문화일보 장르 분류체계

대분류	중분류
*라이프	공연, 전시, 여행, 스타일/패션, 생활/요리, 의학/건강
오피니언	사설, 시론, 시평, 포럼, 뉴스와 시각, 오후여담, 문화논단, 기고, 살며생각하며, 여론마당
피플	인물일반, 국내인물, 국외인물, 인사, 동정, 부음
기획/시리즈	-

〈표 11〉 본 연구에서 구축한 뉴스 기사 장르 분류체계

대분류	중분류
인물	사람과, 인물일반, 부고, 동정, 인사
라이프	문화공연, 운세, 트래블, 스타일, 문화/과학, 건강/의학, 요리/맛집, 이슈/토픽, 생활일반
오피니언	오피니언 알림, 오피니언 일반(칼럼), 기고, 여적, 사설

언론사별 분류체계를 살펴본 결과, 대분류에 오피니언, 라이프, 사람(피플) 카테고리가 공통으로 존재하고 있었다. 이에 본 연구에서는 장르 대분류를 인물, 라이프, 오피니언 3개로 설정하여 각각의 대분류에 따른 중분류를 〈표 11〉과 같이 구성하였다. 각 대분류와 그에 따른 중분류에 대한 정의는 다음과 같다.

라이프는 건강, 여행, 음식, 패션 등 일상생활과 연관된 다양한 주제를 다루고 있으며 삶의 질을 향상하는 데 도움을 주는 정보를 제공하는 카테고리이다. 라이프의 중분류 카테고리는 문화공연, 운세, 트래블, 스타일, 문화/과학, 건강/의학, 요리/맛집, 이슈/토픽, 생활일반 카테고리로 구분했다. 문화공연 카테고리에는 공연, 전시, 연극/클래식 등의 다양한 공연 예술 관련 주제가 포함되었으며 스타일 카테고리에는 뷰티, 인테리어, 패션 등 라이프스타일과 관련된 다양한 주제를 포함하였다. 문화/과학 카테고리는 문화와 과학 분야에 대한 주제들이 포괄적으로 담겨있으며 문학적 작품에서 과학적 요

소를 다룬 경우 혹은 예술적 표현으로 자연 현상을 해석한 경우 등 문화와 과학이 교차하는 복합적인 주제도 포함하고 있다. 이슈/토픽 카테고리는 본래 라이프에 속하지 않고 개별적으로 존재하였으나 일상생활에 영향을 미치는 다양한 사회적, 문화적 요소들을 포함하는 주제의 기사들을 주로 담고 있어 라이프의 중분류 카테고리에 추가하였다.

오피니언은 사설, 칼럼, 논평 등 저자의 견해나 분석을 중심으로 하는 기사로, 독자에게 새로운 시각이나 비판적 사고를 제공하는 카테고리다. 오피니언의 중분류 카테고리는 기사 작성 특징에 따라 오피니언 일반(칼럼), 기고, 여적, 사설, 오피니언 알림 5개로 분류하였다. 카테고리별 뉴스 기사의 특징은 다음과 같다. 오피니언 일반(칼럼)에는 사건, 문화 정치 등 다양한 주제에 대해 개인적 견해나 전문적 분석을 담은 기사로 시론, 시평, 포럼 등이 해당한다. 주로 특정 전문가가 주기적으로 작성된 글이 많아 특정 개인의 기사 작성 스타일이 두드러지게 나타난다. 기고 카테고리는 외부 전문가, 일반인 등이 특정 이슈나 주제에 대해 의견을 서술한 것으로 일회성 글이 많으며 기사 작성 스타일도 다양하게 나타났다. 여적 카테고리는 비교적 짧은 글로 흥미롭거나 가벼운 일상과 관련된 다양한 주제를 다루며 대게 가볍고 친근한 어조로 작성된다. 사설 카테고리에는 사회적, 정치적 이슈 및 사건에 대한 언론사의 공적인 입장이나 견해를 반영한 기사가 존재한다. 마지막으로 오피니언 알림 카테고리에는 행사 및 공연 등에 대한 공지, 현재 진행 중인 중요한 사건/사고에 대한 긴급 속보, 뉴스 사이트 관련 공지, 독자 참여 요청 등 독자에게 필요한 정보나 공지를 전달하는 내용의 기사가 존재한다.

인물의 중분류 카테고리는 인물일반, 인사, 동정, 부고, 사람과 5개로 구분하였다. 인사 카테고리는 특정 조직 및 기업 내 인사의 이동, 승진, 임명 등과 관련된 주제를 다루고 있고, 동정 카테고리는 조직 및 기업 내 임원, 직원 혹은 조직 및 기업의 활동, 소식 등에 관한 기사가 해당한다. 부고 카테고리는 사망한 인물에 대한 공지, 사람과 카테고리에는 사회적 영향력이 큰 인물, 유명인 등의 사람에 관련된 인터뷰, 사건, 업적 등에 관련된 주제를 다루고 있다.

3.2.3 뉴스 기사 지역 분류체계

뉴스 기사 지역 분류 모델을 구성하기 위하여 지역 분류체계를 가지고 있는 언론사를 대상으로 지역 분류체계를 분석하였다. 분석에 사용된 언론사는 연합뉴스(연합뉴스, 2024), 경향신문(경향신문, 2024), 전국매일신문(전국매일신문, 2024), 국제뉴스(국제뉴스, 2024)가 있다.

〈표 12〉 국내 언론사별 지역 분류체계

언론사	지역 분류체계
연합뉴스	경기, 인천, 부산, 울산, 경남, 대구/경북, 광주/전남, 전북, 대전/충남/세종, 충북, 강원, 제주
경향신문	서울·수도권, 충청, 강원, 영남, 호남, 제주, 지역 일반
전국매일신문	서울, 경기, 인천, 대전, 충청, 세종, 부산, 경남, 대구, 경북, 광주, 호남, 강원, 제주
국제뉴스	서울, 경기남부, 경기북부, 인천, 강원, 대전, 충남, 충북, 세종, 전북, 광주, 전남, 경남, 대구, 경북, 울산, 부산, 제주

각 언론사의 지역 카테고리를 살펴본 결과, 언론사마다 지리적 구분의 범위가 상이함을 확인하였다. 이에 본 연구에서는 균형 잡힌 뉴스 기사 분류를 위해 대한민국의 행정구역을 기준으로 하여 〈표 13〉과 같이 총 18개의 카테고리로 지역 분류 체계를 작성하였다.

