박사학위청구논문
2023학년도

군의 인터넷 언론대응 및 관리를 위한 방위사업 관련 문서분류 모델 개발

Development of a document classification model related to defense projects for the military's response and management of the Internet media

광운대학교 대학원
방위사업학과
장 상 훈

군의 인터넷 언론대응 및 관리를 위한 방위사업 관련 문서분류 모델 개발

Development of a document classification model related to defense projects for the military's response and management of the Internet media

광운대학교 대학원
방위사업학과
장 상 훈

군의 인터넷 언론대응 및 관리를 위한 방위사업 관련 문서분류 모델 개발

Development of a document classification model related to defense projects for the military's response and management of the Internet media

지도교수 이 석 준

이 논문을 국방경영학 박사학위 청구논문으로 제출함.

2023년 6월

광운대학교 대학원
방위사업학과
장 상 훈

장상훈의 국방경영학 박사학위논문을 인준함

심사위원장 _____ 인

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

광운대학교 대학원

2023년 6월

국문 요약

최근 증가한 인터넷 신문의 보급과 자동추천 알고리즘의 발달은 인터넷 사용자가 대상에 대한 판단을 할 때 실체와 다르게 볼 뿐만 아니라 이를 정정할 가능성을 차단한다. 군의 이미지 역시 대부분 인터넷을 통해 형성되고 있기 때문에 군의 이미지 관리를 위해서는 인터넷 신문과 사용자에 대해 적절하게 대응하는 것이 중요하다. 기존의 종이 신문에 비하여 인터넷 신문이 가지는 특징 중 하나는 높은 생산성과 파급속도인데, 이 때문에 부정적인 기사가 생산되면 급속히 퍼져나가고, 부정적인 이미지가 형성되고 나면 회복이 어렵다. 따라서 이미지를 관리하기 위해서 군은 인터넷 언론 중 대응 및 관리가 필요한 기사를 신속하게 분류하고 관련 부서에 전달하여 적절하게 대응하는 것이 필요하다.

선행연구 검토 결과, 군에서는 긍정적인 이미지를 형성하기 위한 이미지 관리가 조직의 사기와 정책적 의사결정의 추진력을 위해서 중요하다는 점을 인식하고 있었으며, 이미지의 관리란 인터넷 언론과 사용자의 증가가 이루어지는 오늘날, 제3자가 인터넷 정보를 통해 만들어지는 이미지인 디지털 정체성의 관리를 의미하였다. 또한 부정적인 이미지가 형성되지 않도록 디지털 정체성을 관리하기 위해서는 긍정적이거나 중립적인 성향의 기사보다는 부정적인 성향의 기사를 비교적 더 주목해야 하는데 이를 위해서는 기사, 즉 자연어가 가지고 있는 긍정, 중립, 부정의 태도를 분석하는 감성분석 기법 중 하나인 극성분석이 필요하였다. 감성분석과 극성분석에 대한 연구는 트위터나 영화리뷰 등 일반적인 문장에 대한 연구는 활발하게 이루어졌지만, 군과 방위사업과 관련한 글, 특히 기사 분서에 관

한 연구는 전무 한 수준이었다. 또한, 분류된 기사들은 대응부서로 분류하기 위해서는 각 대응부서별 수행하는 업무 등을 기준으로 문서를 분류해주는 문서분류 기법이 필요하였다. 문서분류에 대한 연구 역시 다양한 분야에서 연구되었지만, 군과 방위사업에 관한 데이터를 활용한 연구는 없었으며, 극성분석과 문서분류 모델을 동시적으로 분석한 연구 역시 확인되지 않았다. 또한 대부분의 연구 역시 분류 정확도를 높이는 것에 연구의 중점을 두었기 때문에 실무에서 관리하는 데이터의 적절성을 검토한 내용은 부족하였다.

따라서 본 연구에서는 방위사업과 관련된 디지털 정체성을 관리하기 위해 인터넷 기사 중 어떠한 내용의 기사들을 관리하여야 하는지에 대한 분류 및 기사 대응에 적합한 부서를 분류하기 위한 모델을 개발하고자 한다. 이를 위해 실제 방위사업청 대변인실에서 담당자가 수집하여 관리하고 있는 '16년~'22년 방위사업청 관련기사 2,153건을 활용하였다. 그리고 이를 원본 데이터, 원본 데이터를 딥러닝에 적합하게 정제한 데이터, 정제한 데이터의 수량을 증강한 증강 정제 데이터, 정제한 데이터를 다시 분류한 재라벨 데이터로 나누어 실험을 진행하였다. 그 결과 방위사업과 관련된 기사 중 27.6%가 보다 관리에 중점을 두어야 하는 부정적인 성격을 가지고 있는 것으로 나타났다. 또한, 분류 정확도에 있어서는 원본 데이터가 50.1%, 정제된 데이터가 60.7%, 증강 정제 데이터가 60.3%, 재라벨 데이터가 70.7%로 나타났다.

실증연구를 통해, 인공지능 학습에 적합한 데이터로 정제하고 라벨링하여 관리하는 것이 문서분류 모델의 성능향상에 도움이 된다는 것과 현재 실무에서 관리하고 있는 데이터 관리방법은 일부 개선이 필요하다는 결론을 도출할 수 있었다. 본 연구에서는 감성분석 시 한국어 감성사전의 적용이 제한되었고, 기준 데이터를 대변인실에서 제공받은 기사로 활용하였

기 때문에 기사의 성향별 명확한 기준을 수립하여 라벨링 하였을 때 성능 확인은 할 수 없었다. 이는 추후 감성사전의 구축과, 방위사업 관련 기사 분류의 기준을 제시하는 연구를 통해 극복할 수 있을 것으로 판단된다.

주제어: 방위사업, 디지털정체성, 인터넷언론, 극성분석, 문서분류 모델

ABSTRACT

Jang, Sanghoon
Dept. of Defense Acquisition Program
The Graduate School of Kwangwoon University

The recent increase in the spread of Internet newspapers and the development of automatic recommendation algorithms blocks the possibility of Internet users not only seeing it differently from the entity but also correcting it when making judgments about objects. Since most of the military's images are also formed through the Internet, it is important to respond appropriately to Internet newspapers and users for the military management. One of the characteristics of Internet newspapers compared to conventional paper newspapers is high productivity and ripple speed, which spreads rapidly when negative articles are produced, and it is difficult to recover after negative images are formed. Therefore, in order to manage images, the military needs to quickly classify articles that require response and management among Internet media and deliver them to related departments to respond appropriately

As a result of the review of previous studies, the military recognized that image management to form a positive image was important for organizational fraud and policy decision-making, and image management meant the management of digital identity, an image created through Internet information. In addition, in order to manage digital identity so that negative images are not formed, it is necessary

to pay relatively more attention to negative articles than positive or neutral articles, which is one of the emotional analysis techniques that analyzes the positive, neutral, and negative attitudes of natural language. Emotional analysis and polar analysis have been actively conducted on general sentences such as Twitter and movie reviews, but there have been no studies on articles related to military and defense projects, especially articles. In addition, in order to classify the classified articles into response departments, a document classification technique was needed to classify documents based on the tasks performed by each response department. Research on document classification has also been studied in various fields, but there have been no studies using data on military and defense projects, and studies that simultaneously analyzed polarity analysis and document classification models have not been confirmed. In addition, since most studies also focused on increasing classification accuracy, there was a lack of review of the appropriateness of data managed in practice.

Therefore, in this study, we would like to develop a model for classifying which content of Internet articles should be managed and classifying departments suitable for responding to articles in order to manage digital identity related to defense projects. To this end, 2,153 articles related to the Defense Acquisition Program Administration from '16 to '22' collected and managed by the person in charge in the actual spokesman's office of the Defense Acquisition Program Administration were used. In addition, the experiment was conducted by dividing it into original data, data that refined the original data appropriately for deep learning, augmented refining data that augmented the quantity of purified data, and re-labeled data that reclassified the purified data. As a result, 27.6% of articles related to defense projects have a negative character that should focus more on

management. In addition, in terms of classification accuracy, the original data was 50.1%, the refined data was 60.7%, the augmented refining data was 60.3%, and the re-labeled data was 70.7%.

Through experimental research, it was concluded that refining, labeling, and managing data suitable for artificial intelligence learning helps improve the performance of the document classification model and that some improvements are needed in the data management method currently managed in practice. In this study, the application of the Korean emotional dictionary was limited in emotional analysis, and since the reference data was used as an article provided by the spokesperson's office, performance could not be confirmed when clear standards for each article's propensity were established and labeled. It is believed that this can be overcome through the establishment of an emotional dictionary in the future and research that presents the criteria for the classification of articles related to defense projects.

Keywords : Defense Project, Digital Identity, Internet Media, Polar Analysis, Document Classification Model

목 차

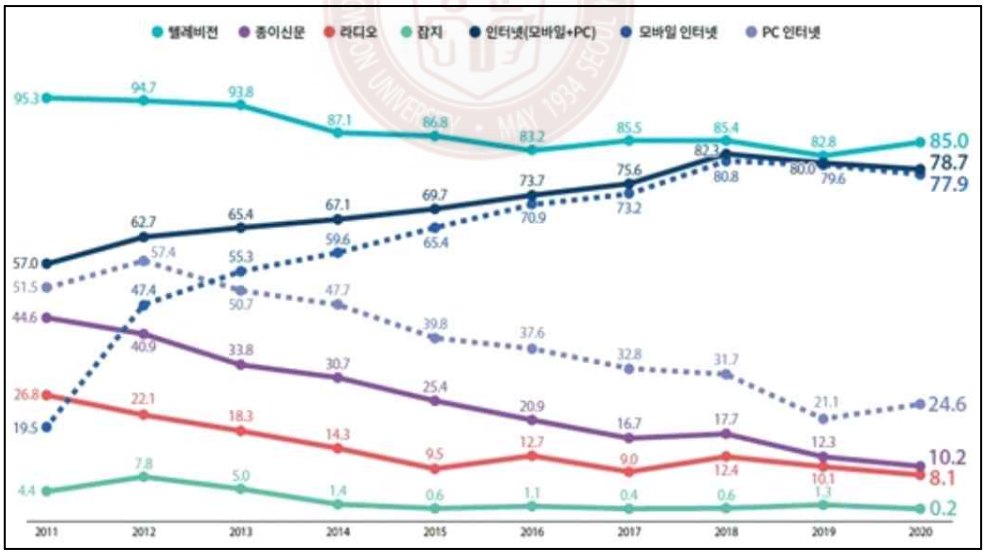
그림 목차

표 목차

제1장 서 론

제1절 연구의 배경 및 목적

인터넷 미디어의 폭발적인 증가와 스마트 기기의 대중화는 제3자가 특정 대상에 대한 이미지를 판단할 때 실제 경험을 바탕으로 하기보다, 인터넷 기사나 뉴스 등 인터넷 매체의 정보에 의존하는 성향을 매우 증가시켰다.[1] <그림 1>은 한국인이 이용하는 미디어가 기존의 텔레비전이나 종이신문, 잡지 등에서 모바일과 PC를 활용한 인터넷 매체로 변화되고 있음을 보여 준다.



* 출처 : 한국언론진흥재단

<그림 1> 미디어 이용률 추이

1) 장상훈.(2019).국방부 디지털 정체성 관리를 위한 실시간 인터넷 문서 자동검색 및 분석 프로그램 발명에 관한 연구.선진국방연구,2(3),1-21.

또한 <표 1>은 연령대별 뉴스 미디어별 이용시간을 나타내는데, 40대 이전의 연령층의 경우 인터넷 미디어를 이용하여 뉴스를 이용하는 인원이 50% 이상으로, 시간이 흐름에 따라 인터넷 매체를 활용한 동영상이나 신문을 소비하는 인구는 지속 증가할 것으로 판단된다.

<표 1> 연령대별 뉴스 미디어별 이용시간 점유율

(단위 : %)

미디어구분	전체 (n=5,061)	20대 (n=929)	30대 (n=968)	40대 (n=1,059)	50대 (n=989)	60대 이상 (n=989)
텔레비전	48.5	25.7	34.4	46.3	54.6	74.5
인터넷(A+B)	31.7	57.5	50.4	32.9	19.8	6.1
이동형인터넷(A)	17.5	33.9	28.2	18.2	9.8	2.6
고정형인터넷(B)	14.2	23.6	22.3	14.7	10.0	3.5
종이신문	10.9	4.9	6.2	11.7	16.4	13.0
라디오	4.6	1.0	3.0	5.2	7.1	5.7
소셜미디어	4.1	10.8	5.8	3.5	2.0	0.3
종이잡지	0.2	0.1	0.2	0.4	0.1	0.1
합계	100.0	100.0	100.0	100.0	100.0	100

* 출처 : 한국 언론진흥재단

인터넷 매체 중 인터넷 신문은 기존의 종이 신문에 비하여 빠르게 생산 및 전파되고, 이를 접한 사용자가 동일한 정보를 재생산하여 전파하거나 유사한 정보를 생산할 수 있는 등 상호작용을 통해 정보가 유통되고 축적된다.[2] 이러한 인터넷 세상 속에서 개인이 스스로를 정의하는 자아나 어떤 대상에 대하여 인터넷 정보들이 모여져 만들어진 이미지를 디지털 정

2) 윤호영.(2011).한국 인터넷의 특징: 소통기반 정보축적 및 유통 문화.한국사회학,45(5),61-104.

체성이라고 하는데, 본 연구에서 의미하는 후자의 디지털 정체성의 경우 대상의 이미지는 실체가 아니라 제한된 인터넷 정보를 기반으로 형성되기 때문에 현실과 차이가 존재할 수밖에 없다.[3] 그리고 이렇게 생성된 디지털 정체성과 관련된 정보들은 다시 인터넷 매체를 통해 전파되고, 기록이 저장된다. 특정 대상에 대하여 유사한 이미지가 반복적으로 노출되면 그 대상의 실체와 디지털 정체성의 차이를 잘 이해하고 있는 사람이 아니라면 대상에 대하여 지속적으로 노출된 이미지를 떠올리게 된다.[4] 또, 이미 형성된 디지털 정체성은 인터넷에서 관련 내용을 접하였을 때 인터넷 사용자가 인지하는 긍정과 부정반응에도 영향을 준다.[5]

특히 제목을 통해 짧게 이슈를 표현해야 하는 인터넷 신문은 간략한 단어로 특정 대상을 수식하는 경우가 많다. 예를 들어 방위사업청에 대해 방산비리, 부패, 군피아 등의 수식어를 합성하는 것이다. 이러한 행위가 반복되면, 방위사업청이라는 단어만 보아도 비리와 부패라는 단어가 자연스럽게 떠오르고, 향후 방위사업청과 관련된 정보를 접하였을 때 부정적인 감성을 가지게 되는데, 언어학자 조지 레이코프는 이를 '특정언어와 연결되어 연상되는 사고의 체계'인 '프레임'이라고 주장하였다.[6]

또한 최근 증가한 인터넷 신문과 매체의 자동추천 알고리즘의 발달은 인터넷 사용자가 대상에 대한 판단을 할 때 실체와 다르게 볼 뿐만 아니

---

3) 한창진(Changjin Han),조민수(minsu Cho),and 이중식(Joonseek Lee). "인터넷에서의 설화(舌禍)뉴스 생산의 확산에 대한 연구." 한국HCI학회 학술대회 2010.1 (2010): 681-685.
4) 한국 언론진흥재단. "2014 언론수용자 의식조사", 검색일 : 2022.10. 8. 출처 : https://www.slideshare.net/girujang/2014-45645888
5) 정이상, 이석용. "인터넷 쇼핑몰의 기업 이미지와 품질특성과 만족도, 충성도의 구조관계에 관한 실증적 연구." 경영과 정보연구 28.4 (2009): 175-197.
6) 나익주, 조지 레이코프, 커뮤니케이션북스, 2017년

라 이를 정정할 가능성을 차단하게 한다. 누리꾼과 언론이 집중하지 않을 대상이라면 이 현상에 대한 부작용이 크지 않겠지만, 언론의 대상이 공인이나 기업, 정부 기관이 된다면 지속적일 뿐만 아니라 고유의 가치에 대한 회복 자체가 어려운 경우로 확대될 수 있는데 그 이유는 다음과 같다.[7)]

첫째 공인이나 기업, 정부 기관은 고객이나 국민의 지지가 운영에 큰 영향을 미치는데, 이를 위해서는 적절한 이미지 관리가 필수적이기 때문이다.[8)] 따라서 현재의 공인이나 기관의 상당수 역시 긍정적인 이미지 형성을 위해 노력한다. 인터넷 매체 및 사용자의 증가에 따라 이러한 이미지 관리는 점차 아날로그 매체에서 인터넷으로 이동하고 있는 성향이 강하다. 과거에는 표지신문, TV의 광고 등을 활용해서 성과와 활동을 홍보하였다면 지금은 인터넷을 이용해서 적극적으로 홍보하는 추세로 전환하고 있다. 즉, 디지털 정체성을 고객들이 선호할 대상으로 만들기 위해 노력 하고 있다.[9)] 그러나, 디지털 정체성을 형성하기 위한 정보는 그들 스스로 뿐만 아니라, 불특정 다수도 어렵지 않게 생산할 수 있다. 이때, 관련 정보는 대상이 직접 주장하는 내용이 아니거나, 정확하지 않은 경우도 많으며, 의도를 가지고 전혀 다른 사실, 즉 가짜뉴스를 만드는 때도 있다. 그러나 인터넷의 경우 몇몇 방송사에 의해 제어되는 매체가 아니며, 신문의 특성 상 국민의 알 권리를 보장하여야 하기에 원하지 않는 부정적인

---

7) 김인식, 김자미. 유튜브 알고리즘과 확증편향. 한국컴퓨터교육학회 학술발표대회논문집 ,25(1(A)) 71-74. 2021

8) 황창호(Hwang Changho). "정부역량에 대한 국민인식이 정부성과인식에 미치는 영향 : 정부의 내·외부역량을 중심으로." 지방정부연구 23.4 (2020): 167-189.

9) 이미나, 박천일, 왕상한. (2021). 국내 주요 기업의 유튜브 분석: 홍보 활동과 현황. 광고PR실학연구, 14(1), 33-54.

정보가 생산되더라도 이를 원천적으로 제거하거나, 법으로 규제할 방안은 현재까지는 존재하지 않는다.[10] 그러나 가짜 뉴스가 사회적으로 미치는 악영향은 <표 2>와 같이 그 피해가 분명하다. 방위사업과 관련된 예로는, 새로운 무기체계 개발소식이나 방위사업과 관련된 내용에 대해 언급하더라도 방산비리와 연관하여 희화화하거나 의도적으로 성과를 낮추어 보는 시선은 종종 발생하며, 이는 인터넷에 축적되어 방위사업과 관련된 디지털 정체성을 관리하는데 부정적인 영향을 준다.[11]

<표 2> 가짜 뉴스의 경제적 비용 추정

구분		피해금액
당사자 피해금액		22조 7,700억 원
	개인	5,400억 원
	기업	22조 2,300억 원
사회적 피해금액		7조 3,200억 원
합계		30조 900억 원

* 출처 : 정민, 백다미(2017)

둘째 인터넷 신문의 특징은 노출이 쉽고 접근이 용이한 만큼 재생산도 유리하다. 한 기사가 일정 인원에게 노출이 된다면 그 인원 중 일부가 다시 해당 기사에 접근 및 복사 또는 일부 수정한 내용이나, 자신의 감성을 덧붙인 정보를 원하는 만큼 재생산하여 실시간으로 전파할 수 있다. 예를

10) 미디어 오늘, “첨예한 갈등 떠들썩했지만 조용히 사라진 언론중재법 개정안”, 검색일 : 2023. 6. 4., 출처 : http://www.mediatoday.co.kr/news/articleView.html?idxno=304915
11) 연합뉴스, “모포털기 사라지나...군, ‘평시 솜이불·전시 침낭’ 대체 추진”, 검색일 : 2023. 5.24, 출처 : https://www.yna.co.kr/view/AKR20210711023951504

들어 어떠한 공인이나 조직과 관련된 기사가 보도되는 경우 꼬리표처럼 과거의 논란이나 이슈가 댓글에 언급되고, 이 내용에 대해 잘 모르는 인터넷 사용자가 다시 관심을 가지고 이와 관련된 정보를 생산하는 것은 이미 일반적인 현상이며, 전문적으로 이를 이용하여 관련 내용을 방송 주제로 삼는 유튜브 크리에이터들까지 존재한다. 신조어로 '사이버 렉카'라고 불리는 이들은, 자기 채널의 조회 수 향상을 목적으로 대중들이 관심 있을 것 같은 공인이나 기관을 대상으로 각종 자극적인 의혹을 제기하는데, 때로는 확인되지 않은 내용을 사실인 것처럼 보도하고도 거짓으로 밝혀져도 특별한 책임을 지지 않는 경우가 있다.[12] 게다가 해당 보도가 잘못된 것으로 밝혀져 해당 내용이 삭제되더라도 인터넷 매체의 특성상 다른 인터넷 사이트에 동일한 내용이 남아있거나, 유사한 내용이 재생산되어 인터넷 상에 상존해 있거나 오히려 확산될 가능성이 높다.[13] 따라서 디지털 정체성 관리에 있어서, 한번 형성된 부정적인 기사는 삭제되지 않고 지속적으로 부정적인 요소로 작용할 수 있다.