3.3 실험데이터 구축

본 연구에서 재구성한 뉴스 기사 분류체계에 따라 수집한 뉴스 기사 데이터를 전처리하였다. 그 결과, 뉴스 기사 데이터는 주제별 뉴스 기사 총 147,114건, 장르별 뉴스 기사 16,993건, 지역별 뉴스 기사 24,896건으로 정리되었다. 이후, 뉴스 기사 데이터를 주제별, 장르별, 지역별로 학습데이터와 평가데이터로 적절히 분할하여 모델의 학습과 성능 평가에 사용하였다.

3.3.1 주제 분류 모델 실험데이터

국제, 문화, IT, 경제, 사회, 과학/환경, 스포츠, 정치 총 8개의 주제 카테고리에 대하여 분류를 수행하는 대분류 모델 구축을 위하여 주제별 뉴스 기사 데이터 147,114건에서 각 대분류 카테고리별로 약 3,000건의 데이터를 추출하여 실험데이터로 활용하였다. 이에 학습데이터 19,702건, 평가데이터 4,926건을 구성하여 총 24,628건을 대분류 모델의 실험데이터로 활용하였다.

주제 중분류 모델의 실험데이터는 일부 카테고리에 대하여 상대적으로 적은 뉴스 기사가 수집된 결과, 카테고리별로 데이터 불균형이 존재한다. 이는 데이터 수집의 한계뿐만 아니라 일부 주제가 다른 주제에 비해 사람들의 관심이 상대적으로 적어 뉴스 보도의 빈도가 낮았기 때문으로 판단된다. 각 중분류 모델별 학습 및 평가에 사용한 실험데이터는 〈표 14〉와 같다.

3.3.2 장르 분류 모델 실험데이터

뉴스 기사 장르 분류 모델의 실험데이터는 총 16,993건이 구축되었다. 장르 분류 실험데이

〈표 13〉 본 연구에서 구축한 뉴스 기사 지역 분류체계

뉴스 기사 지역 분류체계
전국일반, 서울, 대전, 부산, 경기, 경남, 충남, 대구, 전남, 제주, 광주, 세종, 강원, 경북, 충북, 전북, 인천, 울산

〈표 14〉 주제 중분류 모델별 실험데이터

	학습데이터	평가데이터	실험데이터
사회	15,644	3,912	19,556
정치	18,540	4,636	23,176
문화	9,935	2,484	12,419
국제	21,285	5,322	26,607
스포츠	18,100	4,525	22,625
과학/환경	3,148	788	3,936
IT	4,107	1,027	5,134
경제	26,928	6,733	33,661

터 또한, 카테고리별 데이터 불균형이 존재하였다. 이를 해결하기 위하여 대분류 모델의 실험데이터 구성 시, 데이터 수가 가장 많은 카테고리의 데이터 수를 가장 적은 카테고리의 데이터 수보다 최대 3,000건 더 많도록 설정하였다. 이에 인물 카테고리에 3,438건, 라이프 카테고리에 5,171건, 오피니언 카테고리에 2,171건의 데이터를 구축하여 학습데이터 8,624건, 평가데이터 2,156건을 장르 대분류 모델 실험에 활용하였다.

장르의 각 중분류 모델별 학습 및 평가에 사용한 실험데이터는 〈표 15〉와 같다.

3.3.3 지역 분류 모델 실험데이터

지역 분류 모델은 대분류 모델 1개로 18개의 카테고리 분류를 수행한다. 지역 분류 모델의 실험데이터는 〈표 16〉과 같이 지역 카테고리별로 데이터 불균형이 존재한다. 실험데이터는 총 24,896건으로 학습데이터 19,916건, 평가데이터 4,980건으로 구성하여 실험에 활용하였다.

3.4 뉴스 기사 분류 모델 구축

본 연구에서는 뉴스 기사 분류 모델 구축을 위해 BERT 기반의 한국어 사전학습 언어모델을 미세조정하였다. BERT는 Google에서 제안한 트랜스포머 아키텍처 기반의 언어모델로 자연어 처리 작업에서 높은 성능을 기록한 바 있다(Devlin et al., 2019). 특히 BERT는 문맥에 따라 단어의 의미를 양방향으로 이해하는 능력이 뛰어나 텍스트 분류와 같은 자연어 처리 작업에 적합하여 뉴스 기사 분류에서도 사용할 수 있다. 이에 본 연구의 실험에서는 한국어 뉴

〈표 15〉 장르 중분류 모델별 실험데이터

	학습데이터	평가데이터	실험데이터
인물	2,750	688	3,438
라이프	9,107	2,277	11,384
오피니언	1,736	435	2,171

〈표 16〉 지역 분류 모델 실험데이터의 카테고리별 데이터 분포

카테고리	데이터 개수	카테고리	데이터 개수
전국일반	3,168	인천	1,145
서울	3,168	울산	942
경기	3,114	대전	882
제주	2,000	경북	784
부산	1,747	충남	653
대구	1,682	전남	547
경남	1,666	광주	336
강원	1,511	세종	208
충북	1,175	전북	168
실험데이터			24,896

스 기사 분류를 위해 대규모 한국어 데이터로 미세조정된 BERT 기반 한국어 사전학습 언어모델을 사용하였다. 사용한 언어모델은 다음과 같다.

KoBERT 모델은 SKTBrain에서 BERT base multilingual cased의 한국어 데이터 처리 성능의 한계를 극복하기 위해 개발하였다. 위키피디아, 뉴스 등에서 수집한 수백만 개의 한국어 문장으로 이루어진 대규모 말뭉치를 학습한 한국어에 특화된 모델이다(sktelecom, 2021).

KPF-BERT 모델은 한국언론진흥재단이 보유하고 있는 빅카인즈 기사 데이터 중 2000년부터 2021년 8월까지의 기사 약 4,000만 건을 학습한 모델로 언론사 및 뉴스 기사에 특화된 한국형 표준 뉴스 기사 인공지능 언어모델이다(한국언론진흥재단, 2022).

KLUEBERT 모델은 MODU, 나무위키, NEWSCRAWL, CC-100-kor, PETITION의 코퍼스를 사용하여 사전학습 시킨 BERT 기반 모델이다(Park Sungjoon et al., 2021).

3.5 실험 환경 및 평가 척도

실험에서 사용된 뉴스 기사 분류 모델은 AMD 라이젠 스레드리퍼 PRO 5975WX(샤갈 프로)와 NVIDIA RTX 4090 GPU 2개가 장착된 PC에서 NVIDIA RTX 4090 GPU 1개를 사용하여 구축되었다. 실험에 사용된 사전학습 언어모델의 파라미터는 〈표 17〉과 같다.