셋째는 정보를 접한 후, 사실관계 및 실체를 검증하는 과정 역시 인터넷 정보에 의존하는 경향이 높아진다는 것이다. 예를 들어 공인이나 기업에 대한 정보를 확인하고자 할 때, 대상과 특정한 계약을 체결하려는 관계의 상대자나 공식적인 업무를 수행하는 기관 등의 상황이 아닌 경우에는 대상이 제공하는 정식 공문, 진술 등을 이용하는 경우보다는, 간편하게

12) IT동아, "사이버 렉카 "또 떴다!..유튜브만 믿으면 되나?", 검색일 : 2023. 5.24, 출처 : https://it.donga.com/101778/

13) 최윤성 (Younsung Choi), 권오걸 (Oh-geol Kwon),and 원동호 (Dongho Won). "인터넷 쿠키로 인한 프라이버시 침해와 잊혀질 권리에 관한 연구." 인터넷정보학회논문지 17.2 (2016): 77-85.

인터넷을 이용하여 그 대상에 대해 검색해 보는 것이 성향이 높다.[14] 그러나 이때는 상기 첫째와 둘째 사유로 인하여 검색한 대상에 대해 특정한 성향을 가진 인터넷 정보가 이미 만들어져있을 가능성이 있다. 따라서 인터넷에서 검색한 결과는 이미 형성된 디지털 정체성을 변화시킬 요소가 많지 않다.

지속적으로 원하는 디지털 정체성을 형성하지 못하고 의도하지 않은 정보가 재생산되고 확산되면, 인터넷에서의 공인이나 기업, 정부 조직의 이미지는 이미 인터넷에 확산된 정보를 기반으로 형성되어 있고, 고착화될 가능성이 높다. 대상에 대한 정보를 획득하고자 하는 새로운 인물이 인터넷으로 실체를 확인하고자 검색하면 이미 만들어진 디지털 정체성이 노출되고, 불특정 다수에게 동일한 과정이 반복되기 때문에 이미 형성된 디지털 정체성에서 벗어나기란 몹시 제한될 것이다.

이러한 인터넷 매체의 특성으로 인해 원하지 않는 성향의 인터넷 정보가 생산되고 확산되어 부정적인 디지털 정체성이 형성되는 현상에 대응하는 것은 개인과 조직 모두 분명한 한계점이 존재한다.[15] 이는 최근에 발생한 문제가 아니라 2000년 이후 한국에서 인터넷이 발달하면서 인터넷 자료에 근거한 언론의 활성화에 따라 일어나고 있으며, 발생한 오보에 대해 강력하게 처벌하는 등 적극적인 구제방안은 현재 부재한 상태이다.[16] 일례로 '21년 8월 당시 여당인 더불어 민주당은 언론의 허위, 조작 보도에

14) 조성태. (2012). 스마트기기의 이용량 증가에 따른 인터넷 포털뉴스 편집환경 연구 - 국내 포털뉴스 사이트를 중심으로 -. 한국디자인포럼, 35, 441-450.

15) 박재현 (Jae-hyun Park),and 최호규 (Ho-gyu Choi). "인터넷 불매운동에 대한 소비자 의식과 불매운동이 기업의 이미지와 매출에 미치는 영향." 기업경영리뷰 1.2 (2010): 161-180.

16) 이용성. (2008), 인터넷 자료에 근거한 언론보도의 문제점과 개선방안-인터넷 자료 근거한 오보의 발생구조를 중심으로, 언론중재, 107, 41-49.

대한 특칙을 마련하고자 하였으나 현재까지 입법은 중단된 상태이다. 주요 골자는 「언론중재 및 피해구제 등에 관한 법률(법률 제16060호, 2018.12.24.)」 에서 언론 등의 고의 또는 과실로 인한 위법행위로 피해를 받은 경우 징벌적 손해배상을 할 수 있는 항목을 추가하는 것이었다. 그러나 이는 사회 권력에 대한 비판, 감시 기능의 약화, 국민의 알 권리 침해로 이어져 민주주의 발전에 걸림돌이 될 수 있다는 논리 등이 야당 뿐 아니라 발의한 더불어 민주당 당내에서도 주장되어, 추진할 수 없었다.[17] 그러나, 한국사회여론연구소의 조사에 따르면 국민 중 54.1%는 해당 법안의 입법을 찬성했었다.[18] 해당사안을 살펴보면, 이렇게 과반의 국민이 해당 법안에 동의하였고, 정부 여당에서 언론중재법의 개정을 추진한 것 자체가 언론의 허위·조작보도·오보와 그로 인한 피해가 매우 크며, 피해에 대한 대응이 적절하게 이루어지기 어렵다는 것을 보여 준다.

방위사업분야 역시 상황은 동일하다. 오늘날에도 해군의 전력사업, 즉 방위사업청의 함정사업 추진에 대해서는 방위산업비리('15년), 육군의 새로운 군수품 도입에 대해서는 오래된 식수통 이야기('13년) 등 과거 국방획득분야의 각종 비리와 연계하여 형성된 부정적인 이미지가 인터넷에 축적, 유통되면서 지속적으로 재생산되어 디지털 정체성은 부정적 이미지 고착화가 이루어질 가능성이 높기 때문에 특별한 관리가 필요한 실정이다.[19] 특히 방위사업의 주도 정부기관인 방위사업청의 경우, 최근 방산업

17) 한겨레, "언론중재법 공개 반대..."법이 언론개혁 공감대 훼손"", 검색일 : 2023. 5.24, 출처 : https://www.hani.co.kr/arti/politics/assembly/1009024.html

18) 이투데이, "'언론중재법' 처리 앞두고 '팽팽'...의견 어떻게 다른가", 검색일 : 2023. 5.24, 출처 : https://www.etoday.co.kr/news/view/2056078

19) 김태호. "방위사업청 온라인(On-line) 홍보 활성화를 위한 연구." 국내석사학위논문 광운대학교 대학원, 2009. 서울

체에 직접 청장이 방문하여 애로사항을 청취하는 '다파고'나, 수출진흥에 대한 홍보 등 많은 노력을 기울였음에도 불구하고 2022년 설문조사 결과 국민들은 방위사업청을 '신뢰하지 않는다'는 비율이 32.2%로 '신뢰한다'의 20.3%를 웃돌았다. 이때, 방위사업청과 연상되는 부정적 단어와 표현은 방산비리·부패(78.4%), 폐쇄적 조직(7.5%), 로비스트(4.3.%) 등으로 과거 방산비리와 관련된 내용이었다.[20] 그러나, 이는 과거 방위사업 전반에 대한 무리한 수사, 조사, 감사의 행태로 인해 왜곡된 사실이 지속적으로 보도되었던 영향성이 크다고 판단된다.[21] 게다가 인터넷 신문의 경우 인터넷 공간에 지속 축적되어 지속적으로 영향력을 행사하며, 또다시 부정적인 인터넷 기사가 발생할 경우 빠르게 전파되고 재생산 되는 과정 중 과거의 왜곡된 보도와 혼합되어 부정적인 영향성이 더욱 커지기 쉽다. 따라서 인터넷 신문이 생성되었을 때 이 기사가 전파되기 전에 관리가 필요한 내용인지 파악하고, 적절하게 대응하는 것이 중요하다.

따라서 본 연구에서는 방위사업과 관련된 디지털 정체성을 관리하기 위해 인터넷 기사 중 어떠한 내용의 기사들을 관리하여야 하는지에 대한 분류 및 기사 대응에 적합한 부서를 분류하기 위한 모델을 개발하고자 한다.

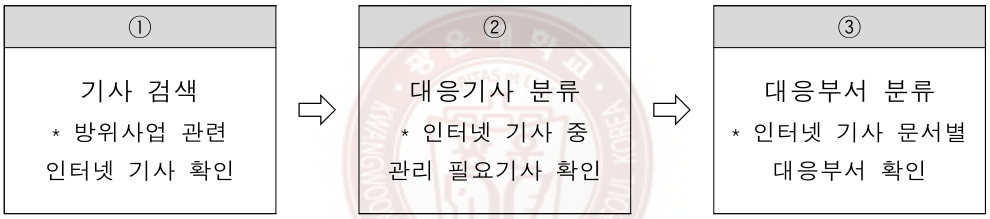
---

20) SPN서울평양뉴스, "북 미사일 70발 쏘자...국민 10명 중 7명 '3축 체계 강화'지지", 검색일 : 2023. 6. 4, 출처 : https://www.spnews.co.kr/news/articleView.html?idxno=60426

21) 최기일, 채우석. (2018). 방위사업 비리 관련 처벌 현황 진단 및 분석 연구. 한국방위산업학회지, 25(4), 13-31.

제2절 연구범위 및 절차

방위사업 관련 디지털 정체성을 관리하기 위해서는 인터넷 기사의 작성으로 방위사업과 관련된 정보가 생산되면 대응 필요성을 결정하여 이 정보가 전파되기 전에 대응할 수 있는 준비를 하여야 한다. 즉, <그림 2>와 같이, 방위사업 관련 기사검색, 대응기사 분류, 대응부서 분류의 과정이 필요하다.[22] 본 연구에서는 방위사업의 대표적인 기관인 방위사업청과 이와 관련된 공식적인 인터넷 기사에 한정하여 연구를 진행하였다.



<그림 2> 방위사업 관련 인터넷 기사 대응단계

첫 번째 단계인 관련 기사 검색의 경우 최근 인터넷 신문의 발달과 함께 이용자의 편리성을 위해 개발된 도구들을 활용할 수 있다. 예를 들어, 대표적인 검색엔진을 자랑하는 구글의 경우에는 '구글 알리미'라는 기능을 이용해서 원하는 키워드가 검색되는 경우 사용자의 메일로 관련 검색결과를 송신해주는 서비스를 제공하고 있다. 구체적으로, 사용자가 구글 알리미에 관심이 있는 키워드를 설정하면, 사용자가 접속해 있지 않더라도 구글 포털 검색창에 자동으로 해당 키워드를 반복 입력하는 효과를 가진다.

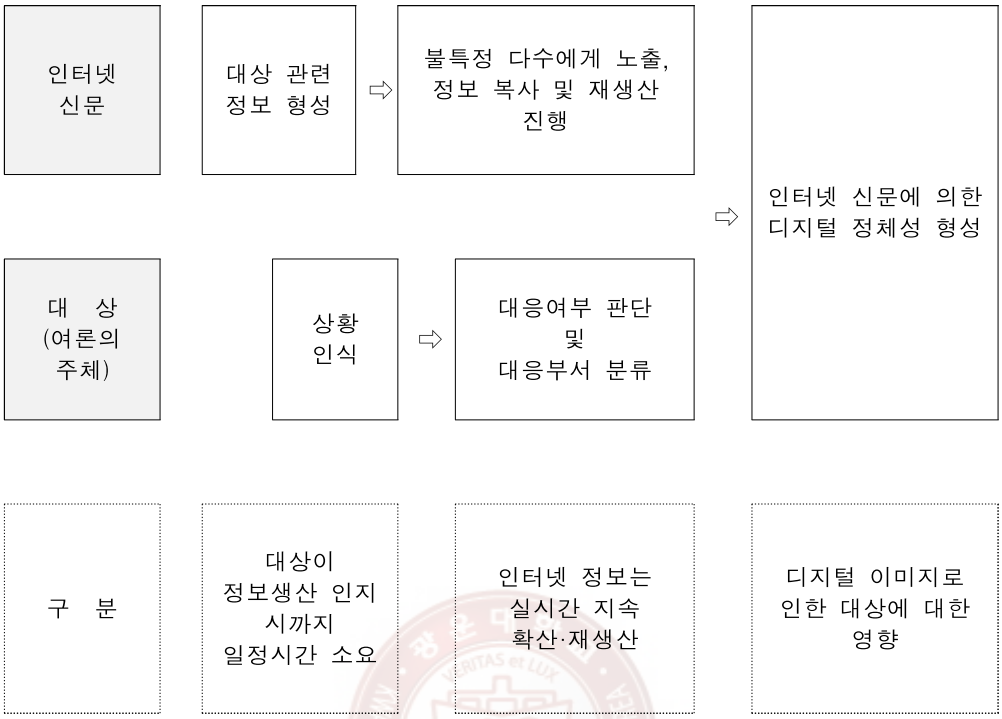
22) 장상훈, 전자대변인시스템, 제10-1812933호, 출원 : 2017. 12. 6., 등록 : 2017.12.20.

이 과정에서 해당 키워드가 검색된 문서를 발견하면 사용자에게 메일로 검색된 문서들의 링크를 보내주는 것이다. 특히 수신빈도, 출처(자동, 블로그, 뉴스, 웹, 비디오, 도서, 토론, 금융), 지역 등을 설정할 수 있어 사용법이 익숙해지면 원하는 시간에, 원하는 출처에서의 필요한 정보를 파악하는데 적절하게 이용할 수 있다.[23] 또한 YTN의 경우에도 '뉴스 알림(웹푸시)' 기능을 이용해서 속보 및 주요뉴스를 실시간으로 사용자에게 전달해주며,[24] 카카오톡에서도 역시 '뉴스봇'이라는 기능을 활용하여 키워드와 관련한 뉴스를 찾아주고, 연관된 주제까지 파생시켜 추가 검색할 수 있다.[25]

두 번째 단계인 대응여부 판단이란, 방위사업청의 디지털 정체성 관리를 위해 검색된 인터넷 신문 내용이 별도의 대응이 필요한지를 분류하는 것이다. 방위사업청에서는 담당자가 지속적으로 신문 내용을 확인하면서 이 분류작업을 수행하고 있다. 그러나 연구의 목적인 디지털 정체성 관리를 위해서는 인터넷에서 생성된 모든 신문에 대해서 대응할 필요성은 없고, 부정적인 이미지를 형성시킬 수 있는 내용만 추출해야 한다. 특히, 인터넷 신문에 대응하기 위한 핵심 중 하나는 대응속도다. 따라서 <그림 3>과 같이, 쏟아지는 여론이 전파되기 전에 관리하기 위해서는 현재와 같이 담당자가 직접 모든 내용을 검토해보고 대응 필요성을 검토하는 것은 확산속도를 고려할 때 한계점이 존재한다고 판단된다.[26]

---

23) 출처 : https://www.google.co.kr/alerts
24) 출처 : https://www.ytn.co.kr/info/webpush.php
25) 출처 : https://pf.kakao.com/_WlSxbu
26) 경향신문, "가짜뉴스 SNS 전파 속도 '진짜'보다 최고 20배 빨라", 검색일 : 2023. 6. 4, 출처 : https://www.khan.co.kr/economy/economy-general/article/201803090400005



<그림 3> 현행 인터넷 신문에 대한 대응 도식도

세 번째 단계인 대응부서 분류는 관리가 필요한 인터넷 신문에 대해 대응할 적절한 대응부서를 정하는 것이다. 방위사업청에 대해 부정적인 디지털 정체성을 형성하는 인터넷 기사의 방법을 다양한데, 이때는 해당 기사와 관련된 담당자의 검토를 통해 기사의 정정 요청, 사과, 개선행위 약속, 보상 등 방위사업청과 담당조직의 의사가 포함된 적절한 대책을 수립해야 하기 때문이다.[27] 이 역시 현재는 방위사업청 담당자가 기사의 성격과 내용을 검토하여 대응할 담당 부서를 분류하고 있다.

대응할 담당 부서를 분류하는 것은 첫 번째 단계인 인터넷 기사를 검색

27) 나재훈. "軍의 이미지 회복 전략에 관한 연구." 국내석사학위논문 고려대학교 대학원, 2008. 서울

하는 단계에서도 담당 부서별 수행하는 업무 등을 기준으로 키워드로 설정하여 일정 수준은 수행할 수 있다. 예를 들어 기사 검색 시 키워드를 '방위사업청'과 '군함'이라고 설정하면 검색된 기사의 대응부서는 적어도 군함을 건조하는 사업 담당 부서와 관련한 내용이라고 판단할 수 있다. 그러나 내용적인 측면에서, 수상함에서 발사한 유도탄이 불량하다는 취지의 기사라면 수상함을 만드는 함정사업부가 담당해야 하는지, 유도탄을 만드는 유도무기사업부에서 검토해야 하는지, 불량한 사실관계를 파악하기 위한 감사관실에서 확인해야 하는지는 단순한 키워드 설정으로는 분류에 제한점이 있다. 이러한 현실적인 제한사항으로 담당자는 실질적으로 분류에 많은 시간을 할애하고 있다.[28] 결국 대응이 필요한 인터넷 기사에 대하여 적절하게 대응할 수 있는 담당을 지정하는 것은 필수적인데, 문서분류와 동일하게 이 작업 역시 대응시간에 영향을 미친다. 인터넷 문서의 경우 많은 사진과 글을 담고 있고, 이슈에 대해서 지속 생산되고 전파되는데, 이 작업을 사람이 읽어보고 담당을 분류하기 때문이다. 또한 인터넷 기사는 실시간 디지털 미디어로 전달된다는 점에서 경쟁적인 속보성을 가지게 되며, 보도자료 기사 등은 지면의 영향이 없기에 기존 종이 기사에 비하여 기사의 길이가 길다는 특징이 있어 더욱 많은 시간이 소요될 수 있다.[29] 따라서 기계학습을 통해서 단순한 키워드 설정에 의한 분류보다 문서분류의 성능이 뛰어난 모델의 개발이 필요하다.