뉴스 기사 분류 모델의 성능 평가를 위해 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1-Score를 활용하였다. 각 지표에 대한 설명은 다음과 같다.

정확도(Accuracy)는 모델이 올바르게 예측한 데이터의 수를 전체 데이터의 수로 나눈 값이다. 이는 전체 데이터에서 모델이 정확하게 예측한 데이터의 비율을 의미한다. 정밀도(Precision)는 모델이 참이라고 예측한 데이터 중에 실제로 참인 데이터의 비율을 의미한다. 재현율(Recall)은 실제 참인 데이터 중에 모델이 참이라고 예측한 것의 비율을 의미한다. F1 Score는 정밀도와 재현율의 조화 평균으로 두 지표 간의 균

〈표 17〉 사전학습 언어모델별 학습 파라미터

	kykim/bert-kor-base	klue/bert-base	jinmang2/kpfbert
Input Size	512	512	512
Heads	12	12	12
Embedding Size	768	768	768
Hidden Size	768	768	768
Dropout	0.1	0.1	0.1
FFNN Size	3072	3072	3072
Layer Size	12	12	12
Optimizer	AdamW	AdamW	AdamW
Learning Rate	5e-5	5e-5	5e-5
Epsilon	1e-8	1e-8	1e-8
Vocab Size	42000	32000	36440

형을 평가하는 데 사용된다.

4. 분류 방법별 실험 및 평가

4.1 뉴스 기사 주제 분류 모델

계층적 구조의 뉴스 기사 주제 분류 모델을 구현하기 위하여 먼저, 대분류 모델과 중분류 모델들을 구축하고 성능 평가를 진행하였다. 〈표 18〉은 주제 대분류 모델의 성능을 평가한 결과이다. 'kykim/bert-kor-base' 모델이 가장 높은 성능을 도출하고 있으나 전반적으로 3가지 사전학습 언어모델 모두 95% 이상의 높은 성능을 보였다. 이는 주제 분류별로 사용되는 어휘가 다르기 때문으로 판단된다. 즉, 각 카테고리의 기사는 특정 주제와 관련된 용어와 표현이 자주 사용되며, 이러한 어휘적 특징이 분류 모델의 성능에 영향을 끼친 것으로 판단된다.

〈표 19〉는 주제 중분류 모델별 성능 평가 결과를 나타낸 것이다.

사회 중분류 모델은 8개의 사회 중분류 카테고리에 대하여 분류를 수행하였다. 'kykim/bert-kor-base' 모델이 Accuracy, F1 Score, Precision에서 가장 높은 성능을 보였고, Recall에서는 'klue/bert-base' 모델이 가장 높은 성능을 보였다. 그러나 전반적으로 사회 중분류 모델의 성능은 90%에 도달하지 못했다. 이는 카테고리의 개수가 많고, 각 카테고리 간의 모호성이 존재하는 것이 주된 원인으로 분석된다. 특히, 혼동 행렬을 분석한 결과, 사회일반 카테고리가 다른 모든 카테고리와 혼동을 일으켰고, 법원/검찰과 사건/사고 카테고리 간의 혼동도 잦은 것으로 관찰되었다. 이는 사회일반 카테고리는 일반적이고 폭넓은 정보를 담는 경향이 있어 다른 카테고리와 구별하기 어렵고, 법원/검찰과 사건/사고 카테고리가 공통으로 법적인 내용과 사건 중심적인 정보를 다루고 있어 혼동이 발생하는 것으로 보인다.

정치 중분류 모델은 정치 관련 주제를 6개의 세부 카테고리로 분류하는 작업을 수행하였다. Recall 지표는 'klue/bert-base' 모델이 89.65%로 가장 높은 성능을 기록하였고, 이외 지표에서는 'kykim/bert-kor-base' 모델의 성능이 가장 높았다. 그러나 정치 중분류 모델 또한 전반적으로 모델의 성능이 90% 미만인 것으로 나왔다. 이는 정치 일반 카테고리가 다른 모든 카테고리와 빈번하게 혼동을 일으켰고, 국방/외교와 북한/한반도 카테고리에서도 혼동이 잦아 나타난 결과로 보인다.

문화 중분류 모델은 6개의 문화 세부 카테고리에 대하여 분류를 수행하였다. 실험 결과, 'jinmang2/kpfbert' 모델에서 Accuracy와 F1 Score가 가장 높았다. Recall은 'klue/bert-base'에서 93.39%

〈표 18〉 주제 대분류 모델 성능 평가 결과

	acc	f1	recall	precision
jinmang2/kpfbert	96.589	96.585	96.586	96.587
klue/bert-base	96.528	96.520	96.525	96.522
kykim/bert-kor-base	96.731	96.723	96.728	96.729

〈표 19〉 주제 중분류 모델별 성능 평가 결과

	모델	acc	f1	recall	precision
사회	jinmang2/kpfbert	88.343	87.160	87.688	86.787
	klue/bert-base	88.420	87.901	88.968	87.040
	kykim/bert-kor-base	88.931	88.034	88.326	87.797
정치	jinmang2/kpfbert	87.878	88.764	88.961	88.586
	klue/bert-base	88.481	89.322	89.657	89.076
	kykim/bert-kor-base	88.740	89.643	89.598	89.726
문화	jinmang2/kpfbert	92.391	92.572	93.098	92.088
	klue/bert-base	92.230	92.351	93.398	91.407
	kykim/bert-kor-base	92.310	92.398	92.617	92.247
국제	jinmang2/kpfbert	91.055	91.435	91.653	91.285
	klue/bert-base	90.924	91.411	91.176	91.677
	kykim/bert-kor-base	91.112	91.648	91.458	91.869
스포츠	jinmang2/kpfbert	95.469	92.823	93.213	92.718
	klue/bert-base	95.513	92.391	92.831	92.635
	kykim/bert-kor-base	95.116	92.047	92.263	92.243
과학/환경	jinmang2/kpfbert	92.639	92.311	92.860	91.964
	klue/bert-base	92.639	91.625	91.933	91.481
	kykim/bert-kor-base	93.020	92.311	92.545	92.277
IT	jinmang2/kpfbert	97.370	95.714	95.055	96.554
	klue/bert-base	97.663	96.037	95.960	96.280
	kykim/bert-kor-base	96.689	94.606	93.932	95.418
경제	jinmang2/kpfbert	83.692	82.499	81.709	83.993
	klue/bert-base	83.959	83.018	82.509	83.745
	kykim/bert-kor-base	84.033	83.395	82.584	84.439

로 가장 높았으며, Precision은 92.24%로 'kykim/ber-kor-base'에서 가장 높은 성능을 보였다. 문화 중분류 모델은 모든 지표에 대해 비교적 균등한 성능을 보여줌으로써 각 세부 카테고리에 대해 일관성 있는 분류 능력을 갖춘 것으로 판단된다.

국제 중분류 모델은 국제 관련 주제를 7개의 세부 카테고리로 분류하는 작업을 수행하였다. 'jinmang2/kpfbert' 모델에서 Recall 지표가 91.65%로 가장 높았으며 그 외 지표에서는 'kykim/bert-kor-base' 모델이 미세한 차이로 가장 높은 성능을 보였다. 국제 중분류 모델 또한, 모든 모델에서 모든 지표에 대하여 균등한 성능을 보여 세부 카테고리에 대하여 일관성 있는 분류 능력을 갖추었다고 판단된다.

스포츠 중분류 모델은 11개의 세부 카테고리에 대하여 분류를 수행하였다. 그 결과, 'klue/bert-base' 모델에서 Accuracy가 가장 높았으며, 그 외 지표에서는 'jinmang2/kpfbert'가 높은 성능을 기록했다. 스포츠 중분류 모델의 경우, Accuracy에 비해 F1 Scocre, Recall, Precision은 약 3% 정도 낮았다. 이러한 결과는 클래스 불균형 문제에 기인한 것으로 스포츠의 일부 카테고리의 데이터 개수가 상대적으로 적어 해

당 카테고리의 학습이 충분히 이루어지지 않은 것으로 판단된다.

과학/환경 중분류 모델은 총 4개의 세부 카테고리에 대하여 분류를 수행하였다. 분류 결과, 'kykim/bert-kor-base' 모델에서 Accuracy가 93.02%로 가장 높았다. F1 Score는 'jinmang2/kpfbert'와 'kykim/bert-kor-base' 모델에서 92.31%로 가장 높게 나타났으며, Recall과 Precision은 'jinmang2/kpfbert'에서 가장 높은 성능을 보였다.

IT 중분류 모델은 총 4개의 세부 카테고리에 대해 분류를 수행하였다. 분류 결과, Precision은 'jinmang2/kpfbert' 모델에서 96.55%로 가장 높은 성능을 기록했으며, 그 외 지표는 'klue/bert-base' 모델이 가장 높은 성능을 보였다. IT 중분류 모델의 경우, 모든 지표에서 평균적으로 약 95.94% 정도의 성능을 보였다. 이는 IT의 세부 카테고리 간의 뉴스 기사 특징이 명확하여, 분류 모델이 카테고리를 식별하는 데 있어 혼동이 적게 발생했다는 것을 시사한다.

경제 중분류 모델은 총 9개의 세부 카테고리에 대하여 분류를 수행하였다. 'kykim/bert-kor-base' 모델이 모든 지표에서 가장 우수한 성능을 보였다. 경제 중분류 모델의 성능은 80% 초반에 그치고 있다. 이는 경제의 세부 카테고리 간 모호성이 존재하여 발생한 결과로 판단된다. 예를 들어, 기업소식 카테고리 내 존재하는 대기업의 신규 투자 소식 관련 뉴스 기사의 경우, 금융적 관점에서는 금융/재테크 카테고리에 분류될 수 있지만 해당 투자가 산업에 미치는 영향을 고려할 때는 산업/통상으로도 분류될 수 있다. 이처럼 경제 중분류 모델은 세부 카테고리 간의 유사성이 다수 존재하여 모델이 정확한 분류를 수행하는 데 어려움이 있는 것으로 판단된다.

4.2 뉴스 기사 장르 분류 모델

계층적 구조의 뉴스 기사 장르 분류 모델의 구현을 위해 장르 대분류 모델과 장르 중분류 모델들을 구축하고 평가하였다. 장르 대분류 모델은 모든 지표에서 평균 약 97.49%의 우수한 성능을 보이고 있으며, 'jinmang2/kpfbert' 모델이 미세한 차이로 모든 지표에서 가장 높은 성능을 기록했다. 이러한 결과는 모델이 분류를 수행해야 할 카테고리의 개수가 적고, 카테고리 간의 범주가 명확하고 간단하여 나타나는 것으로 보인다.

〈표 21〉은 장르의 중분류 모델별 성능 평가 결과를 나타낸 것이다.

인물 중분류 모델은 총 5개의 세부 카테고리에 대하여 분류를 수행하였다. 그 결과, 모든 지표에서 'kykim/bert-kor-base' 모델이 가장 높은 성능을 보였다. 인물 중분류 모델은 Accuracy에서

〈표 20〉 장르 대분류 모델 성능 평가 결과

	acc	f1	recall	precision
jinmang2/kpfbert	97.912	97.736	97.661	97.813
klue/bert-base	97.263	97.107	97.105	97.118
kykim/bert-kor-base	97.727	97.480	97.326	97.645

〈표 21〉 장르 중분류 모델별 성능 평가 결과

	모델	acc	f1	recall	precision
인물	jinmang2/kpfbert	97.383	93.488	92.233	95.048
	klue/bert-base	96.947	91.513	91.125	91.942
	kykim/bert-kor-base	98.110	94.919	93.528	96.498
오피니언	jinmang2/kpfbert	97.241	93.586	90.705	97.859
	klue/bert-base	96.551	94.755	93.701	96.126
	kykim/bert-kor-base	96.551	89.223	87.643	91.842
라이프	jinmang2/kpfbert	94.861	93.315	94.200	92.502
	klue/bert-base	93.983	92.406	92.850	92.053
	kykim/bert-kor-base	94.554	92.959	92.832	93.116

평균 약 97% 정도의 높은 성능을 보이고 있으나 그 외 지표에서는 Accuracy 대비 약 3~5% 정도 낮은 성능을 보였다. 이는 인물 중분류 카테고리의 인물일반, 동정, 사람과 카테고리의 데이터가 개수가 부고, 인사 카테고리에 비해 상대적으로 적어 충분한 학습이 이루어지지 않은 것을 보인다.

오피니언 중분류 모델은 총 5개의 세부 카테고리에 대하여 분류를 수행하였다. Accuracy와 Precision은 'jinmang2/kpfbert' 모델의 성능이 가장 높았고, F1 Score와 Recall은 'klue/bert-base' 모델의 성능이 가장 우수했다. 오피니언 중분류 모델 또한, 카테고리 간의 데이터 불균형이 존재하여 Accuracy와 F1 Score, Recall 간의 점수 차이가 발생한 것으로 판단된다.