정리하면 방위사업청의 디지털 정체성 관리를 위해 부정적인 인터넷 기

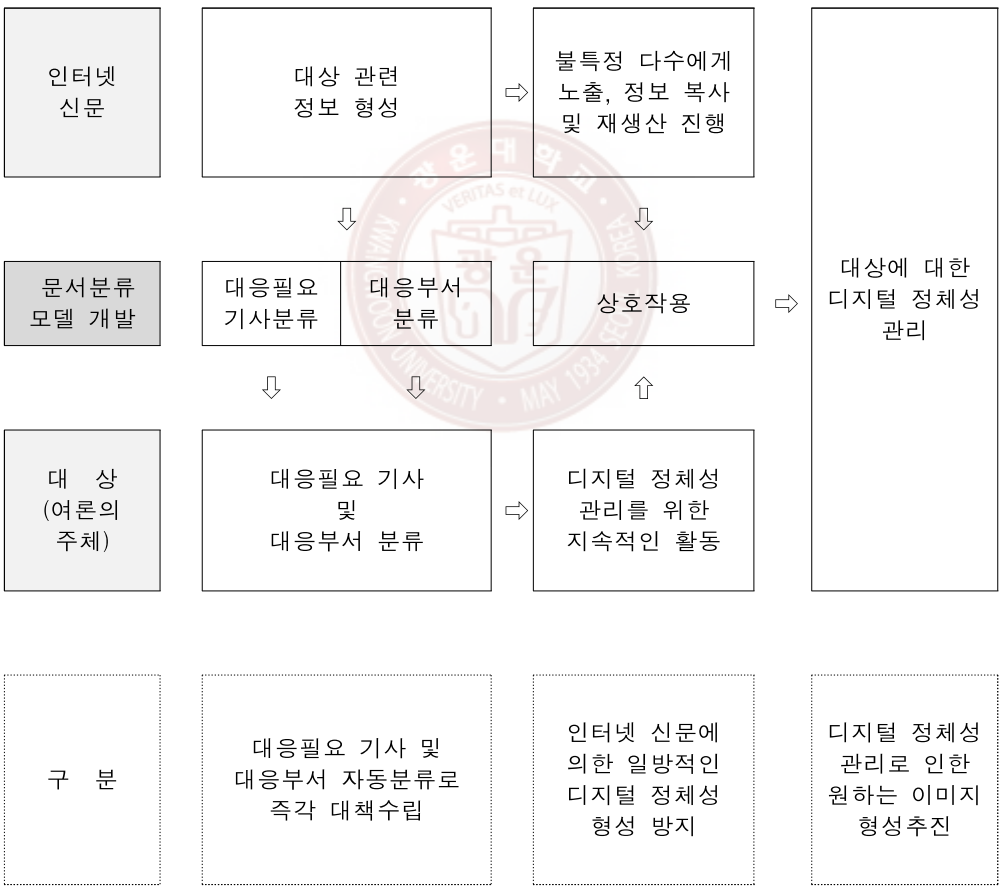
28) 조덕현. "중앙정부부처 홍보조직의 우수성에 관한 연구." 국내석사학위논문 경희대학교 언론정보대학원, 2011. 서울

29) 김익현. "인터넷의 매체특성이 인터넷신문 기사에 미치는 영향." 국내석사학위논문 연세대학교 언론홍보대학원, 2003. 서울

사의 내용에 대응하기 위해서는 다음과 같은 인터넷 기사의 특징을 고려해야 한다. 첫째, 대상에 대하여 속보로 보도할만한 특별한 사건이나 이슈가 발생하면, 수많은 매체에서 동시다발적으로 유사한 내용의 기사가 작성된다. 그리고 일정 시간을 두고 일부 수정된 기사가 다시 생성된다. 둘째, 사안에 따라서는 해당 내용에 대해서 각 매체별로 각자의 시선으로 다양한 심층 분석을 수행하여 기사를 생성한다. 이 경우, 명확한 디지털 정체성 관리를 위해서는 기사의 대상이 되는 개인이나 조직은 동시다발적으로 생성되는 기사의 내용을 모두 확인하여야 한다. 매체별로 중점을 상이하게 다루는 속보들에 대하여 각각의 대응방안을 모색해야 하기 때문이다[30]. 또한 언론사에서 추가분석이 수행된 기사의 경우 분량이 길어지고 내용도 일부 변경되는 경향이 있기 때문에 어떠한 내용인지 검토하는데 시간이 소요되고, 최초의 보도내용과 바라보는 각도가 달라질 수 있어 대응방안 역시 달라질 수 있다. 예를 들어 방위산업 업체에 사고가 발생하였다면, 최초에는 어느 업체에서 어떠한 사고가 발생했다는 보도가 지속적으로 생성될 것이다. 그 가운데에는 단순 사실을 보도하는 경우도 있고, 중대재해처벌법 등 다른 이슈와 연계지어서 보도할 수도 있다. 또한 심층분석을 추진하게 되면, 해당 업체에서 추진하는 방위산업 수출과 관련된 내용의 신빙성이나 정상적인 추진 여부에 대한 기사까지도 작성될 수 있다. 이러한 경우, 단순 사실이라면 해당 업체와 그와 관련된 사업을 하는 부서가, 중대재해처벌법 등 정책적인 내용이 위주라면 방위사업정책을 수행하는 부서가 담당하여야 하며, 그 업체가 수행하는 내용이 주요 방위산

30) 진행남. "인쇄신문과 독립 인터넷신문의 기사특성에 관한 비교연구." 국내박사학위논문 경희대학교 대학원, 2002. 서울

업수출 등에 대한 내용이라면 또 다른 부서가 배정되어서 담당하여야 한다. 그러나 언급한 바와 같이 사람이 이러한 업무를 수행하기 위해서는 지속적으로 기사를 검색 및 감시하고, 분류하는 물리적인 시간이 필요하며, 이 작업 중에도 새로운 기사들은 생성되고 있을 것이다. 따라서 <그림 4>와 같이 대응이 필요한 기사를 분류하고, 이 기사를 담당해야하는 부서를 분류하는 작업을 수행할 수 있는 모델의 개발이 필요하다.



<그림 4> 문서분류 모델을 이용한 인터넷 신문에 대한 대응 도식도

본 연구에서는 방위사업청 관련 기사검색 단계는 방위사업청에서 관리하고 있는 인터넷 기사 데이터를 이용하였다. 문서분류가 적절한지 비교하기 위해서는 담당자에 의하여 이미 문서분류가 이루어진 데이터가 필요하기 때문이다. 그리고 <그림 5>와 같이 이 기사들을 현재 관리되고 있는 원본 데이터와 기계학습을 위해 기사내용 중 학습과 불필요한 내용(노이즈) 등을 제거한 정제 데이터, 정제 데이터 수량을 증강한 증강정제 데이터, 문서분류의 기준을 재검토한 재라벨 데이터로 분류하였다. 이후 이 중 기계학습의 기준이 되는 정제 데이터로 극성분석을 추진하고, 모든 데이터를 활용하여 개발한 문서분류 모델의 성능을 비교하였다.

<그림 5> 연구 절차

제2장 이론적 배경 및 선행연구 고찰

제1절 군 디지털 정체성 관리

디지털 정체성이란 인터넷과 현실이 연결된 사회에서 제3자에 의하여 생성된 정보 등에 의하여 재구성된 '자아'를 뜻하고,[31] 다양한 디지털 미디어에서 본인과 관련된 정보의 조각들을 통합·분석하여 제3자가 제시한 자아로서 디지털 기반의 자아(Digital data-based self)라고 명명하기도 한다.[32] 이 디지털 정체성은 개인뿐만 아니라 기업도 가지게 된다. 인터넷 뉴스 포탈에서 탈세 논란 등이 발생하였을 때, 이 미디어 정보를 접한 이용자가 기업에 대한 고객으로서 작용하여 해당 기업 가치에 영향을 미친다는 사실이 확인 되었다.[33] 기업들은 이러한 영향성을 이미 인지하여 디지털에서 좋은 이미지를 부각시키고 고객들이 선호하는 채널과 서비스를 제공하는 등의 활동을 통해서 고객의 니즈를 만족시키는 기업의 이미지를 홍보할 뿐만 아니라, 인터넷의 피드백을 통하여 고객에게 제공하는 서비스를 효율적으로 발전시키기 위해 노력하고 있다.[34] 반면, 인터넷에 의해 결정되는 디지털 정체성 관리는 부작용도 심각한데, 조회 수로 광고수익

---

31) WEF, Technology Tipping Points and Societal Impact, 검색일 : 2023. 5.24, 출처 : https://www3.weforum.org/docs/WEF_GAC15_Technological_Tipping_Points_report_2015

32) 안상훈, 한은영, 장근영, & 김선희. (2013). 초연결 사회에서 디지털 자아의 정체성 연구. 정책연구, 2013(51), 1-167.

33) 이상민 (Sang-min Lee), 박명호 (Myung-ho Park), 김병준 (Byung-jun Kim),and 박대근 (Dae-keun Park). "빅데이터 분석을 통한 인터넷 뉴스 포털에서의 탈세 논란이 기업 가치에 미치는 영향 연구." 인터넷정보학회논문지 22.6 (2021): 51-57.

34) Adobe Summit, ADOBE DIGITAL MARKETING SUMMIT. 검색일 : 2023. 5.24, 출처 : https://business.adobe.com/summit/adobe-summit.html

을 창출하는 플랫폼의 증가에 따라 허위, 왜곡, 확대 등을 통해 많은 조회수를 유도하는 가짜 뉴스나 자극성 기사들도 적극적으로 생성하기 때문이다. 정민, 백다미(2017)는 2017년 연간 가짜뉴스 추정 발생 수는 13만 건에 이르고 이에 의한 피해는 30조로 추산하였으며, MIT 연구에 따르면 가짜 뉴스의 경우 진짜 뉴스에 비해 최대 약 20배의 전파속도를 가지고 있다고 확인된다.[35] 그럼에도 불구하고 법적으로 이를 규제하기는 쉽지 않은데, 고성수(2018)는 표현의 자유 및 국민의 알 권리와 언론에 의한 피해를 예방하고 보장받을 권리가 대치되기 때문으로 분석하였다.[36]

군에 있어 국민의 신뢰는 중요하므로, 이미지를 관리하고자 하는 연구는 지속되었다. 강덕찬(1993)은 군과 국민을 이어주는 매개체는 언론으로 판단하였으며,[37] 서정근(2001) 역시 부정적인 언론은 부정적인 군 이미지를 형성하는데 직결된다고 판단하였다.[38] 그러나 김태웅(2008)은 군은 자체의 이미지 관리를 과학적인 경영보다는 임무 위주의 통솔에 의존해왔으며, 이에 따라 홍보활동 역시 퍼블리시티나 선전의 개념에 치중되어 왔다는 점을 지적하였다.[39] 그러나 시간이 흐름과 함께 김수진(2009)의 연구와 같이, 군내에서도 대중에게 원하는 정책과 이미지 전달을 위해서는 전달하는 이미지와 대상에 대한 세분화 등의 연구가 필요하다는 의견이 제기되었다.[40] 또한, 변의혁(2013)은 군 자체의 홍보보다, 주변인들의 구전

35) 윤인아. “로봇저널리즘의 이해와 전망”, 제4차 산업혁명과 소프트파워 이슈리포트, 2018-18. 정보통신산업진흥원, 1-14
36) 고성수. "가짜뉴스 규제법안 실효성에 관한 연구." 국내석사학위논문 서울과학기술대학교, 2018. 서울
37) 강덕찬. "軍 이미지 類型과 形成要因에 대한 硏究." 국내석사학위논문 高麗大學校, 1992. 서울
38) 서정근. "국내 신문에 반영된 군 이미지와 보도 성향에 관한 연구." 국내석사학위논문 동국대학교 언론정보대학원, 2001. 서울
39) 김태웅. "청소년의 정보원 이용이 군 이미지, 복무의사, 신뢰도에 미치는 영향에 관한 연구." 국내석사학위논문 서울대학교 대학원, 2008. 서울

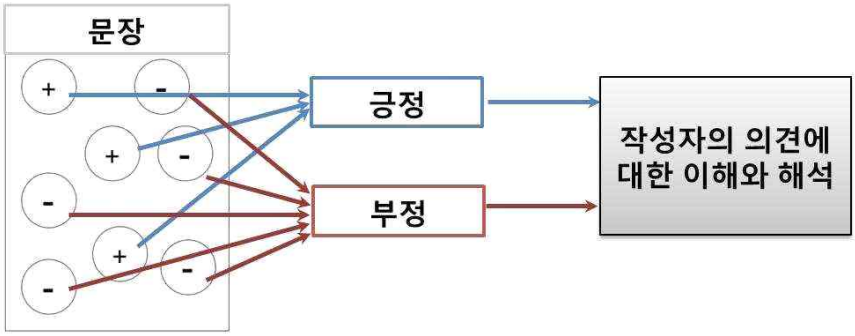
이 군의 긍정적 이미지 형성에 더 중요하다는 결론을 도출하였다.[41] 한편, 군과 방위사업 관련 조직에서의 디지털 정체성 관리에 대한 연구는 많이 확인되지 않았다. 다만, 조인상(2014)은 군에서도 군대는 존재 기반을 확보하기 위한 국민과의 공감대 형성과 전문직업주의가 정착되는 단계에서 유능한 인재의 참여 등을 위해 군 이미지 관리가 매우 중요하다는 점, 조직의 이미지가 디지털 혁명과 정보화의 영향을 받으며 이 가운데 실제적 가치를 상실하지 않기 위한 노력이 필요하다는 주장을 하고 있었다.[42] 또한, 장상훈과 이석준(2019)은 인터넷 미디어의 발전에 따라 정부와 국방 분야에서도 조직의 실제 이미지 자체가 디지털 정체성에 의하여 좌우될 가능성이 증가하기 때문에 본격적인 디지털 정체성의 관리가 필요하다고 주장하였다. 따라서, 대부분의 개인이 인터넷 사용자가 되어있고 인터넷 기사를 통해 언론을 접하는 오늘날, 인터넷 기사를 통한 군의 디지털 정체성 관리는 필수적인 요소로 확인된다.

제2절 극성분석

극성분석은 텍스트를 구성하는 각 단어별 긍정과 부정의 가중치를 정하고 빈도수 등을 계산하여 텍스트의 감성을 분석하는 감성분석 기법 중, 가중치의 판단을 통해 텍스트가 긍정, 부정, 중립 중 어떠한 감성을 가지는지를 판단하는 방법이다. 즉 <그림 6>와 같은 오피니언 마이닝의 한

40) 김수진. "초급장교 교육훈련 홍보 보도기사에 나타난 장교상이 군 이미지와 신뢰에 미치는 영향." 국내석사학위논문 서울대학교 대학원, 2009. 서울
41) 변의혁. "군 홍보가 ROTC 이미지 및 지원의사에 미치는 영향에 관한 연구." 국내석사학위논문 연세대학교 정경대학원, 2013. 서울
42) 조인상. "군 이미지에 관한 통합적 연구." 국내박사학위논문 大田大學校, 2014. 대전

분야인데, 이는 자연어 처리 기술을 바탕으로 텍스트에 내제된 태도를 기준으로 텍스트를 분류하는 것이다.[43]



<그림 6> 오피니언 마이닝 개념

Kamps et al.(2004)은 WordNet 감성사전을 이용하여 문서의 긍정과 부정의 극성을 판단하였다.[44] Esuli & Sebastiani(2006)는 WordNet을 이용하여, 단순 긍정과 부정이 아닌 감성의 정도 값을 정의한 SentWordNet 관련 연구를 진행하였다.[45] 한국어에 대한 자연어 처리를 위한 연구의 경우, 김명규(2010)는 인터넷 텍스트들의 감성을 분석하여 긍정과 부정의 두 극성으로 분류하는 시스템을 개발하였다.[46] 김승우와 김남규(2014)는 감성사전을 활용하여 오피니언 분류에 의미 있는 결과를 보여주었다.[47] 온

43) 김건아. "빅 데이터를 이용한 제품디자인의 감성반응 분석." 국내박사학위논문 부산대학교 대학원, 2016. 부산

44) Kamps, J., Marx, M., Mokken, R. J., Rijke, M., "Using WordNet to Measure Semantic Orientation of Adjectives," Proc. of the International Conference on Language Resources and Evaluation, Vol.4, 2004, pp.1115-1118.

45) Esuli, A., Sebastiani. F., "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining," Proc. of the 5th International Conference on Language Resources and Evaluation, 2006pp. 417-422.

46) 김명규, "인터넷 감성 텍스트에 대한 극성 분류 시스템," 한국항공대 대학원 컴퓨터공학 박사학위 논문, 2010.

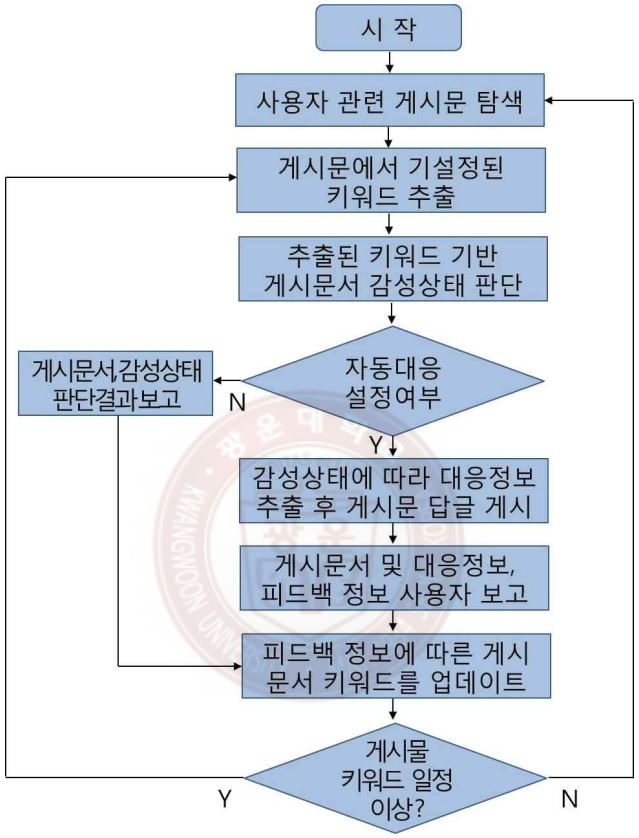
병원 외(2018)는 14,843개의 관용구, 문형, 축약어, 이모티콘 등에 대한 긍정, 중립, 부정 판별 및 정도값을 계산하여 한국어 감성 사전을 구축하였다.[48] 상기 연구내용 및 <그림 7>과 같이 오피니언 마이닝은 다양한 분야에서 활용되고 있다.

<그림 7> 오피니언 마이닝 활용분야

또한, 장상훈(2017)은 국방분야의 디지털 정체성 관리를 위해 감성분석의 기법을 활용한 모델을 제시하였다. 이는 인터넷에서 디지털 정체성을 훼손할 수 있는 오보나 부정적인 텍스트를 실시간으로 검색하고, 그 텍스트에 내제된 감성을 분석하여 수치화한 후, 그 결과를 사용자에게 자동으로 알려주는 모델이다. 전반적인 체계는 <그림 8>과 같이, 사용자가 설정한 자료를 크롤링하고, 텍스트의 감성을 분석 후 필요시 피드백하는 것이

47) 김승우, 김남규. "오피니언 분류의 감성사전 활용효과에 대한 연구," 지능정보연구, 제20권 제1호, 2014, pp.133-148.
48) 박상민, 나철원, 최민성, 이다희, 온병원.(2018).Bi-LSTM 기반의 한국어 감성사전 구축 방안. 지능정보연구,24(4),219-240.

다. 이때 감성 수치, 즉 가중치를 설정하는 방법은 키워드가 되는 단어에 가중치를 설정하여 산출 한다.



* 출처 : 장상훈(2019)

<그림 8> 전자대변인 시스템 체계

전자대변인 시스템은 <표 3>와 <표 4>의 가중치를 기반으로 성향분석 결과를 감성지수(E)를 나타내는데, 제1가중치는 긍정과 부정의 키워드 합에 의해 결정되는 부호와 가중치의 평균점수에서 제2가중치에서 해당 문서의 영향력을 고려한 값을 전체 키워드로 표현된다. 즉, 제1가중치에서

적용된 양의 상수를 a, 음의 상수를 b, 제1가중치 평균값을 계산한 결과를 A, 제2가중치를 B, 제2가중치 키워드의 평균값을 C라고 하고, 양의 상수인 긍정과 중립 키워드의 전체 개수를 n_1, 음의 상수인 부정의 키워드 전체 개수를 n_2, 전체 키워드의 총 개수를 n_3 이라고 하면 E는 식 (2-1)과 같이 산출된다.

$$E = \frac{a(A \cdot B)n_1 + b(C \cdot B)n_2 \sum}{n_3} \qquad (2\text{-}1)$$

다만, 제2 가중치의 경우에는 선행 작업이 상당하고, 관련 연구가 부족하여 최근의 감성분석 연구는 감성사전을 활용하는 방법이 지배적인 것으로 판단된다. 특히 전자대변인 시스템의 경우에는 키워드인 한국어 단어에 대한 가중치가 텍스트 감성을 분석하는 결정적인 요인으로, 관련근거를 마련할 수 있는 연구가 필요하다.

<표 3> 전자대변인 시스템 화면 제1가중치 선정

구 분	부호	가중치	키워드 예시	비 고
중립	+	0	규모, 예산, 조직, 병력, 위치, ...	단순정보
긍정		0.8	고마운, 즐거운, 도움이 되었습니다...	긍정표현
		1	항상 감사합니다, 가장 보람찼던 경험...	큰 긍정표현
부정	-	0.8	수사를 받았다, 혐의가 있다...	부정표현
		1	역량부족, 비리, 부패, 경질을 요구...	큰 부정표현

* 출처 : 장상훈(2019)

해당 가중치를 계산할 때 단순한 단어뿐 아니라, 그 출처의 영향성도 고려할 수 있다. 즉 온라인상에서의 조회 수 등을 고려하여 키워드에 대한 제2 가중치를 반영할 수 있다.