라이프 중분류 모델은 총 9개의 세부 카테고리에 대하여 분류를 수행하였다. 실험 결과, 모든 지표에 대하여 'jinmang2/kpfbert' 모델의 성능이 가장 우수하였다. 라이프 중분류 모델도 카테고리 간의 데이터 불균형이 존재하나 인물, 오피니언 중분류 모델처럼 지표 간의 점수 차이가 크지 않다. 이는 라이프의 세부 카테고리 간의 경계가 상대적으로 명확하여 모델이 일관된 성능을 유지할 수 있던 것으로 판단된다.

4.3 뉴스 기사 지역 분류 모델

18개의 지역 카테고리에 대하여 분류를 수행하는 지역 분류 모델을 구축하였다. 실험 결과, 모든 지표에서 평균 약 95% 정도의 성능을 보였으며 그중에서 'klue/bert-base' 모델이 모든 지표에서 가장 높은 성능을 기록하였다.

지역 분류 모델의 경우, 카테고리의 개수가 많고, 카테고리 간 데이터 불균형이 존재했으

〈표 22〉 뉴스 기사 지역 분류 모델 성능 평가 결과

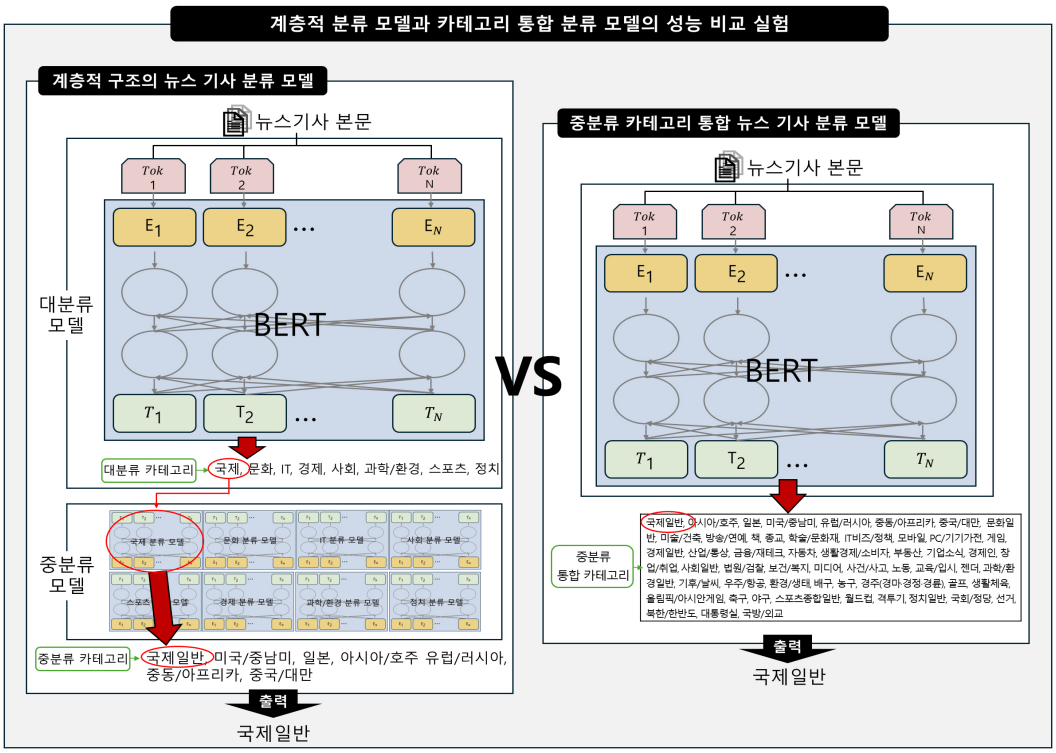
	acc	f1	recall	precision
jinmang2/kpfbert	95.542	95.364	95.875	94.886
klue/bert-base	95.823	96.113	96.602	95.680
kykim/bert-kor-base	95.662	95.658	96.068	95.295

나 모델의 성능은 약 95%로 전반적으로 매우 높게 나타났다. 이는 지역 기사의 내용에 춘천시, 대전시 등 지역의 이름이나 관련 용어가 명확하게 언급되어 있을 뿐만 아니라, 특정 지역 관련 행정 기관의 활동 및 정책, 지역별 특정 행사 및 프로젝트 등의 지역적 특색이 반영되어 있어 모델이 효율적으로 지역 기사를 식별한 것으로 판단된다.

5. 계층적 구조의 뉴스 기사 분류 모델의 성능 평가

본 절에서는 계층적 구조로 이루어진 주제 및 장르 뉴스 기사 분류 모델의 분류 정확도와 효율성을 평가하고자 한다. 이를 위해 앞서 주제 및 장르 분류 모델 실험에서 구축한 대분류와 중분류 모델을 연결하여 계층적 구조의 분류 모델을 구성하고, 이를 단일 모델로 모든 대분류의 중분류 카테고리를 분류하는 모델과 비교하고자 한다.

본 연구에서는 〈그림 2〉와 같이 뉴스 기사를 먼저 대분류 모델에 입력하여 상위 카테고리로 분류한 후, 해당 대분류 결과를 바탕으로 중분류 모델을 사용하여 세부 카테고리로 분류하는 모델을 계층적 구조의 뉴스 기사 분류 모델(이하 계층적 주제/장르 분류 모델)로 지칭하였다. 모든 대분류의 중분류 카테고리를 하나의 모델



〈그림 2〉 계층적 분류 모델과 카테고리 통합 분류 모델의 성능 비교 실험

로 처리하여 뉴스 기사를 중분류 수준에서 한 번에 분류하는 모델은 중분류 카테고리 통합 뉴스 기사 분류 모델(이하 주제/장르 카테고리 통합 분류 모델)로 지칭하였다.

5.1 계층적 구조의 뉴스 기사 주제 분류 모델 성능 평가 비교 실험

5.1.1 계층적 구조의 뉴스 기사 주제 분류 모델 구현

계층적 구조의 뉴스 기사 주제 분류 모델은 대분류 모델과 8개의 중분류 모델을 연결하여 구현하였다. 주제 대분류 모델은 성능이 가장 우수한 것으로 평가된 'kykim/bert-kor-base' 모델을 활용하였다. 주제 중분류 모델은 대분류 카테고리별로 Accuracy와 F1 Score의 평균이 가장 높은 모델을 선택하였다. 이에 사회, 정치, 국제, 과학/환경, 경제 중분류 모델에는 'kykim/bert-kor-base', 문화와 스포츠 중분류 모델에는 'jinmang2/kpfbert', IT 중분류 모델에는 'klue/bert-base' 모델을 선택하였다.