<표 4> 전자대변인 시스템 화면 제2가중치 선정

구분	조회 수(좋아요)	제2가중치
100만 회원 사이트	100회 이하	1.1
	101~300회	1.2
	301~400회	1.3
1만 회원 사이트	50회 이하	1.1
	51~80회	1.2
	81~100회	1.3

* 출처 : 장상훈(2019)

VADER(Valence Aware Dictionary and Sentiment Reasoner)는 감성 분석을 위해 NLTK(Natural Language Toolkit)에서 제공하는 프로그램으로, 감성 단어 사전과 규칙 기반의 감성 점수 계산 방법을 활용하여 텍스트의 감성을 파악할 수 있는 프로그램이다. VADER는 기존에 단어만으로 감성을 평가하는 방식에서 발전하여, 문장의 감성 분석에 뛰어난 성능을 발휘한다.[49] 즉, 감성 단어 사전과 규칙 기반의 감성 점수 계산을 활용하여 텍스트의 감성을 파악하는데, 우선 사전에 등록된 감성 단어를 기반으로 문장의 감성 점수를 계산한다. 감성 단어 사전은 긍정적인 의미의 단

49) 유혜연. "텍스트 스토리에서 이벤트의 감정과 등장인물의 역할 인식." 국내박사학위논문 성균관대학교 일반대학원, 2022. 서울

어와 부정적인 단어, 그리고 중립적인 의미의 단어로 구분되어 있다. 각 단어는 <표 5>와 같이 미리 정의된 감성 점수를 가진다. VADER는 문장 내에서 부정적인 단어와 긍정적인 단어의 등장 패턴을 분석하여 문장의 감성을 예측하는데, 예를 들어 부정적인 단어의 등장이 긍정이나 중립적인 단어의 등장보다 많다면 해당 문장은 부정적인 감성으로 분류 한다.[50]

<표 5> VAER 감성사전의 예시

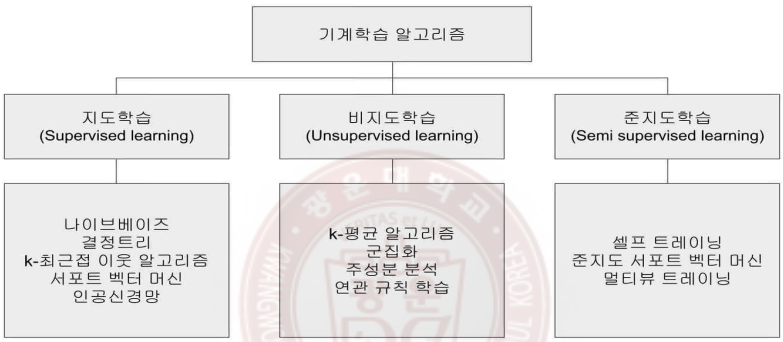
단어	접미사	기능	감정 점수
cutely	-ly	cute의 부사화	1.4
cuteness	-ness	cute의 명사화	2.3
cutenesses	-nesses	cute의 명사화	1.9
cuter	-r	cute의 비교급	2.3
cuteise	-sie	cute의 인격화	1.0
cutesier	-sier	cute의 비교급	1.5
cutesiest	-siest	cute의 최상급	2.2

이때 VADER의 감성사전에는 어간 추출(stemming)이나 원형화(lemmatization)를 진행하지 않았기 때문에 같은 어간을 지닌 단어라도 접미사에 따라 감정의 극성 또는 강도가 변화할 수 있다.

50) 강아미. "VADER와 성향 점수를 이용한 텍스트 분류." 국내석사학위논문 이화여자대학교 대학원, 2021. 서울

제3절 문서분류

자동분류는 1960년대에 시작되어 1980년까지는 전문가에 의해 작성된 규칙을 통한 분류를 주로 사용하였으나 1990년 이후부터는 컴퓨터 성능의 향상과 함께 <그림 9>와 같은 기계학습 기반의 자동분류가 연구되었다 (최윤수, 2019).



* 출처 : Bhavani, Kumar(2021)

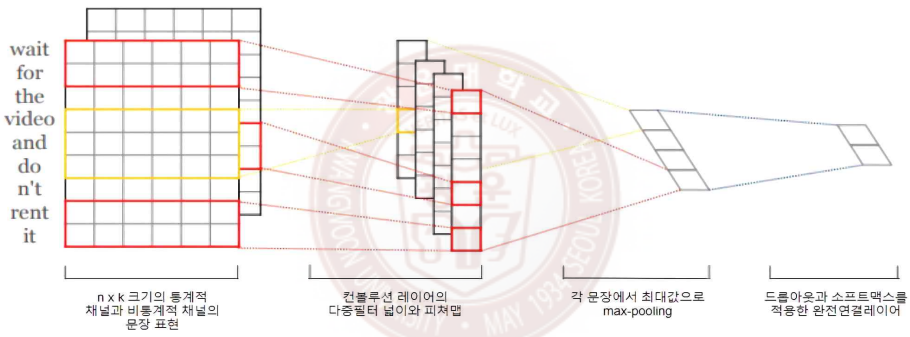
<그림 9> 기계학습 알고리즘

BERT(Bidirectional Encoder Representations from Transformers)는 2018년 구글에서 발표한 모델로, 사전학습 언어모델(pre-trained language model)인 GPT, ELMo 등과 함께 뛰어난 성능을 보이고 있는 모델이다 Devlin et al., 2019.). 책 등과 같은 대용량 말뭉치로부터 비지도 학습으로 미리 학습한 후, 자연어 처리 특정 하위 분야 문제에 지도학습으로 모델을 구성하는 방법을 사용한다.

기존 딥러닝 모델의 자연어 처리는 크게 CNN과 RNN, 워드 임베딩을 주로 사용하여왔다. CNN(convolutional neural network) 알고리즘이 이미

지 처리에 주로 사용되는 반면, 텍스트 데이터 처리에는 순서를 고려한 모델인 순환신경망, 즉 RNN(recurrent neural network)을 사용하는 것이 일반적이다.[51]

그러나 <그림 10>과 같이 CNN 알고리즘으로도 문장 데이터를 모델링 할 수 있는데, Convolution layer를 붙이고 필터로 사용, 입력 데이터 간 자질을 포착하여 추출한 것을 풀링(pooling)과 완전연결층(fully-connected layers)으로 이어 분류를 진행하는 방식을 사용한다.[52]



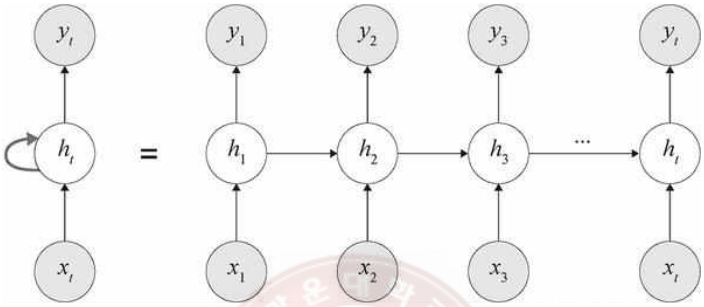
* 출처 : kim(2014.)

<그림 10> 문장 데이터 CNN 알고리즘 모델

RNN은 <그림 11>과 같은 구조를 가지고 있어 비교적 짧은 시퀀스 데이터에서 효과적인데 이는 장기 의존성 문제(the problem of long-term dependencies)를 가졌으나, 데이터를 계산하는 과정에서 각 상태값을 조

51) Bhavani, A. & Kumar, B. S. (2021). A Review of state Art of Text Classification Algorithms, Proceedings of the 25th International Conference on Computing Methodologies and Communication (ICCMC), 1484-1490.
52) 임희석, 고려대학교 자연어처리연구실 (2019). 자연어처리 바이블-핵심이론 · 응용시스템 · 딥러닝. 서울: 휴먼싸이언스

작하는 입력, 망각, 출력게이트를 사용하여 불필요한 연산, 오차 등을 줄이는 LSTM(Long Short Term Memory) 알고리즘을 사용하여 해결할 수 있다. 김정미와 이주홍(2017)은 이를 통해 LSTM의 긴 시퀀스에서의 능력을 입증 하였다.[53]



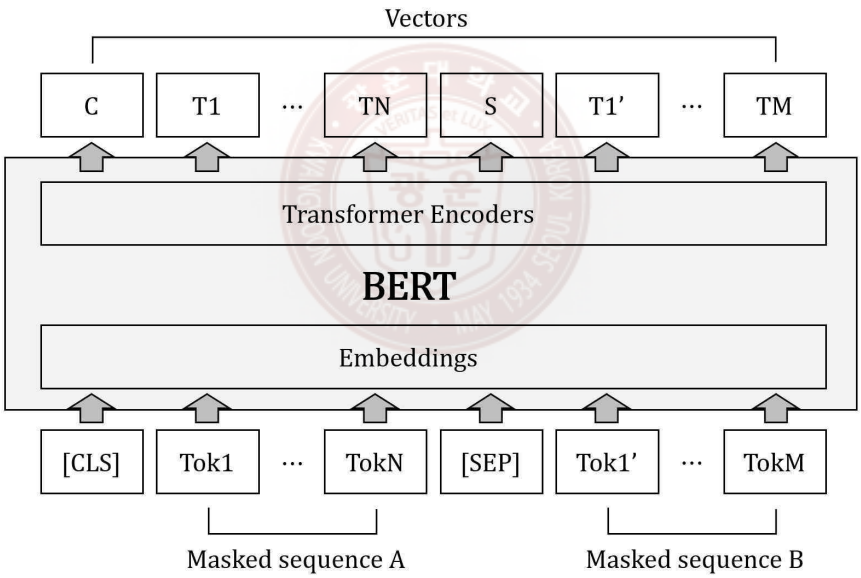
<그림 11> 순환신경망 알고리즘

워드 임베딩(word embedding)은 단어를 백터(vector) 형태로 변환하여 백터 공간에 텍스트를 표현하는 방법으로 Static Word Embedding과 Contextualized Word Representations로 나눌 수 있다. Static Word Embedding은 단일 벡터로 단어의 의미가 변하지 않는 특성을 가지고 있으나, Contextualized Word Representations는 단어표현이 문맥에 반응하여 변화한다는 차이가 있다.[54]

53) 김정미, 이주홍 (2017). Word2vec을 활용한 RNN기반의 문서 분류에 관한 연구. 한국지능시스템학회 논문지, 27(6), 560-565.

54) M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, andL. Zettlemoyer, “Deep contextualized word representations,” in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Jun. 2018, pp. 2227 - 2237.

BERT는 위키피디아나 BookCorpus 등의 범용적인 말뭉치 데이터들을 학습시켜 만든 사전학습모델이 존재하고, 이 모델 위에 출력층을 추가하여 미세조정 과정을 거치는 전이학습 과정을 통해 원하는 분석능력을 갖추게 한다. <그림 12>와 같은 구조로 이루어져 있어, 인코더(encoder)와 임베딩 레이더(embedding layer)로 구성되어 입력 시퀀스를 입력하면 같은 길이의 벡터를 출력한다. 입력 시퀀스는 텍스트를 토큰화한 토큰(token) 시퀀스에 [CLS], [SEP] 토큰을 추가하여 구성한다.[55]

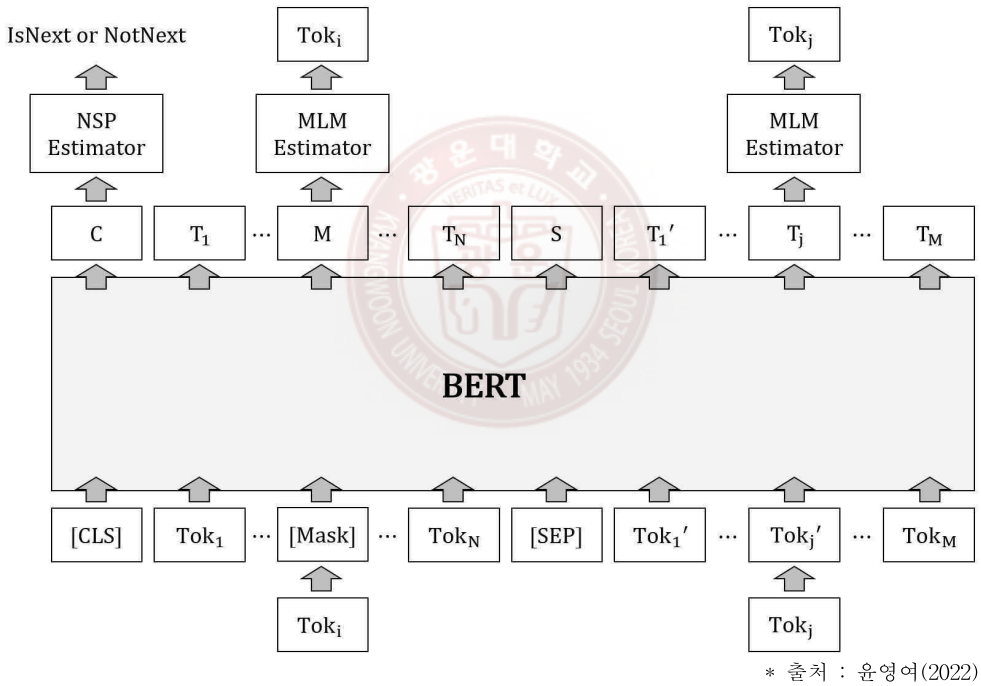


* 출처 : 윤영여(2022)

<그림 12> BERT의 구조

55) Khan, A., Baharudin B., Lee L. H. & Khan K. (2010). A Review of Machine Learning Algorithms for Text-Documents Classification. Journal of Advances in Information Technology, 1(1), 4-20

BERT는 두 개의 입력 토큰을 각 토큰 위치의 출력 벡터로부터 Masked Language Model(MLM), Next Sentence Prediction(NSP) 문제를 해결하도록 학습 한다. 즉, MLM은 토큰의 일부를 [Mask] 토큰 등으로 마스킹하고 마스킹 전의 토큰을 예측한다. NSP는 문장 간의 관계를 이해하기 위해 두 개의 문장이 이어지는 문장인지 예측한다. 이 학습 과정을 표현하면 <그림 13>과 같다.

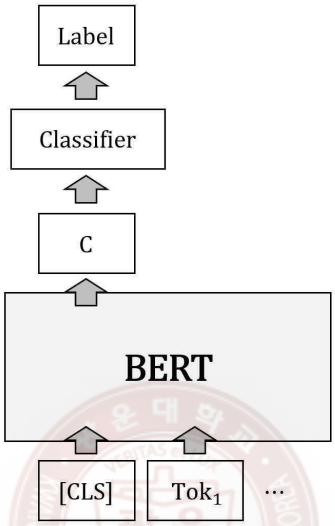


* 출처 : 윤영여(2022)

<그림 13> BERT 사전 학습 과정

이후 <그림 14>와 같은 미세 조정 과정을 거쳐 원하는 학습을 추진할 수 있다. BERT를 활용하여 자연어를 분류하기 위해서는 [CLS] 토큰과 대응하는 BERT의 출력 벡터를 사용하여 입력 시퀀스에 대응하는 레이블

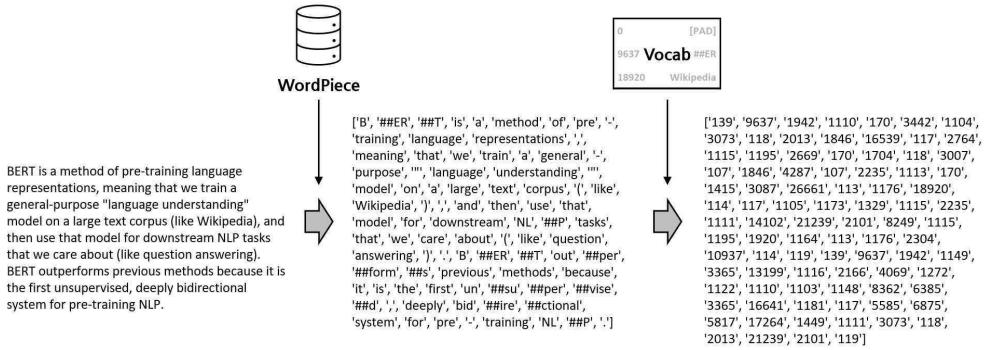
을 예측 하도록 파라미터를 조정하는 절차를 거친다. 이때 BERT를 학습시키기 위해서는 텍스트의 토큰화 과정이 필요하다.



* 출처 : 윤영여(2022)

<그림 14> BERT의 미세 조정 과정

BERT의 토큰화 과정은 <그림 15>와 같이, 텍스트의 화이트스페이스(whitespace)를 스페이스로 변경하고 구두점을 기준으로 분리한 뒤 서브워드 토크나이저(subword tokenizer) WordPiece를 사용한다. WordPiece는 likelihood 기반의 Byte Pair Encoding(BPE) 방법으로 텍스트를 토큰으로 만드는 과정을 학습한다.



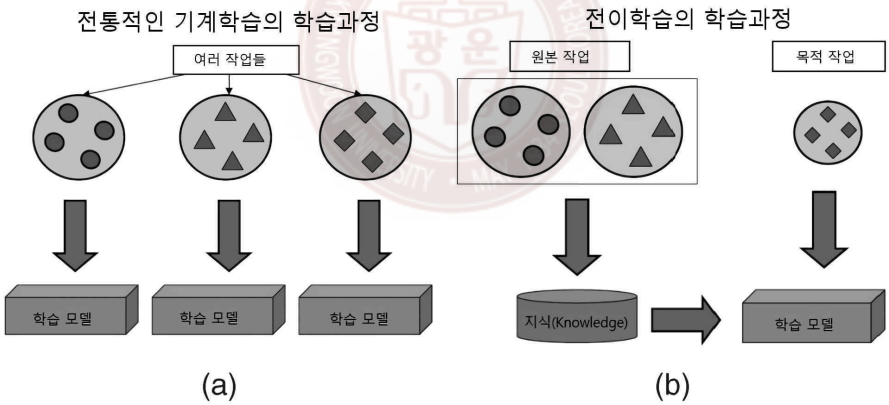
* 출처 : 윤영여(2022)

<그림 15> 텍스트의 토큰화 과정

텍스트를 문자(character) 단위로 분리한 뒤 문자 페어(pair)의 likelihood에 따라서 캐릭터를 페어링(pairing)하여 하나의 문자로 취급한다. 문자를 페어링하는 과정을 일정 조건 동안 반복하여 토크나이저와 토큰 사전(vocab)을 생성한다. 추가적으로 텍스트 데이터 상태에서 화이트스페이스가 아니지만 분리된 경우 토큰에서 텍스트 데이터로 쉽게 복원하기 위해 토큰 앞에 “##”을 추가한다. 예를 들어 “BERT”라는 단어가 WordPiece 토크나이저에 의해 “B”와 “ERT”로 나뉜다면 “ERT” 앞에 “##”을 붙여 “##ERT” 형태의 토큰으로 사용된다. 토큰을 텍스트 데이터로 복원할 때 “##”이 포함된 토큰은 붙이고 다른 경우에는 스페이스를 추가하여 쉽게 복원할 수 있다. Vocab은 토큰과 정수값이 1:1로 대응되는 사전형(dictionary) 데이터로 각 토큰에는 그 토큰과 대응하는 정수값이 존재한다. 생성된 토크나이저와 vocab을 사용하여 텍스트 데이터를 학습 가능한 형태인 정수 타입의 토큰 시퀀스로 토큰화한다

현재 KoBERT, BioBERT, RoBERT 등 BERT를 이용한 다양한 파생연

구가 이루어지고 있어 관련 연구가 활발하게 진행되고 있다.[56] BERT는 대표적인 전이학습을 수행하는데, 전이학습이란 하나 혹은 그 이상의 원본 작업 데이터(source tasks)에서 지식(knowledge)을 추출하여 해당 지식 정보를 목적 작업 데이터(target task)에 적용하여 학습을 진행하는 것을 의미한다. 전통적인 기계학습은 각각의 작업을 처음부터 실행하여 각각의 학습 모델이 존재하지만, <그림 16>과 같이 전이학습은 작업의 목표 데이터를 처리하기 전에 다른 작업 데이터에 해당하는 데이터들을 학습하여 지식을 목표 데이터에 전이하여 학습을 진행할 수 있다는 차이를 가진다.[57]



* 출처 : Pan & Yang(2010)

<그림 16> 기계학습과 전이학습의 차이점

56) 유소엽, 정옥란 (2019) BERT 모델과 지식 그래프를 활용한 지능형 챗봇. 한국전자거래학회지, 24(3), 87-98

57) Pan, S. J. & Yang, Q. (2010). A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345-1359

BERT를 이용하여 텍스트, 즉 자연어 뭉치를 원하는 카테고리로 자동적으로 분류할 수 있는데, 사람이 모두 읽고 이를 분류하는 것에 많은 시간이 소요되는 경우에 유용하다. 김판준(2018)은 국내 학술지들로 문헌 집합을 구성하여 자동분류의 성능을 검토하였고, 학습집합의 크기가 증가할수록 정확도가 향상되는 현상을 확인했다.[58] 박규훤과 정영섭(2021)은 일상대화문을 분류하면서 사전학습 모델별 성능을 비교하였는데 Random Forest, XGBoost, KoBERT를 사용한 결과 KoBERT의 성능이 가장 우수하였음을 확인했다.[59] 박진배(2020) 역시 LSTM(Long-Short-Term Momory), CNN, BERT 관련 사전학습 모델들을 비교한 결과 KoBERT의 성능이 가장 좋았다.[60] 이수빈 등(2021)은 소셜 미디어 텍스트에서 질병을 기준으로 데이터를 자동 분류하는 실험을 진행하였고, 공황장애 말뭉치를 구축하여 경향 문헌을 분류하였다.[61] 최윤수(2019)는 특허 문헌을 수집하여 자동분류 실험을 수행한 결과, 워드 임베딩이 전통적인 기계학습보다 더 뛰어난 성능을 보임을 확인하였다.[62] 김인후(2022)는 학술문헌 자동분류에 BERT를 활용하여 의미 있는 결과를 보여주었다.

선행연구를 통해서 다양한 도구를 활용하여 극성분석과 문서분류의 모델을 개발하였음을 알 수 있었으며, 특히 텍스트 분야의 문서분류에

58) 김판준 (2018). 기계학습에 기초한 국내 학술지 논문의 자동분류에 관한 연구. 한국정보관리학회, 35(2), 37-62.
59) 박규훤, 정영섭 (2021). KoBERT를 사용한 한국어 일상 주제 분류, 2021년 한국컴퓨터종합학술대회 논문집, 1735-1737.
60) 박진배 (2020) 사전훈련 된 모델을 통한 한국어 임베딩 성능 비교, 한국국방기술학회 논문지, 2(3), 1-4.
61) 이수빈,김성덕,이주희,고영수,송민,Lee Soobin,Kim Seongdeok,Lee Juhee,Ko Youngsoo,and Song Min. "딥러닝 자동 분류 모델을 위한 공황장애 소셜미디어 코퍼스 구축 및 분석." 정보관리학회지 38.2 (2021): 153-172.
62) 최윤수 (2018). 기술용어에 대한 분산표현과 딥러닝 모델을 이용한 특허 문헌 자동 분류에 관한 연구. 박사학위논문. 경기대학교 대학원 문헌정보학과.

BERT가 우수한 성능을 보임은 입증되었다. 그러나 첫째, 군이나 방위사업 분야와 관련된 조직에 대해 그 중요성에 비하여 디지털 정체성 관리와 자연어 분석과 관련된 연구는 전무 할 정도로 추진된 연구가 부족하였다. 둘째, 극성분석과 문서분류 모델 개발을 통해 인터넷 기사의 신속성과 전파력을 극복하고자 하는 시도 역시 부족하였다. 셋째, 대부분의 문서분류 모델의 초점이 데이터셋 관리를 통한 예측 정확도를 증가시키기거나 활용분야를 검증하는 것에 맞춰져 있어 현재 실제로 관리되고 있는 데이터들을 인공지능에 학습시켰을 때의 한계점을 확인하고, 이를 개선하기 위한 방안을 도출하려는 노력은 상대적으로 부족했었다고 보여 진다.

제3장 연구의 방법

본 연구의 목적은 현재 방위사업청 담당자가 수행하고 있는 인터넷 기사 중 대응이 필요한 문서를 추출하고, 이 기사에 적절하게 대응하여야 하는 부서를 분류하는 업무를 수행하는 모델을 개발하는 것이다. 더불어, 현재 실무적으로 관리하는 데이터가 인공지능 학습에 적절한지를 확인하고 부족한 점이 있다면 그 개선방안을 확인하는 것이다. 따라서 대변인실에서 관리하는 인터넷 언론을 활용하여 극성분석 모델과 문서분류 모델을 제안하고, 대변인실에서 관리하는 원본 데이터와, 원본 데이터를 인공지능 학습에 적절하도록 정제한 데이터로 나누어 모델의 성능을 비교하였다.

제1절 데이터 정제 및 분류

방위사업청 대변인실에서 담당자가 관리하던 인터넷 기사들의 모음을 원본 데이터라고 하고, 이 데이터를 세 가지 방법으로 정제하여 총 4개 유형의 데이터를 활용하여 연구를 수행하였다. 각 데이터의 활용목적은 다음과 같다.

첫째, 원본 데이터는 실무에서 사용하는 기사의 형태를 유지하는 자료로, 현재 관리상태의 적절성을 확인하기 위해 다른 데이터들과의 비교군으로 활용하였다.

둘째, 정제 데이터는 원본 데이터에서 인공지능 학습을 위해 원본기사의 노이즈와 중복된 기사를 제거하고, 담당부서별 데이터양의 균등성을

고려하여 수행업무를 기준으로 라벨링한 데이터다. 즉, A업무를 수행하는 a팀과 a'팀의 인터넷 기사량의 차이가 크다면 이는 'A'로 라벨링하였다. 그리고 이 데이터는 감성분석과 문서분류 모델 개발의 기본 데이터로 활용하였다.

셋째, 정제 증강 데이터는 정제 데이터에서, 증강기법을 이용하여 학습의 양을 늘린 자료로 학습 데이터 자체의 부족으로 발생할 수 있는 문제를 해결하기 위해 사용하였다.

넷째, 재라벨 데이터는 정제 데이터에서, 기사와 대응부서의 적절성을 재검토 후 라벨링하여 모델개발을 위한 방법으로 활용하였다.

제2절 극성분석 모델개발

대응문서 분류 모델의 개발을 위해서는 어떤 문서, 본 연구에서는 어떤 인터넷 기사가 대응이 필요한지, 즉 중점적인 관리가 필요한지를 먼저 구분하여야 한다. 본 연구에서는 그 기준을 기사가 담고 있는 태도, 즉 감성분석(Sentiment Analysis)의 결과에 있다고 판단하였다. 기사의 특성상 어떤 복잡한 감성으로 이루어진 글이라기보다는 사실과 주장에 대한 내용이기 때문에 긍정과 부정, 즉 극성분석(Polarity analysis)으로 진행하였다. 본 연구의 목적이 디지털 정체성 관리를 위해 사람이 해당 기사를 대응하는 것에 기술적 도움을 주는 것이므로, 긍정적이거나 중립적으로 사실을 보도하는 경우에는 반응할 필요성이 없으며, 부정적인 내용의 기사가 생성된 경우에는 디지털 정체성을 관리하기 위한 어떠한 활동이 요구된다. 또한, 방위사업청 담당부서에 대한 기사의 긍정과 부정의 비율은 몹시 중

요한데, 부정적인 기사의 비율이 높을수록 해당부서에 대한 신뢰도는 하락하기 때문이다.[63]

본 연구에서 분석하는 대상은 공식기사이므로 사전에 구축된 한국어 감성사전을 활용하는 것이 적절하다고 판단하였다. 한국어 감성사전에 대한 연구는 SentiWord나 SentiWordNet을 활용하거나 오픈한글 등을 이용한 실험이 많았으나 이들의 경우는 한국 감성 어휘들의 특징을 잘 반영하지 못한다는 문제점이 있었고, 오픈 한글의 경우 서비스가 종료된 상태였다. 한편, 14,843개의 한국어 감성어휘를 사용한 KNU 한국어 감성사전과 VADER의 영어 감성사전은 활용이 가능한 상태였다.[64] 다만, KNU 한국어 감성사전의 경우 수행 중 본 방위사업과 관련한 인터넷 기사에 대한 분석에 제한이 있었기 때문에 VADER를 활용하여 인터넷 기사의 극성분석을 수행하였다.

VADER의 경우에는 한글 지원 기능이 없기 때문에 본 연구에서는 수집된 기사들을 구글번역(Google Translation)을 활용하여 영어로 번역 후 VADER를 적용하였다. 극성분석과 같은 감성분석 시 한국어를 영어로 번역 후 수행해도 의미 있는 분석결과가 추출된다.[65] VADER는 비지도 감성 분석 라이브러리로 입력한 문장의 감정 지수를 문장 내 단어가 가지는 감정 지수의 조합으로 산출한다. 본 연구의 정제 데이터를 영어로 번역하

63) 심하영, 오수진, “김응모(2018), 감성분석 기반 호텔 리뷰의 특성별 극성분석 및 유저의 선호도 반영 시스템”, 성균관대학교 2018년 춘계학술발표대회논문집, 제25권 제1호

64) 박상민, 나철원, 최민성, 이다희 and 온병원. (2018). Bi-LSTM 기반의 한국어 감성사전 구축 방안. 지능정보연구, 24(4), 219-240.

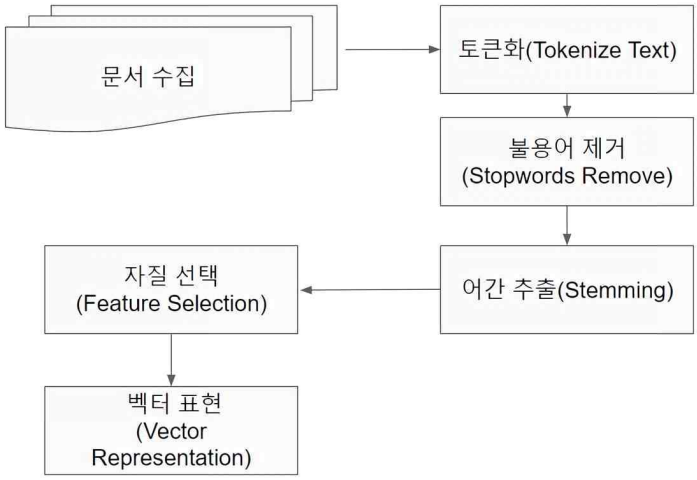
65) 김영민 (Young Min Kim), 정석재 (Suk Jae Jeong), and 이석준 (Suk Jun Lee). "소셜 미디어 감성분석을 통한 주가 등락 예측에 관한 연구." Entrue Journal of Information Technology 13.3 (2014): 59-69.

여 모델에 입력하면 기사 내 문장의 긍정 감정 지수(positive), 부정 감정 지수(negative), 중립 감정 지수(neutral)의 지수를 조합하여 -1~1사이의 감정 지수(compound)가 score로 표현된다. 분석하는 텍스트의 종류에 따라 감정 지수의 임계값을 조정할 수 있지만 기사의 경우 사실관계를 전달하는 것이 목적이므로 긍정적인 기사와 부정적인 기사가 편향성 없이 혼재되어 있을 것으로 판단하여 감정 지수가 0이상 이면 긍정, 미만이면 부정으로 평가한다.

제3절 문서분류 모델개발

본 연구에서 문서분류 모델의 적용대상이 되는 기사의 경우, 일상적인 문장으로 구성되어 있기 때문에 사전학습모델을 사용하는 것이 효율적이고, 일상문으로 구성된 트위터, 영화 리뷰, 논문 등의 분류에서 이미 성능이 확인되었기 때문에 BERT 사전학습모델을 활용한다.

텍스트 기사 데이터를 BERT에 적용하여 분류 모델을 개발하기 위해서는 자연어를 기계가 이해할 수 있도록 토큰화부터 어간 추출까지의 과정을 거쳐서 선택한 자질들을 특정 모델로 학습시키는 작업인 전처리 작업이 <그림 17>과 같이 필요하다.



<그림 17> 텍스트 데이터 전처리 과정

우선, 텍스트를 학습하기 위한 형태인 토큰 시퀀스로 바꾸는 토크나이징이라는 과정을 거친다. 이 BERT의 토큰화 과정은 텍스트의 모든 화이트스페이스(whitespace)를 스페이스로 변경하고 구두점을 기준으로 분리한 뒤 서브워드 토크나이저(subword tokenizer) WordPiece를 사용한다. WordPiece는 likelihood 기반의 Byte Pair Encoding(BPE) 방법으로 텍스트를 토큰으로 만드는 과정을 학습한다. 텍스트를 문자(character) 단위로 분리한 뒤 문자 페어(pair)의 likelihood에 따라서 캐릭터를 페어링(pairing)하여 하나의 문자로 취급한다. 문자를 페어링하는 과정을 일정 조건 동안 반복하여 토크나이저와 토큰 사전(vocab)을 생성한다. 추가적으로 텍스트 데이터 상태에서 화이트스페이스가 아니지만 분리된 경우 토큰에서 텍스트 데이터로 쉽게 복원하기 위해 토큰 앞에 “##”을 추가한다.

예를 들어 “BERT”라는 단어가 WordPiece 토크나이저에 의해 “B”와 “ERT”로 나뉜다면 “ERT” 앞에 “##”을 붙여 “##ERT” 형태의 토큰으로 사용된다. 토큰을 텍스트 데이터로 복원할 때 “##”이 포함된 토큰은 붙이고 다른 경우에는 스페이스를 추가하여 쉽게 복원할 수 있다. Vocab은 토큰과 정수값이 1:1로 대응되는 사전형(dictionary) 데이터로 각 토큰에는 그 토큰과 대응하는 정수값이 존재한다. 생성된 토크나이저와 vocab을 사용하여 텍스트 데이터를 학습 가능한 형태인 정수 타입의 토큰 시퀀스로 토큰화하는 것이다.

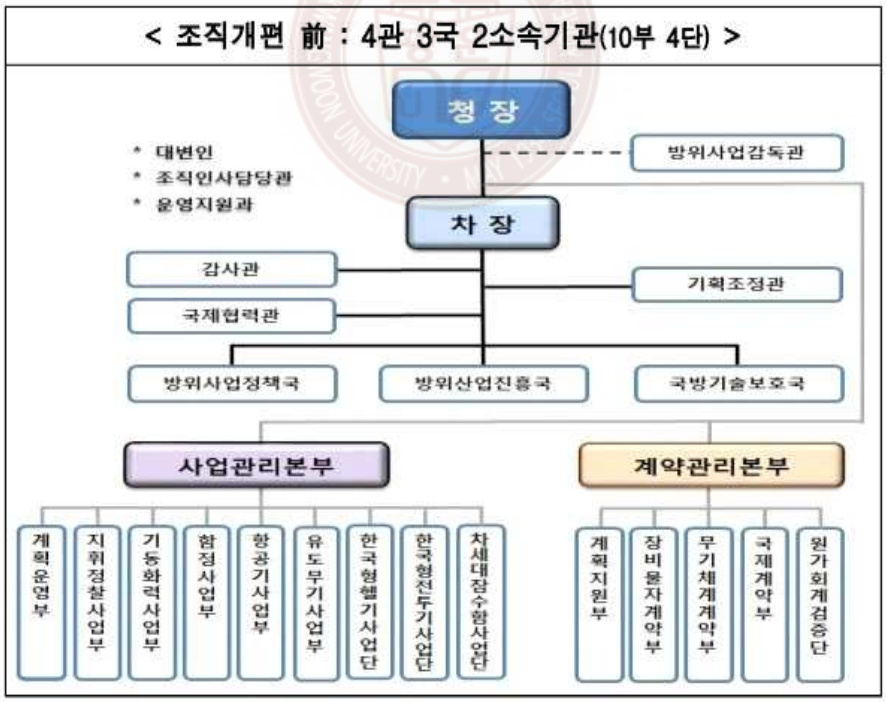
이후 사전 학습된 BERT 모델(model), 최대 시퀀스 길이(max_seq_len), 배치 크기(batch_size)를 입력하여 학습을 위한 입력 시퀀스 데이터를 출력한다. 문서 임베딩 데이터를 생성하기 위해 토큰화, 임베딩, 매칭, 패딩 과정이 필요하다. 토큰화 과정에서 텍스트를 토큰으로 만들고 BERT 모델 입력 형태를 맞추기 위해 앞, 뒤 각각에 [CLS], [SEP] 토큰을 추가하여 정수화하였다. 텍스트의 토큰 수를 BERT 모델에 배치로 입력하기 위하여 부족한 부분에 [PAD] 토큰을 추가하였다. 텍스트를 배치 크기만큼 사전 학습된 BERT 모델에 입력하여 레이블과 매칭하도록 시퀀스를 구성하였다.

제4장 실증연구

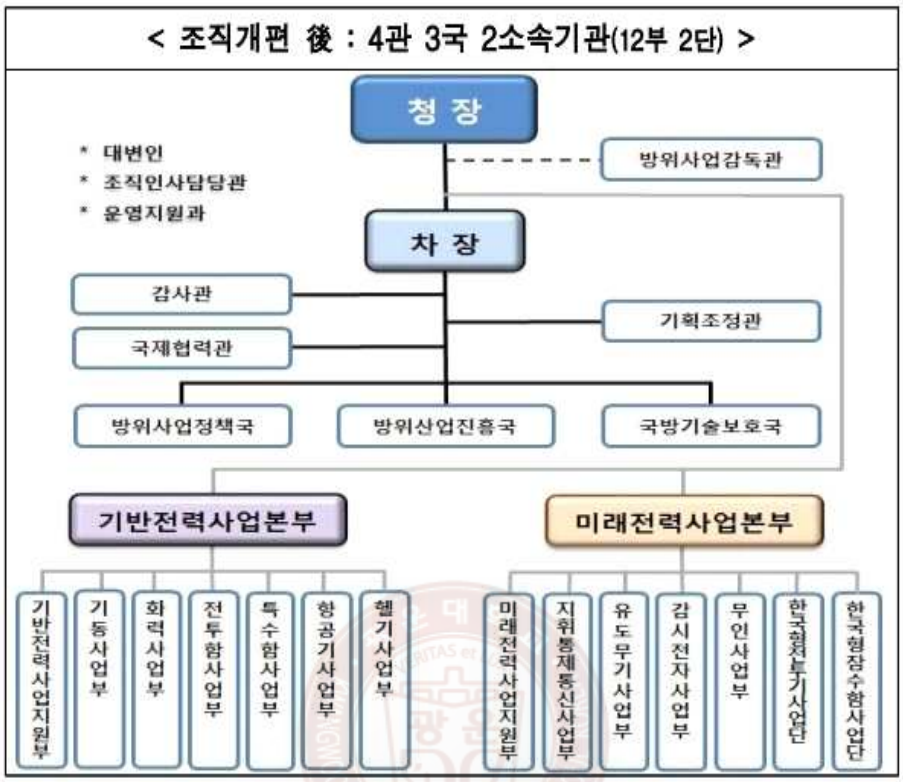
제1절 데이터 정제 및 분류

대변인실에서 기사별 담당부서가 분류된 인터넷 기사 2,153건을 제안 모델개발을 위해 현재를 기준으로 담당부서를 최신화 하였다. 우선, 2016년~2019년 이전 기사의 경우 분류된 담당부서를 2019년 조직 개편 후의 기준에 맞추어 재분류하였다.

<그림 18>과 <그림 19>는 2019년 이루어진 방위사업청 조직 개편 전·후를 보여주고 있다.