5.1.2 주제 중분류 카테고리 통합 뉴스 기사 분류 모델 구축

주제 중분류 카테고리 통합 뉴스 기사 분류 모델은 뉴스 기사 주제 분류체계에 포함된 총 55개의 주제 중분류 카테고리에 대하여 분류 작업을 수행하는 모델이다. 주제 중분류 카테고리 통합 뉴스 기사 분류 모델 구축을 위한 실험데이터는 총 120,724건으로 학습데이터로 96,579건, 평가데이터로 24,145건을 사용하였다.

모델 학습 및 평가 결과는 〈표 23〉과 같다. 모든 모델의 평가 지표 결과 평균이 86% 이상으로 나타났다. Accuracy와 F1 Score, Recall은 'jinmang2/kpfbert' 모델에서 가장 높은 성능을 보였고, Precision은 'klue/bert-base' 모델에서 가장 높았다.

비교 실험에 사용할 주제 중분류 카테고리 통합 뉴스 기사 분류 모델은 Accuracy와 F1 Score의 성능이 가장 우수했던 'jinmang2/kpfbert' 모델을 선택하였다.

5.1.3 비교 실험을 위한 평가데이터 구축

계층적 구조의 뉴스 기사 주제 분류 모델과 주제 중분류 카테고리 통합 뉴스 기사 분류 모델 간의 성능 평가 비교를 위하여 평가데이터는 다음과 같이 구축하였다. 주제 중분류 모델별 평가데이터와 중분류 카테고리 통합 뉴스 기사 분류 모델의 평가데이터에서 공통으로 사용된 데이터를 기반으로 레이블별 최대 100건의 기사 데이터를 구성하였다. 그러나 일부 레이블에서 데이터 불균형이 존재하여 최종적으로 5,092건의 평가데이터가 구축되었다.

〈표 23〉 주제 중분류 카테고리 통합 뉴스 기사 분류 모델 성능 평가 결과

	acc	f1	recall	precision
jinmang2/kpfbert	86.465	86.007	86.597	85.747
klue/bert-base	86.316	85.940	86.029	86.068
kykim/bert-kor-base	86.394	85.863	85.930	85.979

5.1.4 실험 결과

계층적 구조의 뉴스 기사 주제 분류 모델과 주제 중분류 카테고리 통합 뉴스 기사 분류 모델을 평가한 결과는 〈표 24〉와 같다. 계층적 구조의 뉴스 기사 주제 분류 모델은 Accuracy 80%, F1 Score 79.50%, Recall 78.82%, Precision 81.87%를 기록하였다. 반면, 주제 중분류 카테고리 통합 뉴스 기사 분류 모델은 Accuracy 95.81%, F1 Score 95.74%, Recall 95.67%, Precision 95.90%를 기록하여 계층적 구조의 뉴스 기사 분류 모델보다 우수한 성능을 보였다. 이러한 결과는 계층적 구조의 뉴스 기사 주제 분류 모델이 대분류 결과에 따라 중분류가 진행되는 방식으로 이루어져 초기 대분류 단계의 오류가 전체적인 성능에 부정적인 영향을 끼쳤으나 주제 중분류 카테고리 통합 뉴스 기사 분류 모델의 경우, 단일 모델을 사용하여 모든 카테고리를 동시에 분류함으로써 이러한 종속적 오류의 영향을 받지 않아 나타난 것으로 보인다.

5.2 계층적 구조의 뉴스 기사 장르 분류 모델 성능 평가 비교 실험

5.2.1 계층적 구조의 뉴스 기사 장르 분류 모델

계층적 구조의 뉴스 기사 장르 분류 모델은 대분류 모델과 3개의 중분류 모델을 연결하여 구현하였다. 장르 대분류 모델은 성능이 가장 우수한 것으로 평가된 'jinmang2/kpfbert' 모델을 선택하였다. 장르 중분류 모델은 대분류 카테고리별로 Accuracy와 F1 Score가 가장 높은 모델을 선택하였다. 이에 인물 중분류 모델은 'kykim/bert-kor-base', 오피니언 중분류 모델은 'klue/bert-base', 라이프 중분류 모델은 'jinmang2/kpfbert' 모델을 선택하였다.

5.2.2 장르 중분류 카테고리 통합 뉴스 기사 분류 모델 구축

장르 중분류 카테고리 통합 뉴스 기사 분류 모델은 뉴스 기사 장르 분류체계에 포함된 19개의 중분류 카테고리 전체를 대상으로 분류를 수행하는 모델이다. 장르 중분류 카테고리 통합 뉴스 기사 분류 모델 구축을 위한 실험데이터는 총 16,617건으로 13,293건의 데이터로 학습하고, 3,324건의 데이터로 평가한 결과는 〈표 25〉와 같다. 'kykim/bert-kor-base' 모델에서 Accuracy와 precision이 가장 높았고, 'jinmang2/kpfbert' 모델에서 F1 Score와 Recall 점수가 가장 높았다.

장르 중분류 카테고리 통합 뉴스 기사 분류 모델의 경우, 모든 모델에서 Accuracy와 다른 지표들 간 약 ±3%의 점수 차이가 관찰되었다. 이는 카테고리별 데이터의 불균형이 모델 성능에 큰 영향을 미친 것으로 분석된다. 장르 뉴스 기사 데이터의 경우, 카테고리 간 데이터 개수

〈표 24〉 계층적 주제 분류 모델과 주제 카테고리 통합 분류 모델 간의 성능 평가 결과 비교

	acc	f1	recall	precision
계층적 구조의 뉴스 기사 주제 분류 모델	80.007	79.508	78.824	81.878
주제 중분류 카테고리 통합 뉴스 기사 분류 모델	95.816	95.744	95.672	95.909

〈표 25〉 장르 중분류 카테고리 통합 뉴스 기사 분류 모델 성능 평가 결과

	acc	f1	recall	precision
jinmang2/kpfbert	93.561	90.453	91.115	89.979
klue/bert-base	93.080	90.233	90.773	89.890
kykim/bert-kor-base	93.712	90.335	90.81	90.193

의 격차가 컸다. 이러한 상황에서 모델이 분류해야 할 카테고리 수가 많아지면서 학습 과정에서 소수 카테고리에 대한 노출이 충분하지 않아 소수 카테고리에 대한 예측 성능이 상대적으로 낮아지는 경향을 보였다. 이로 인해 Accuracy는 비교적 높게 유지되었으나 F1 Score, Recall, Precision은 상대적으로 점수가 낮게 나타난 것으로 판단된다.

계층적 구조의 뉴스 기사 장르 분류 모델과의 비교 분석을 위해 장르 중분류 카테고리 통합 뉴스 기사 분류 모델은 Accuracy와 F1 score의 평균 성능이 가장 우수한 'kykim/bert-kor-base' 모델을 사용하였다.

5.2.3 비교 실험을 위한 평가데이터 구축

계층적 구조의 뉴스 기사 장르 분류 모델과 장르 중분류 카테고리 통합 뉴스 기사 분류 모델 간의 성능 평가 비교를 위하여 평가데이터는 다음과 같이 구축하였다. 장르 중분류 모델별 평가데이터와 중분류 카테고리 통합 뉴스 기사 분류 모델의 평가데이터에서 공통으로 사용된 데이터를 추출하여 836건의 평가데이터를 구축하였다.

데이터 불균형 문제를 해소하고자 카테고리별 최대 100건의 데이터를 추출하고자 했으나 데이터 부족으로 인해 카테고리 간 데이터 불균형이 존재한다.

5.2.4 실험 결과

계층적 구조의 뉴스 기사 장르 분류 모델과 장르 중분류 카테고리 통합 뉴스 기사 분류 모델을 평가한 결과는 〈표 26〉과 같다. 주제 분류 모델의 결과와 동일하게 장르 중분류 카테고리 통합 뉴스 기사 분류 모델이 모든 평가 지표에서 계층적 구조의 뉴스 기사 장르 분류 모델보다 높은 성능을 나타냈다. 다만, 뉴스 기사 장르 분류 모델이 주제 분류 모델보다 비교적 높은 성능을 보였는데, 이는 뉴스 기사의 내용이 주제에 비해 장르 분야가 다른 카테고리로 분류될 수 있는 모호성이 적은 데이터이며 분류 카테고리의 개수도 적기 때문에 나타난 것으로 판단된다.

6. 결 론

본 연구에서는 사전 학습된 한국어 BERT

〈표 26〉 계층적 장르 분류 모델과 장르 카테고리 통합 분류 모델 간의 성능 평가 결과 비교

	acc	f1	recall	precision
계층적 구조의 뉴스 기사 장르 분류 모델	94.258	91.854	92.041	92.589
장르 중분류 카테고리 통합 뉴스 기사 분류 모델	99.282	98.909	98.810	99.067

모델을 활용하여 뉴스 기사를 주제별, 장르별, 지역별로 자동 분류하는 모델을 구축하고, 성능을 평가하였다. 이를 위하여 국내 언론사의 뉴스 기사 분류체계를 참고하여 주제, 장르, 지역별로 세분화한 새로운 뉴스 기사 분류체계를 작성하였다. 주제 및 장르 분류에서는 대분류와 중분류 모델을 연결하여 계층적 구조의 분류 모델을 구현하고 카테고리 통합 분류 모델과 성능을 비교하였다. 평가 결과, 카테고리 통합 분류 모델이 계층적 구조의 뉴스 기사 분류 모델보다 더 높은 성능을 기록하였다. 이는 단계적 과정을 거치지 않는 카테고리 통합 모델과 달리, 계층적 구조의 분류 모델은 초기 대분류 단계의 분류 결과가 이후 중분류 모델의 선택에 영향을 주기 때문에 발생한 것으로 판단된다. 카테고리 통합 모델은 본 연구에서 95% 이상의 높은 성능을 보여 뉴스 기사의 주제 및 장르를 계층적 구조의 분류 모델보다 정확하게 분류할 수 있음을 보였다. 반면에 본 연구에서 제안한 계층적 구조의 뉴스 기사 분류 모델은 대분류에서 중분류로 단계적으로 나아가며 분류를 진행하기 때문에 모호하거나 중복된 카테고리에 속한 기사들을 카테고리 통합 모델보다 명확하게 분류할 수 있음을 보였다.

지역 분류 모델의 경우, 카테고리의 개수가 많고, 데이터 불균형이 존재하였으나 모든 모델에서 높은 성능을 보였다. 이러한 결과는 지역별 뉴스 기사의 분포와 특성이 상대적으로 명확하게 구분되기 때문으로 해석된다.

본 연구는 국내 언론사의 뉴스 기사 분류체계를 참고하여 새롭게 재구성된 분류체계를 제안함으로써 뉴스 기사의 자동 분류에 적합한 보다 일반화된 체계를 제공하였다는 점에서 의의가 있다. 또한, 주제, 장르, 지역을 포괄하는 다각적인 뉴스 기사 분류체계를 제안함으로써, 사용자의 요구에 부합하는 다차원적 뉴스 기사 분류 서비스의 가능성을 제시한 점에서 중요한 의미를 가진다.

향후 연구에서는 본 연구에서 나타난 데이터 불균형 문제를 해결하기 위한 방안을 모색하고, 뉴스 기사의 지속적인 변화와 다양성을 반영할 수 있는 정교하고 사용자 친화적인 뉴스 기사 분류 시스템 구축을 목표로 해야 한다.

참 고 문 헌

강승태, 장길진 (2023). ChatGPT와 다국어 BERT를 이용한 코로나-19 감염병 다국어 기사 자동 색인 및 분류. 전자공학회논문지, 60(7), 20-29. https://doi.org/10.5573/ieie.2023.60.7.20

경향신문 (2024. 7. 5.). 경향신문 홈페이지 뉴스 기사 카테고리. 출처: https://www.khan.co.kr/

국제뉴스 (2024. 7. 5.). 국제뉴스 홈페이지 뉴스 기사 카테고리. 출처: https://www.gukjenews.com/

김덕기, 온병원 (2023). 한국어 뉴스 기사 분류를 위한 KoBERT 기반의 문맥 벡터 클러스터링 방안 성능 비교 연구. 한국정보기술학회 2023년도 하계종합학술대회 및 대학생논문경진대회, 499-503.

김미선 (2022). 핵심 키워드 추출 기반의 토픽 모델링을 통한 신문기사 분류모델 제안: 한국 농업 신문 기사 데이터를 중심으로. 석사학위논문, 충북대학교 경영정보학과(원).