<그림 18> 조직 개편 전 방위사업청 조직도



<그림 19> 조직 개편 후 방위사업청 조직도

2019년 조직 개편의 골자는 청본부, 사업관리본부, 계약관리본부로 구성된 방위사업청 조직을 청본부, 기반전력사업본부, 미래전력사업본부로 나누면서 사업본부 내부에 계약부서를 포함한 것이다. 따라서 2019년 이전 계약관리본부 하부의 담당부서는 사업에 따라서 기반전력사업본부 소속의 기반전력사업지원부 또는 미래전력사업본부 소속의 미래전력사업지원부 소속으로 변경하였다.[66]

66) 행정안전부, "방위사업청, 사업관리 중심 조직개편으로 제2의 개청!", 검색일 : 2023. 6. 6, 출처 : https://www.mois.go.kr/frt/bbs/type010/commonSelectBoardArticle.do?bbsId=BBSMSTR_000000000008&nttId=72892

또한 새로운 무기체계 소요에 따라서 통합사업관리팀 및 단이나 부가 추가되기도 한다. 예를 들어 과거에는 잠수함관리팀만 존재하였다면, 장보고-III 사업의 소요가 추가되면서 대한민국 자체적으로 잠수함을 건조할 수 있는 기술을 확보하고 성장시키기 위한 한국형잠수함사업단과 그 소속의 팀들이 만들어졌다. 그리고 사업의 흐름에 따라서 동일한 부서의 명칭이 변경되기도 하는데, 사업 초기에는 무기체계 전력명칭을 사용하였다가 이후 사업이 구체화 되면서 사업명을 사용하는 등이다. 예를 들어 현재 한국형전투기사업단의 경우도, 과거에는 차세대전투기사업단이나 KF-X 사업단으로 그 사업명칭이 분류되어 있어 전수조사 및 최신화 작업을 수행하였다. 한편, 동일한 부서가 수행하는 업무를 더 잘 표현하기 위하는 등의 사유로 명칭을 변경하기도 한다. 예를 들어 과거의 조달기획과가 지금은 계약제도발전과로, 고속함사업팀이 무인수상함사업팀으로 그 역할은 동일하지만 명칭이 변경되었다. 마지막으로, 부서는 동일하나 수행하는 업무가 다시 분장되는 경우도 있다. 예를 들면, 차기 상륙함사업이었던 마라도함 사업은 최초 상륙함사업팀에서 담당하다가 2015년부터는 고속함사업팀에서, 2017년부터는 전투체계 사업팀이, 현재는 다시 상륙함사업팀에서 그 성능개량 사업을 수행하고 있다. 이러한 상황을 반영하여 2016년부터의 자료를 2023년 현재의 상태로 최신화하였고, 그 기준은 해당팀에서 수행하는 업무를 기준으로 하였다.

이렇게 대응할 담당부서를 분류하는 과정에서 확인할 수 있었던 것은 부서의 명칭이 달라지고 업무를 재분장하더라도 방위사업청이라는 조직이 수행하는 큰 형태는 동일하다는 점이었다. 즉, 방위사업청은 감독관 및 감

사관·국제협력관·기획조정관·방위사업정책·방위산업진흥·국방기술보호국이 있는 청본부, 실질적으로 군에서 요구하는 무기체계를 연구개발·구매·임차하는 등의 사업을 관리하는 사업관리본부, 사업관리본부의 요구에 따라 원가를 검증하고 예정가격을 작성하여 계약을 추진하거나, 군의 소요를 직접적으로 계약하는 계약관리 본부의 3축에서 벗어나지 않았다. 2019년 9월 17일 방위사업청의 고유 업무인 사업관리 업무의 효율화를 위해 사업 중심의 조직으로 사업관리와 계약관리업무를 통합하였지만, 그것이 본질적으로 수행하고 있는 업무의 변화를 뜻하는 것은 아니었다. 즉, 계약관리본부가 사업관리본부의 예하의 사업지원부로 편입되고, 사업관리본부는 기동·화력·함정(전투함·특수함)·항공기·헬기를 개발하는 사업팀으로 구성된 기반전력사업본부와, 지휘통제통신·유도무기·감시전자·무인·한국형전투기·한국형잠수함 등을 개발하는 사업팀으로 구성된 미래전력사업본부로 개편되었지만, 결국 수행하는 업무의 중점을 어떻게 두느냐의 차이이지 개청 당시의 청본부·사업관리본부·계약관리본부로 구성되어 수행하던 업무는 현재에도 동일하다. 이는 법에서 명시된 방위사업청의 목표와 역할 범위가 명확하기 때문이다.

<표 6>의 방위사업법의 제1장 총칙에서 확인할 수 있는 바와 같이, 방위사업청이 따르는 방위사업법은 선진강국의 육성과 국가경제 발전에 이바지하는 것을 목적으로 한다. 그리고 이를 위해서 방위력 개선, 방위산업 육성 및 군수품 조달 등의 업무를 경쟁력 있게 수행하고자 하고 있다. 이 세 가지 기본적인 임무는 어떠한 점이 강조되느냐의 차이지 전혀 관계없는 임무를 갑자기 추가로 수행하게 되거나, 근본적인 업무가 없어지는 법

은 없다.

<표 6> 방위사업청의 목적

[방위사업법(법률 제18805호, 2022. 2.3.)]
제1조(목적) 이 법은 자주국방의 기반을 마련하기 위한 방위력 개선, 방위산업육성 및 군수품 조달 등 방위사업의 수행에 관한 사항을 규정함으로써 방위산업의 경쟁력 강화를 도모하며 궁극적으로는 선진강군(先進强軍)의 육성과 국가경제의 발전에 이바지하는 것을 목적으로 한다. 제2조(기본이념) 이 법은 국가의 안전보장을 위하여 방위사업에 대한 제도와 능력을 확충하고, 방위사업의 투명성·전문성 및 효율성을 증진하여 방위산업의 경쟁력을 강화함으로써 자주국방 태세를 구축하고 경제성장 잠재력을 확충함을 기본이념으로 한다.

개편 전후의 차이를 보면, 첫째, 방위력 개선 사업이라는 기본업무를 기존에 사업관리본부와 계약관리본부에서 수행하던 것을 무기체계 특성을 고려하여 기반전력사업본부와 미래전력사업본부에서 통합하였고, 둘째, 방위산업육성을 위한 정책적 검토와 노력은 그대로 방위사업정책국과 방위산업진흥국에서 수행하는 것이다. 셋째로 군수품 조달의 역할 역시 기존의 사업관리본부와 계약관리본부의 수행하던 범위를 그대로 기반전력사업본부와 미래전력사업본부에서 수행하며, 이러한 업무를 위해 각종 지원과 감시의 역할을 하는 부서는 변화된 사항이 없다. 결론적으로 조직의 구성이 달라지더라도 방위사업청이 수행하여야 하는 고유 업무는 고정적이므

로, 대응부서를 지정할 때 지나치게 구체화하는 것은 팀이나 과의 명칭이 달라지는 등의 경우 오히려 데이터 관리의 제한사항으로 작용하였다. 또한 기사가 발생하여 대응하기 위한 부서를 지정할 때에는 팀이나 과의 수준에서 이를 결정할 수는 없다. 방위사업 내부 규정에 따라 외부 언론에 대응하기 위해서는 적어도 본부장의 승인을 받아야 하는 사항이기 때문이다.[67]

그러나 현재 대변인실에서 관리하고 있는 인터넷 기사와 이를 대응한 담당부서를 정리한 데이터를 확인하였을 때, 담당 부서가 과와 팀의 단위로 나누어져 있었다. 즉, 본부급에서 어떤 부서에서 관련 내용을 대응하는 것이 적절한지 판단하여 최종적으로 대응한 과나 팀을 대응부서로 표현한 것이다. 따라서 상대적으로 인터넷 기사 그리고 언론에서 주목받는 사업과 그렇지 않은 사업 간의 인터넷 기사 건수의 편차가 컸고 당시 각종 상황으로 인해 동일한 사안이라도 다른 부서에서 대응한 경우도 많았다. 예를 들어 방위사업청장에 대한 비판적인 기사를 대응할 때 대부분은 대변인실에서 대응하였으나, 20년도에는 감사관실에서 대응하였다. 이는 당시 방위사업청장이 감사원 사무총장 출신으로, 감사관실을 잘 활용하였기 때문이지 인터넷 기사의 내용에 따른 것은 아니었다. 따라서 사실 대변인실에서 관리하는 데이터만으로는 대응부서를 분류하기 위한 인터넷 기사에 대한 명확한 기준을 제시하는 것은 제한되었다. 이러한 문제를 보완하기 위해 팀과 과의 단위로 정리되어있는 기사 대응부서에서, 방위사업청이 수행하는 업무의 성격이 유사한 팀과 과는 그 상위 개념인 부와 단의 단

67) 방위사업청훈령 제591호(2020.4.21.) 방위사업 홍보 규정

위로 통합하였다. 예를 들어 사업감사담당관실, 공직감사담당관실로 나누어져있다면 감사관실로 통합한 것이다. 또한, 출연 기관인 국방기술품질원과 국방과학연구소는 방위사업청을 중심으로 종합된 기사이므로 기관 자체로 통합하였다. 그 결과 <표 7>에서 볼 수 있듯이 전체 기사를 총 22개의 관련 부서로 분류할 수 있었다.

<표 7> 원본 데이터

구 분	대응부서	데이터
1	감사관실	124
2	감시전자사업부	85
3	국방과학연구소	105
4	국방기술보호국	66
5	국방기술품질원	114
6	국제협력관	145
7	기동사업부	107
8	기획조정관	120
9	대변인실	78
10	방위사업정책국	128
11	방위산업진흥국	215
12	우주지휘통신사업부	47
13	운영지원과	24
14	유도무기사업부	126
15	조직인사담당관	64
16	첨단기술사업단	34
17	한국형잠수함사업단	36
18	한국형전투기사업단	103
19	함정사업부	93
20	항공기사업부	179
21	헬기사업부	136
22	화력사업부	24
계		2,153

이 첫 번째 데이터는 대변인실에서 제공한 데이터를 최대한 유지하였다. 이는 단순히 대응부서를 지정한 내용, 즉 라벨링의 변화를 최소화한다는 의미만을 뜻하지 않는다. 인터넷에 있는 기사를 그대로 사용하기 때문에 기자의 소개 정보, 기사의 출처, 그림 소개 정보 등 딥 러닝에 노이즈로 작용할 수 있는 텍스트가 있지만 이를 제거하지 않았다. 이는 자연어 처리의 모델링에는 데이터의 관리가 중요한 요소인데, 현재의 상태와 모델링을 위한 데이터 수정의 차이가 어떻게 작용하는지에 대한 차이를 보기 위함이다.

다음에는 기계학습을 위해 기사별 한 개의 대응부서를 가질 수 있도록 정리하였다. 예를 들어 방위사업의 추진에 대한 의사결정을 하는 회의인 방위사업추진위원회의 결과를 다루는 기사에서, A, B, C 사업을 다루었다면 대변인실은 방위사업추진위원회를 관리하는 부서와, A, B, C 사업 부서를 모두 지정하여 한 개의 데이터에 총 4개의 라벨링이 설정된 상태가 되어 있었다. 이 경우에는 해당 기사에서 각 사업부서와 연계된 내용을 추출하여 기사를 분류하거나, 특별한 내용이 없다면 방위사업추진위원회를 관리하는 부서로 라벨링 하였다. 동일하게 A라는 사업추진의 결함 등을 꾸짖는 기사가 있다면 A 사업부서와 감사관실이 모두 지정되어 있었다. 이러한 경우에는 한 기사에 한 사업만 있는 경우에는 해당 사업 부서를 기사의 대응부서로 구분하였고, 복수의 사업이 포함된 경우에는 감사관실을 대응부서로 구분하였다. 그 결과 전체 데이터의 개수가 2,150개로 변경되었다. 또한 딥 러닝 학습의 특성상 정확도 향상을 위해서는 라벨(관련부서)당 많은 데이터(기사)를 가지고, 편차가 적은 것이 바람직하기

때문에, 추가적인 학습을 위해 통합 가능한 명칭으로 <표 8>과 같이 정리하였다. 즉 22개 담당부서를 9개 통합부서로 묶은 것이다. 이는 최대한 실무와 딥 러닝연구의 균형을 맞추기 위함이었다.

<표 8> 정제 데이터

구 분	대응부서	대응부서 정리결과	데이터
1	함정사업부	해상무기 및 헬기사업	265
	한국형잠수함사업단		
	헬기사업부		
2	항공기사업부	공중무기사업	282
	한국형전투기사업단		
3	기동사업부	지상 및 유도무기사업	257(-1)
	화력사업부		
	유도무기사업부		
4	첨단기술사업단	첨단무기사업	166
	우주지휘통신사업부		
	감시전자사업부		
5	방위산업진흥국	방위산업진흥	212(-2)
6	방위사업정책국	방위사업정책	330
	감사관실		
	대변인실		
7	기획조정관	방위사업기획	208
	조직인사담당관		
	운영지원과		
8	국제협력관	방위사업수출	211
	국방기술보호국		
9	국방기술품질원	출연기관	219
	국방과학연구소		
계			2,150

제2절 극성분석을 활용한 대응기사 분류 모델

대변인실로부터 제공받은 기사 2,153건 중 중복기사를 제거하고, 추후 대응부서 분류모델 개발을 위해 관련기사 당 대응부서를 매치시키는 작업을 거친 2,150건을 9개의 대응부서로 분류한 기사를 모은 정제 데이터를 이용하였다. 문장 내 불용어 제거 및 명사만 추출하여 감성용어로 사용 후 실험을 수행하였다.

우선, 서비스를 제공하고 있는 KNU 한국어 감성을 사용하고자 하였으나, 이는 기대와 다르게 표준국어대사전으로 작성하였음에도 불구하고 방위사업 관련 기사로 샘플링 테스트를 진행하였을 때 <표 9>와 같이 분석 결과가 대부분 산출되지 않았는데, 이는 통상적인 문장과 방위사업과 관련한 기사에 사용되는 단어의 차이에서 기인한 것으로 판단되었다.

<표 9> KNU 한국어 감성사전 테스트 결과

word : 정부 소식통은 “방위사업에 대한 이해 없이 법리적 검토에만 치중할 경우 사업이 원활하게 이뤄지지 않을 것이라는 우려도 있다”고 말했다. 어근 : None 극성 : None (‘None’, ‘None’)
word : 국방부 고위 관계자는 “방위사업청에 별도의 감사부서가 있고 감사원 직원도 파견돼 있지만 방위사업 비리를 막을 순 없었다”면서 “방사청장 직속인 방위사업감독관이 얼마만큼 전문성과 자율성을 확보하느냐가 중요할 것”이라고 말했다. 어근 : None 극성 : None (‘None’, ‘None’)

따라서, 영어 감성사전을 보유하고 있는 VADER의 활용을 추진하였다.

VADER의 감정사전은 긍정의 단어 3,345개, 부정의 단어 4,172개가 있다. 정확한 문장을 지향하는 기사의 특성상 영어로 번역하였을 때 문장해석에 문제가 없었기에 적절한 영어단어가 선택되었다고 판단하였으며, 해당 모델은 <표 10>과 같이 극성분석을 수행한다.

<표 10> 번역 후 VADER를 활용한 극성분석

구 분	세부내용	감성수치
한국어 기사	4월 1일 대법원은 이규태 일광공영 회장의 방위산업(방산) 비리에 대해 무죄를 선고한 원심을 확정했다. 이 회장의 죄목은 납품 사기. 1000억 원대 공군 전자전훈련장비(EWTS)를 도입하면서 원가를 부풀려 부당이익 500여억 원을 챙겼다는 혐의였다...(생략)	-
⇩		
영어 번역	On April 1, the Supreme Court upheld the lower court's verdict of not guilty of corruption in the defense industry (defense industry) of Lee Kyu-tae, chairman of Ilgwang Corporation. Chairman Lee's crime is delivery fraud. They were accused of making unfair profits of 50 billion won by inflating the cost of introducing 100 billion won Air Force Electronic Warfare Training Equipment (EWTS)...	-0.9992

VADER를 통해 극성분석을 수행한 한 결과 긍정(중립)은 1,555건, 부정은 595건(27.6%)으로 분석되었다. 모델 세부 실험결과는 <표 11>과 같다.

<표 11> 대응기사 분류 모델 실험결과

구 분	분류기준	기사수	감성평균	대응필요 기사
1	해상무기 및 헬기사업	265	0.01	130
2	공중무기사업	282	0.53	58
3	지상 및 유도무기사업	257	-0.04	134
4	첨단무기사업	166	0.09	73
5	방위산업진흥	212	0.59	38
6	방위사업정책	330	0.46	81
7	방위사업기획	208	0.62	27
8	방위사업수출	211	0.74	22
9	출연기관	219	0.65	32
계		2,150	0.41	595

감성평균을 기준으로 지상 및 유도무기사업(-0.04)은 디지털 정체성을 관리하기 위해 가장 언론에 대응해야 할 필요성이 크게 나타났고, 반면에 방위사업기획(0.62), 출연기관(0.65), 방위사업수출(0.74)은 디지털 정체성이 잘 관리되고 있는 것으로 보여 졌다. 다만 분석 자료는 2016년부터 2022년까지의 자료를 기반으로 한 것으로, 해당 기간 동안의 평균값으로 현재의 상태를 의미하는 것은 아니다.

제3절 대응부서 분류 모델

1. 원본 데이터 실험

문서분류 실험은 BERT를 통해 수행되었으며, 정제한 데이터들의 성능 차이를 통해 데이터 관리의 방향 등을 확인하고자 하였다. 학습과 테스트 데이터의 비율은 80:20으로 진행되었으며, 배치사이즈(Batch size)는 3, 학습 횟수(epoch)은 100으로 수행하며, 실험은 1절에서 분류한 데이터를 기반으로 총 4회에 걸쳐 진행된다. 문서분류의 성능을 평가할 때는 일반적으로 정확도(accuracy), 재현율(recall)과 조화평균 척도(F1 score)를 사용하고(Bhavani, Kumar, 2021), 자료가 불균형한 상태에서는 정밀도(precision), 재현율(recall), 조화평균 척도를 사용한다(Manning, Raghavan & Schutze, 2010.).[68] 여기서 정밀도, 재현율, 조화평균 척도란 다음과 같다.

모델이 문서분류를 수행하면서 해당 문서가 참이고, 참으로 분류하면 True Positive, 거짓으로 잘못 분류하면 False Negative이고, 반대로 해당 문서가 거짓이고, 거짓으로 분류하면 True Negative이고 실제 참일 경우에는 False Positive에 해당한다. 이를 종합적으로 표현하면 <표 12>와 같다.

68) Manning D. C., Raghavan P. & Schutze H. Introduction to Information Retrieval. 안동언, 김재훈, 남영준, 박혁로, 이상곤 공역 (2010). 최신 정보검색론. 경기도:교보문고

<표 12> 이진 분류 결과

구 분	실제 참	실제 거짓
예측 참	True Positive	False Positive
예측 거짓	False Negative	True Negative

이때 True Positive를 TP, False Positive를 FP라고 할 때, 정밀도인 Precision은 식 (3-1)과 같이 산출된다. 즉, 예측값을 중심으로 평가한 내용이다.

$$\mathrm{Pr}ecision = \frac{TP}{TP+FP} \tag{3-1}$$

한편, False Negative를 FN이라고 하면, 재현율이 Recall은 식 (3-2)에 의해 산출된 값을 의미 한다. 정밀도와 달리 실제 값을 중심으로 평가한 것이다.

$$Recall = \frac{TP}{TP+FN} \tag{3-2}$$

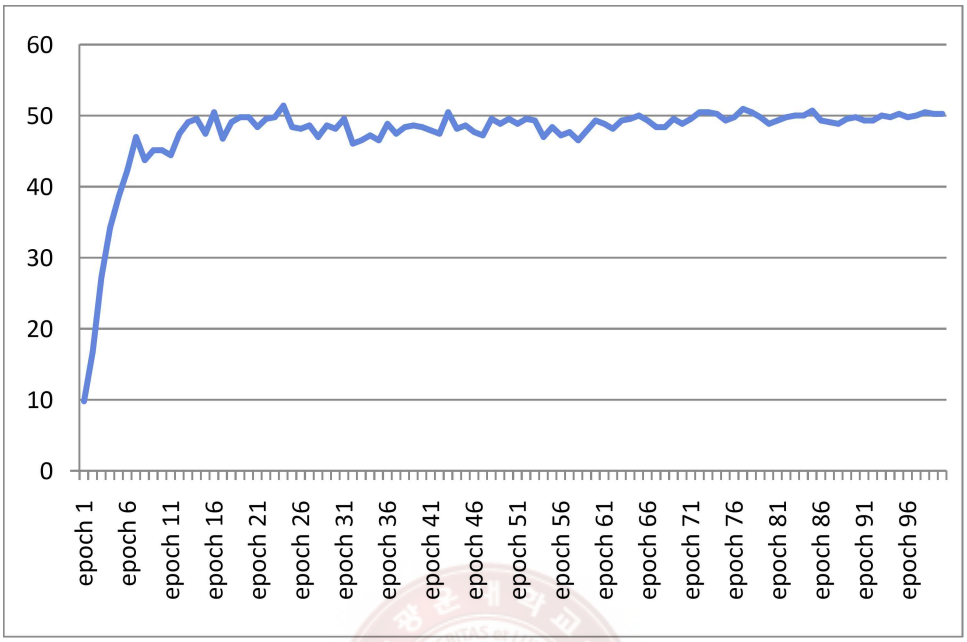
식 (3-1)과 (3-2)에 따라 조화평균 척도인 F-1 score는 다음과 같다. 즉, 정밀도와 재현율 양쪽의 평가지표의 조화평균으로 불균형한 데이터를 수치로 평가한다.

$$F-1\,score = \frac{2 \times \mathrm{Precision} \times Recall}{\mathrm{Precision} + Recall} \quad (3\text{-}3)$$

따라서 실험에서는 분류 정확도를 확인하고 정밀도, 재현율, 조화평균 척도를 제시하여 모델을 평가하였다. 이때, 학습 횟수는 loss 값에 변화가 없을 때 까지 수행하는 것을 목표로 진행하였다. 이를 통해 현재 실제로 관리하고 있는 데이터들이 인공지능 학습에 적절한지 여부를 평가하고, 관리하는 데이터들이 의미 있는 학습결과를 도출하기 위해서는 어떠한 노력이 필요한지 등을 도출하고자 하였다.