김혜영 (2015). 토픽 모델링을 활용한 한국어 신문의 주제별 자동분류. 석사학위논문, 고려대학교 언어학과. https://doi.org/10.23186/korea.000000059980.11009.0000915

뉴시스 (2024. 7. 5.). 뉴시스 홈페이지 뉴스 기사 카테고리. 출처: https://www.newsis.com/

문화일보 (2024. 7. 5.). 문화일보 홈페이지 뉴스 기사 카테고리. 출처: https://www.munhwa.com/

성나영, 구명완 (2018). WPM 기반 MemN2N(End-to-End Memory Networks)을 이용한 한국어 뉴스 내용 분류에 관한 연구. 한국정보과학회 2018 한국컴퓨터종합학술대회 논문집, 994-996.

연합뉴스 (2024. 7. 5.). 연합뉴스 홈페이지 뉴스 기사 카테고리. 출처: https://www.yna.co.kr/

이재욱, 고병규, 김판구 (2016). 상호 정보량과 로그 정규화를 이용한 뉴스 카테고리 분류. 한국정보기술학회논문지, 14(7), 79-85. https://doi.org/10.14801/jkiit.2016.14.7.79

장지형, 홍참길 (2022). 북한 및 통일 관련 뉴스 기사의 자동 주제 분류. 한국정보과학회 2022 한국컴퓨터종합학술대회 논문집, 2126-2128.

전국매일신문 (2024. 7. 5.). 전국매일신문 홈페이지 뉴스 기사 카테고리.
출처: https://www.jeonmae.co.kr/

중앙일보 (2024. 7. 5.). 중앙일보 홈페이지 뉴스 기사 카테고리. 출처: https://www.joongang.co.kr/

한국언론진흥재단 (2022. 2. 23.) 빅카인즈 기사 기반 AI 언어모델 'KPF-BERT' 공개.
출처: https://kpf.or.kr/front/board/boardContentsView.do?board_id=246&contents_id=0efb18236cbe482293f4b366a251f676&link_g_topmenu_id=ccd1e88d6d7345cca51f20ce9f56d652&link_g_submenu_id=8f52dfc509b34e90aa799cf2d8204223&link_g_homepage=F

한국일보 (2024. 7. 5.). 한국일보 홈페이지 뉴스 기사 카테고리.
출처: https://www.hankookilbo.com/

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805v2. https://doi.org/10.48550/arXiv.1810.04805

IPTC (2024. 7. 5.). NewsCodes mediatopic. arts, culture, entertainment and media. Available: https://cv.iptc.org/newscodes/mediatopic/01000000

Park Sungjoon, Moon Jihyung, Kim Sungdong, Cho Won Ik, Han Jiyoon, Park Jangwon, Song Chisung, Kim Junseong, Song Youngsook, Oh Taehwan, Lee Joohong, Oh Juhyun, Lyu Sungwon, Jeong Younghoon, Lee Inkwon, Seo Sangwoo, Lee Dongjun, Kim Hyunwoo, Lee Myeonghwa, Jang Seongbo, Do Seungwon, Kim Sunkyoung, Lim Kyungtae, Lee Jongwon, Park Kyumin, Shin Jamin, Kim Seonghyun, Park Lucy, Oh Alice, Ha Jung-Woo, & Cho Kyunghyun (2021). Klue: Korean language understanding evaluation. arXiv preprint

arXiv:2105.09680. https://doi.org/10.48550/arXiv.2105.09680
sktelecom (2021. 2. 14.). KoBERT. 출처: https://sktelecom.github.io/project/kobert/

• **국문 참고자료의 영어 표기**
(English translation / romanization of references originally written in Korean)

Gukjenews (2024. 7. 5.). Gukjenews homepage news article category.
Available: https://www.gukjenews.com/

Hankookilbo (2024. 7. 5.). Hankookilbo homepage news article category.
Available: https://www.hankookilbo.com/

Jang, Jihyoung & Hong, Charmgil (2021). Automated topic classification of the news articles on North Korea and korean unification. The Korean Institute of Information Scientists and Engineers Proceedings of the 2022 Korea Computer Congress, 2126-2128.

Jeonkuk Maeil Shinmun (2024. 7. 5.). Jeonkuk Maeil Shinmun homepage news article category.
Available: https://www.jeonmae.co.kr/

JoongAng (20204. 7. 5.). JoongAng homepage news article category.
Available: https://www.joongang.co.kr/

Kang, Seungtae & Jang, Gil Jin (2023). COVID-19 multilingual news article auto-indexing and classification using ChatGPT and multilingual BERT. Journal of the Institute of Electronics and Information Engineers, 60(7), 20-29. https://doi.org/10.5573/ieie.2023.60.7.20

Kim, Deok Gi & On, Byung Won (2023). A comparative study on KoBERT-based context Vector clustering methods for Korean news article classification. The Proceedings of the 2023 KIIT Summer Conference, 499-503.

Kim, Heyoung (2015). Automatic Classification of Korean Newspapers by Topic Using Topic Modeling. Master's thesis, Korea University, linguistics.
https://doi.org/10.23186/korea.000000059980.11009.0000915

Kim, Mi Sun (2022). Newspaper Article Classification Model Based on Core Keyword Extraction: Using The Korea Agricultural Newspaper Article. Master's thesis, Chung Buchk National University, Management Information System

Korea Press Foundation (2024. 7. 5.). AI language model 'KPF-BERT' released based on Big Kinds article.
Available: https://kpf.or.kr/front/board/boardContentsView.do?board_id=246&contents_id=0efb18236cbe482293f4b366a251f676&link_g_topmenu_id=ccd1e88d6d7345cca51f20ce9f

56d652&link_g_submenu_id=8f52dfc509b34e90aa799cf2d8204223&link_g_homepage=F

Lee, Jae Uk, Ko, Byeong-Kyou, & Kim, Pan Koo (2016). News category classification using mutual information and log normalization. The Journal of Korean Institute of Information Technology, 14(7), 79-85. https://doi.org/10.14801/jkiit.2016.14.7.79

Newsis (2024. 7. 5.). Newsis homepage news article category. Available: https://www.newsis.com/

Sung, Na Young & Koo, Myoung Wan (2018). A study on category classification of Korean news by a MemN2N(End-to-End Memory Networks) based on WPM. The Korean Institute of Information Scientists and Engineers Proceeding of the 2018 Korea Computer Congress.

The Kyunghyang Shinmun (20204. 7. 5.). Kyunghyang Shinmun homepage news article category. Available: https://www.khan.co.kr/

The Munhwa Ilbo (20204. 7. 5.). The Munhwa Ilbo homepage news article category. Available: https://www.munhwa.com/

Yonhap News Agency (20204. 7. 5.). Yonhap News Agency homepage news article category. Available: https://www.yna.co.kr/