원본 데이터인 2,153개의 데이터를 22개 대응부서로 라벨링한 결과를 그대로 활용하여 실험을 진행한 결과는 <그림 20>과 같다. x축은 학습 횟수(epoch)이고, y축은 분류 정확도(accuracy)이며, 이는 최대 50.1%로 나타났다. 22개 대응부서 중에서 운영지원과나 화력사업부의 경우에는 관련기사가 24개인 반면, 방위산업진흥국의 경우 관련기사가 215개인 등 데이터의 불균형이 심하기 때문에 학습의 효율이 떨어질 수밖에 없는 상황이었다.

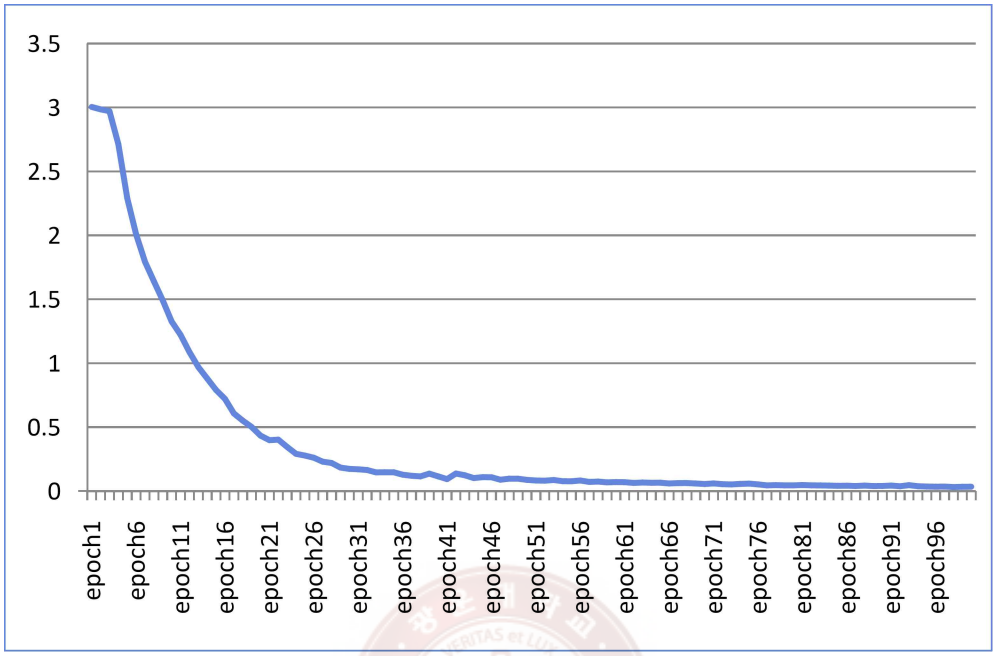
(단위 : %)



<그림 20> 원본 데이터의 대응부서 분류 정확도

학습 횟수가 12를 넘어간 경우부터 정확도가 50% 정도로 상승하였다가 거의 변동이 없었고, <그림 21>과 같이 training loss가 0.03 이하로 변동이 없음을 고려하여 충분한 학습되었다고 판단된다.

(단위 : %)



<그림 21> 원본 데이터의 대응부서 분류 학습 손실

일반적으로 조화평균 척도와 평균적인 정확도가 50% 이상인 경우 의미 있는 분류 성능을 가진다고 판단한다.[69] 그러나 본 실험과정에서 정확도는 최대가 50.1%이고 조화평균 척도 역시 0.49로 의미 있는 성능을 가지고 있다고 할 수 없었다.

<표 13> 원본 데이터의 대응부서 분류 실험결과

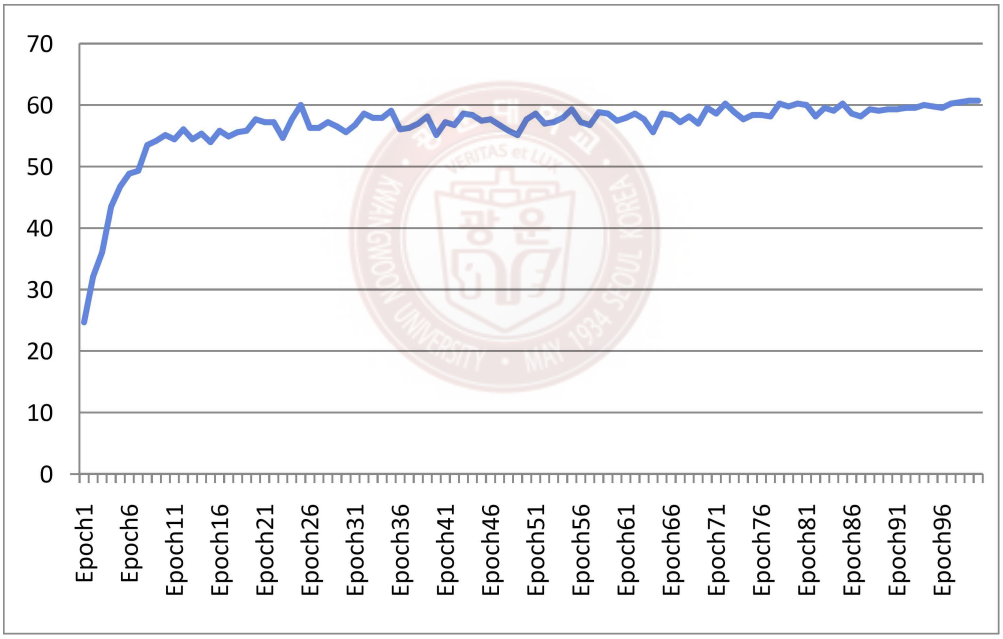
재현율	정밀도	F-1 score
0.5011	0.5057	0.4985

69) 황상흠, 김도현 (2020). 한국어 기술문서 분석을 위한 BERT 기반의 분류모델. 한국전자거래학회지, 25(1), 203-214

2. 정제 데이터 실험

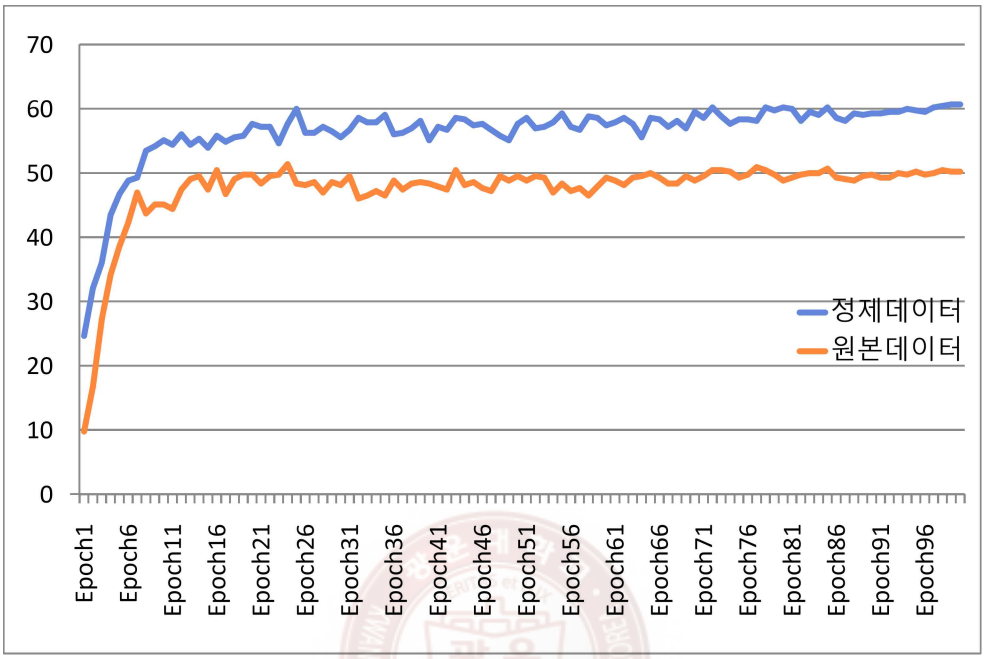
원본 데이터를 일부 교정하고 중복된 기사를 제거한 2,150개의 기사를 9개의 대응부서로 정리한 정제 데이터를 이용하여 실험을 추진한 결과, 분류 정확도는 <그림 22>와 같이 최대 60.7%이며, <그림 23>과 같이 원본 데이터 대비 약 10.6%가 상승하였음을 확인하였다. 조화평균 척도 역시 0.57로 대응부서 원본 데이터에 비해 향상된 성능을 보여주었다.

(단위 : %)



<그림 22> 정제 데이터의 대응부서 분류 정확도

(단위 : %)



<그림 23> 원본 및 정제 데이터의 대응부서 분류 정확도 비교

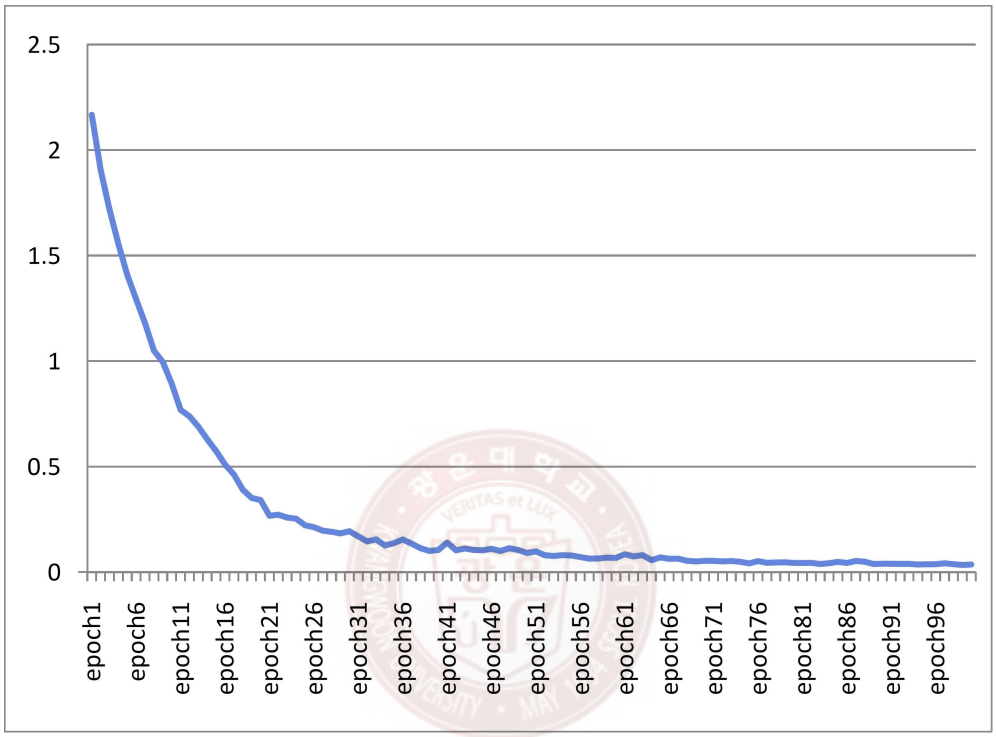
<표 14>와 같이 원본 데이터에 비하여 정제 데이터의 정확도를 포함한 등 전반적인 수치를 비교한 결과 모두 높게 나타났음을 확인할 수 있어, 문서분류의 성능이 향상되었음을 확인할 수 있었다.

<표14> 원본 및 정제 데이터의 대응부서 분류 실험결과

구 분	재현율	정밀도	F-1 score
원본 데이터	0.5011	0.5057	0.4985
정제 데이터	0.6069	0.6025	0.6028

<그림 24>와 같이 실험 중 training loss를 고려하여 0.03 이후 변화가 없어 epoch를 100까지 수행하였다.

(단위 : %)



<그림 24> 정제 데이터의 대응부서 분류 학습 손실

정제 데이터의 실험결과가 원본 데이터의 실험결과에 비하여 분류 정확도가 10.6% 향상되었다는 점을 고려할 때 최소한의 데이터 관리가 필수적이라고 판단된다. 다만, 타 연구에서 확인된 BERT의 성능을 고려할 때 60.7%의 정확도는 높다고 할 수 없다. 따라서 이것이 학습할 데이터의 양이 부족해서 발생한 문제인지, 그 외의 데이터 관리 차원 등에서 검토되어야 할 문제인지 등의 확인이 필요하였다. 따라서 추가실험으로 기사의

데이터 자체 수량을 증가시키는 text augmentation을 수행 후 해당 데이터를 이용하여 학습을 진행하였다.

3. 증강 정제 데이터 실험

본 연구의 데이터는 대변인실에서 제공한 기사 모음 중 중복기사 및 노이즈를 제거한 2,150건의 기사가 기준이 된다. 이 데이터의 수량은 정제 데이터를 기준으로 라벨별 평균 238.8건으로 딥 러닝을 위한 충분한 분량이 아닐 수 있었다. 매개변수를 훈련할 충분한 학습 데이터를 확보하지 않으면 개발한 모델이 훈련 데이터에만 지나치게 적응하여 테스트 데이터 또는 새로운 데이터에는 반응하지 못하는 과적합 현상의 발생하는 등에 따라 정확도 향상을 방해 할 수 있다. 그러나 기사를 대량 확보하기 위해서는 시간이 소요되고, 이는 대변인실에서 수집할 가치가 있어 분류가 이루어질 것이라고 기대할 수 없다. 따라서 같이 랜덤으로 훈련 데이터에 인위적인 변화를 가하는 어그멘테이션(augmentation)이 효과를 발휘할 수 있다.

<표 15>와 같이 텍스트의 경우, 특정 단어를 유의어로 교체하는 SR, 임의의 단어를 삽입하거나 삭제라는 RI 및 RD, 문장 내 임의의 두 단어의 위치를 바꾸는 RS 등을 수행할 수 있다.[70]

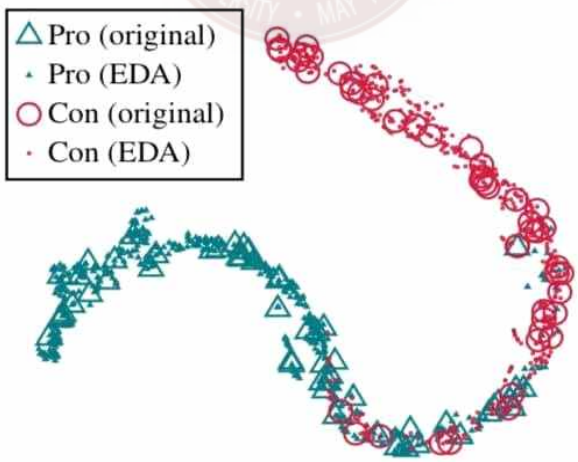
70) Jason Wei, Kai Zou, Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks(2019)

<표 15> 텍스트 어그멘테이션 기법

Operation	Sentence
None	A sad, superior human comedy played out on the back roads of life.
SR	A **lamentable**, superior human comedy played out on the **backward** road of life.
RI	A sad, superior human comedy played out on **funniness** the back roads of life.
RS	A sad, superioir human comedy played out on **roads** back **the** of life.
RD	A sad, superior human out on the roads of life.

* 출처 : Jason Wei(2019)

해당 기술적용을 통해 실제로 성능향상을 유도할 수 있고, 어그멘테이션 과정을 통해 생성된 문장들은 <그림 25>와 같이 본래의 라벨 성질을 잘 따른다는 것이 확인되었다.



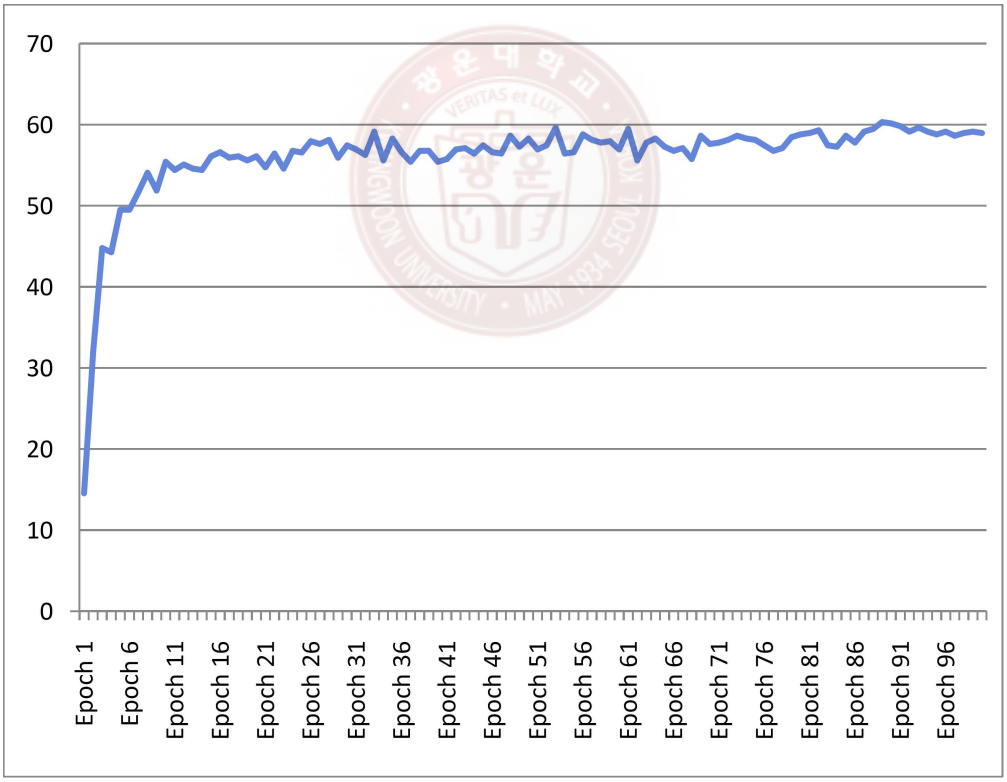
* 출처 : Jason Wei(2019.)

<그림 25> 어그멘테이션 결과

정제 데이터를 기준으로 어그멘테이션을 추진한 결과 2,150개의 기존 텍스트 데이터에 랜덤 생성한 810개(37.7%)의 문서가 추가된 데이터를 확보하여 실험을 진행하였다. 기존의 9라벨 분류모델이었던 정제 데이터를 어그멘테이션한 결과로 이를 증강 정제 데이터로 구분하였다.

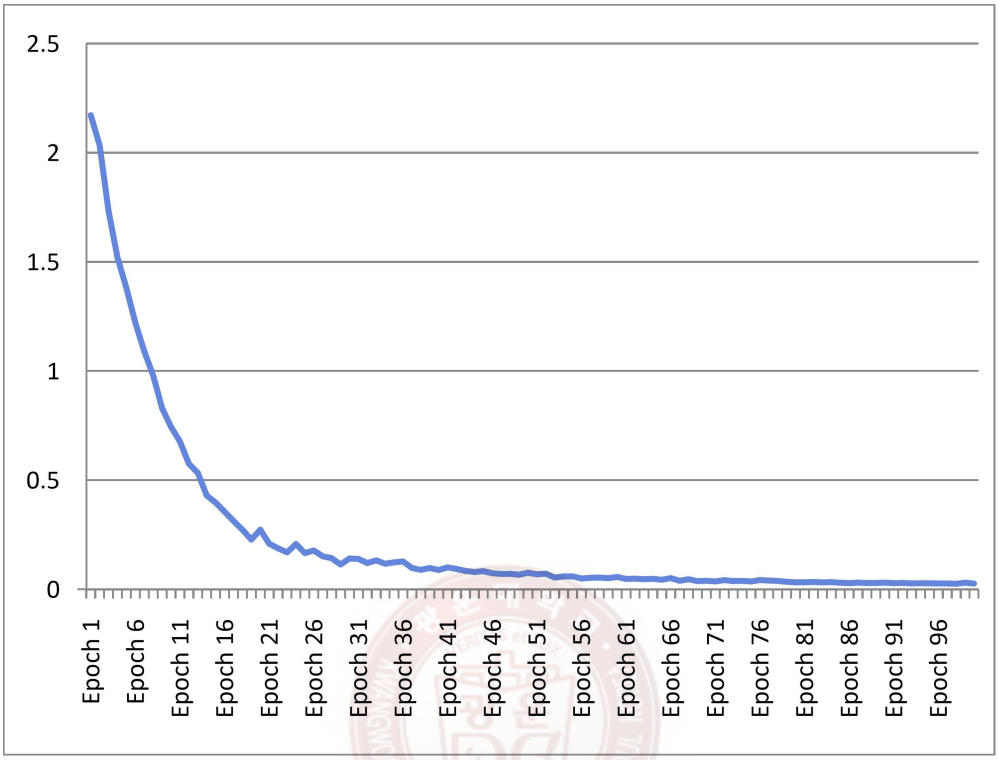
증강 정제데이터를 적용한 결과, <그림 26>과 같이 분류 정확도는 60.1%로 확인되었고 <그림 27>과 같이 training loss가 0.03에서 변화가 없는 점을 고려 epoch100 에서 실험을 종료하였다.

(단위 : %)



<그림 26> 증강 정제 데이터의 대응부서 분류 정확도

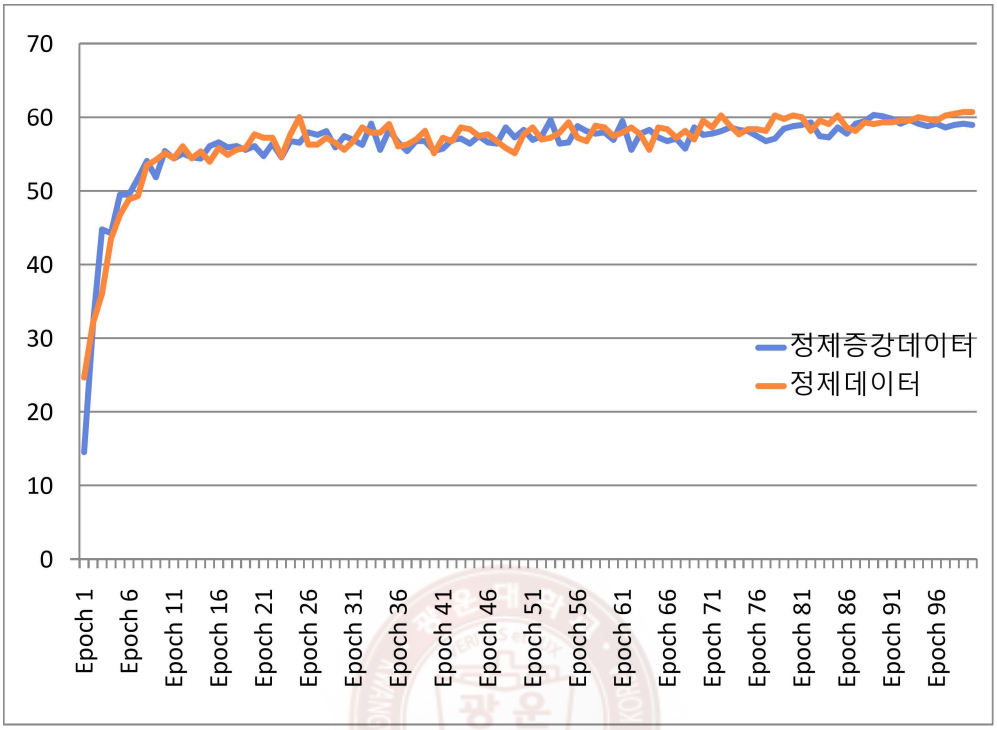
(단위 : %)



<그림 27> 증강 정제 데이터의 대응부서 분류 학습 손실

정제 데이터 실험결과와 증강 정제 데이터의 실험결과를 비교하였을 때, 데이터가 37.7%가 증가하였음에도 불구하고 <그림 28>과 같이 분류 정확도는 의미 있게 증가하지 않았다는 점이 주목되었다.

(단위 : %)



<그림 28> 정제 및 증강 정제 데이터의 대응부서 분류 정확도 비교

기사의 노이즈를 제거한 데이터를 가지고, 이 데이터의 양을 증가시켜 학습시켰음에도 불구하고 <표 16>과 같이 성능이 향상되지 않아 원본 데이터 및 정제 데이터 자체를 검토한 결과 다음의 문제점이 발견되었다.

<표 16> 원본·정제·증강 정제 데이터의 대응부서 분류 실험결과

구 분	재현율	정밀도	F-1 score
원본 데이터	0.5011	0.5057	0.4985
정제 데이터	0.6069	0.6025	0.6028
증강 정제 데이터	0.6013	0.5962	0.5946

첫째, 유사한 내용의 기사라 하더라도 관리해야 하는 대응부서를 다르게 지정한 경우가 다수 발견되었다. 이는 사업부서보다는 청본부, 통상 정책부서에서 두드러지게 나타났는데, 사유는 다음과 같이 분석되었다. 대변인실로부터 제공받은 데이터는 해당기사가 실제로 확인되었을 때, 당시 담당자의 판단 및 관계부서 협조결과 결정된 대응부서를 기입한 내용인데, 현실적으로 특정 정책적 의사결정이나 추진내용에 대하여 조직도에 명시된 업무분장에 따라서 담당하지 않기도 하고, 기사의 내용으로만 볼 때 명확하게 분장할 수 없는 내용이 다수 있었기 때문이다. 따라서 당시의 부서별 업무량 등 상황 및 부서장의 입장에 따라서 담당부서가 결정되는 경우가 많아 기사의 내용과 담당부서의 라벨링 상태의 기준이 모호했다.

둘째, 대변인실을 관리하는 관리자는 시간에 따라 변화하고, 관리자의 성향에 따라 관리하는 기사의 수준이 몹시 다르다는 점이다. 언론에 민감한 관리자의 경우에는 단순한 행사의 개최를 알리는 기사까지도 수집, 관리하는가 하면 어떠한 관리자는 그렇게 하지 않는다. 따라서 민감한 관리자와 그렇지 않은 관리자의 성향 차이에 따라 기사의 관리상태가 크게 달랐고, 이는 학습 패턴에 부정적인 영향을 끼친 것으로 판단되었다.

방위사업청에서 인터넷 기사에 대하여 대응부서를 분류하는 기준이 명확하지 않기 때문에, 김인후(2022)의 연구 중 정확도가 50%에 불과한 문헌정보학의 경우와 같다고 수 있다. 그러나 본 연구에서 정제 데이터 실험결과 60.1%의 정확도를 보이고 조화평균 척도 역시 0.6028로 의미 있는 모델이라고 판단된다.

4. 재라벨 데이터 실험

현재 사업부서의 경우에는 대부분 획득하는 무기체계에 따라 담당 부서가 구분되고, 청본부라 하더라도 정책적인 내용과 수출 및 기술보호 등에 대한 내용은 일정 수준 구분이 가능하였다. 또한, 출연 기관인 국방과학연구소와 국방기술품질원의 경우 역시 타당하게 분류가 되어있다고 보여졌다. 따라서 대변인실에서 분류한 라벨링을 유지하면서, 상위분류의 라벨링을 통해 데이터를 정비할 수 있었다. 이를 정리한 내용은 <표 17>과 같다.

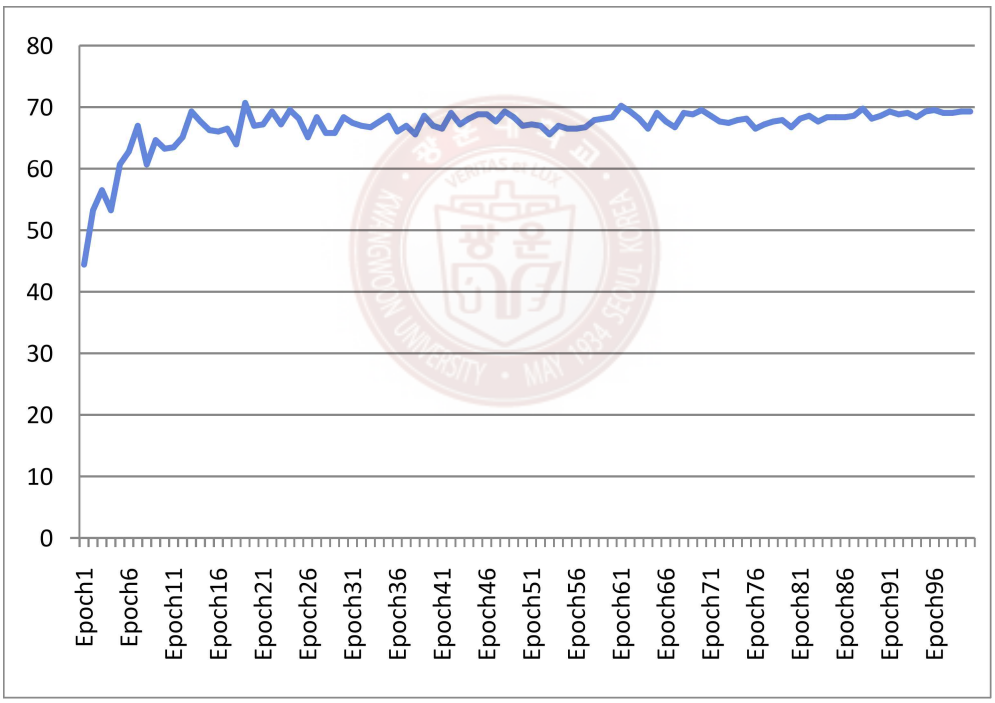
<표 17> 재라벨 데이터

구 분	통합부서	재라벨 결과	데이터
1	기동사업부	기반전력사업본부	539
	함정사업부		
	항공기사업부		
	헬기사업부		
	화력사업부		
2	감시전자사업부	미래전력사업본부	431
	우주지휘통신사업부		
	유도무기사업부		
	첨단기술사업단		
	한국형잠수함사업단		
	한국형전투기사업단		
3	감사관실	청본부(1)	750
	기획조정관		
	대변인실		
	방위산업진흥국		
	방위사업정책국		
	운영지원과		
	조직인사담당관		
4	국방기술보호국	청본부(2)	211
	국제협력관		
5	국방기술품질원	출연기관	219
	국방과학연구소		
계			2,150

재라벨 결과, 실제 방위사업청 조직도의 대분류인 청본부, 기반전력사업본부, 미래전력사업본부, 그 외 출연기관의 형태를 유지할 수 있었다. 결국 기사 2,150건에 대하여 5개 부서로 재라벨링한 데이터를 활용하여 실험을 진행하였다.

<그림 29>에서 볼 수 있듯이 분류 정확도는 70.7%로 정제 데이터 실험결과에 비해 10.5%, 증강 정제 데이터 실험결과에 비해 10.2% 증가된 수치를 보여주었다.

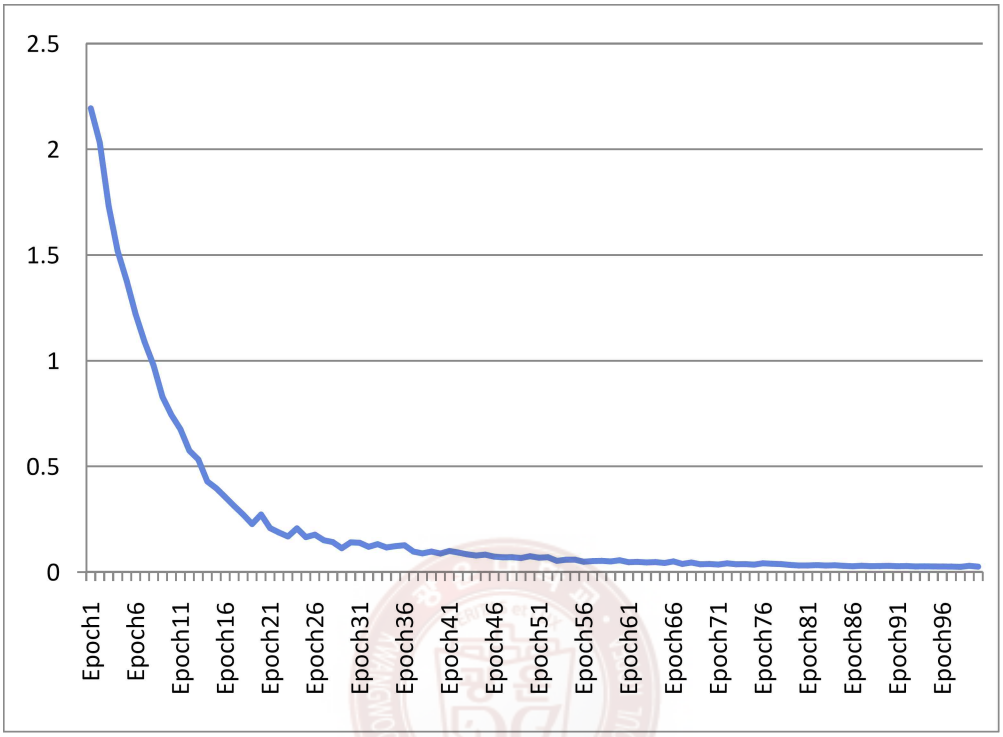
(단위 : %)



<그림 29> 재라벨 데이터의 대응부서 분류 정확도

본 실험은 training loss가 0.02 이후 변화가 없는 점을 고려하여 <그림 30>과 같이 epoch 100까지 수행하였다.

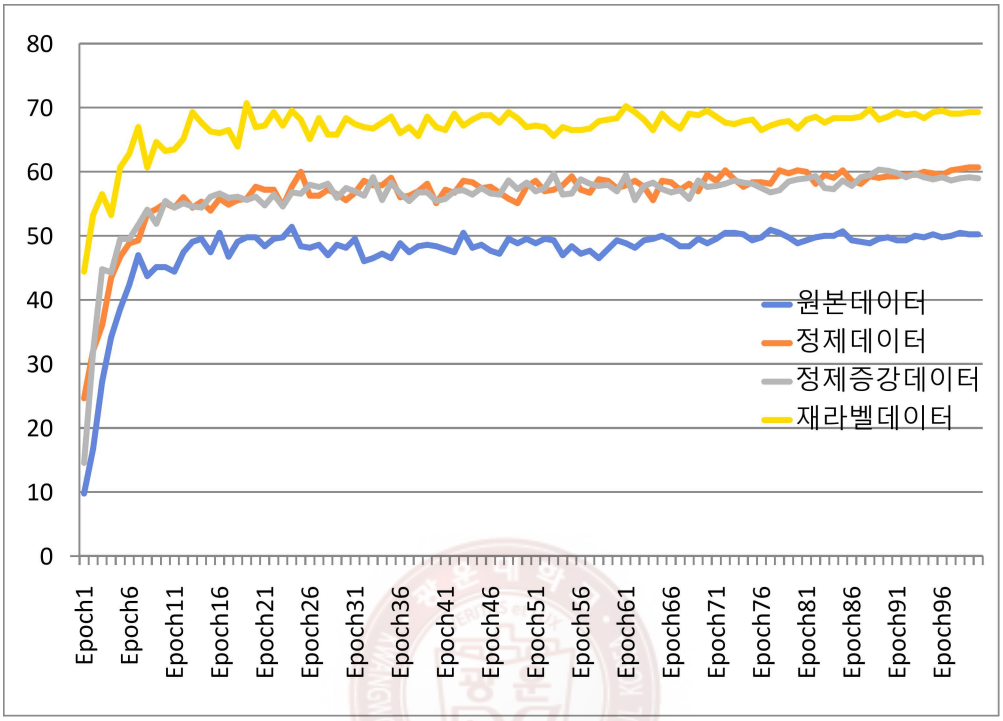
(단위 : %)



<그림 30> 재라벨 데이터의 학습 손실

결론적으로, 대변인실에서 관리하는 데이터를 그대로 사용하는 원본 데이터, 중복 및 노이즈 등을 제거한 정제 데이터, 정제 데이터의 데이터를 증강한 증강 정제 데이터, 정제 데이터에서 기준이 불명확한 라벨링들을 교정한 재라벨 데이터의 실험결과를 종합하면 첫째 재라벨 데이터, 둘째 정제 데이터, 셋째 증강 정제 데이터, 넷째 원본 데이터 순으로 분류 정확도를 가져, <그림 31>과 같이 나타났다.

(단위 : %)



<그림 31> 대응부서 분류모델 데이터별 대응부서 분류 정확도 비교

특히, 정제 데이터 실험결과의 분류 정확도는 60.7%인데, 데이터 수량을 810개(37.7%) 증강시킨 증강 정제 데이터 실험결과의 정확도는 60.3%로 정제 데이터에 비하여 오히려 0.4%의 정확도 감소를 보였다.

반면에, 원본 데이터 실험 정확도는 51.4% 인데, 이를 중복내용 및 노이즈를 제공한 정제 데이터의 실험결과 정확도는 10.6%가 증가하였고, 정제 데이터에서 분류의 기준이 적절하지 않은 데이터 등을 다시 라벨링한 재라벨 데이터의 실험결과 분류 정확도는 70.7%로 정제 데이터에 비하여 10%가 증가하였다.

이 두 가지 사실은 딥 러닝에 있어 학습 데이터의 증가보다는, 분류의 명확한 기준 등 적절한 데이터 관리가 성능에 더 큰 영향을 미칠 수 있다는 점을 보여주었다. 전체 데이터별 실험결과를 나타내면 <표 18>과 같다.

<표 18> 대응부서 분류모델 데이터별 실험결과

구 분	재현율	정밀도	F-1 score
원본 데이터	0.5011	0.5057	0.4985
정제 데이터	0.6069	0.6025	0.6028
증강 정제 데이터	0.6013	0.5962	0.5946
재라벨 데이터	0.7023	0.7057	0.6986

최근 문서분류 연구의 동향을 보면 키워드가 명확하지 않은 문서를 대상으로 분류 모델을 개발하는 경우 그 정확도 자체는 높지 않다. 김인후(2022)가 논문에 대한 분류 모델을 개발한 사례를 보면, 서지학과 같이 초록에 한자가 다량 포함되어 다른 논문과 확연한 특성차이를 보이는 경우 분류 정확도는 90%를 넘기지만, 문헌정보학과 같이 다른 항목으로 분류될 수도 있는 논문 데이터의 경우 분류 정확도가 50%에 불과했다.[71] 기사를 대상으로 한 김미선(2022)이 수행한 연구의 경우 농업신문이라는 한정된 기사 텍스트 데이터를 활용한 결과 61.3%의 정확도를 보였다.[72]

71) 김인후. "딥러닝 기반의 BERT 모델을 활용한 학술문헌 자동 분류." 국내석사학위논문 중앙대학교 대학원, 2022. 서울

72) 김미선. "핵심 키워드 추출 기반의 토픽 모델링을 통한 신문기사 분류모델 제안: 한국 농업신문기사 데이터를 중심으로." 국내석사학위논문 충북대학교, 2022. 충청북도

본 연구에서 수행하는 방위사업 관련 기사 분류의 경우, 청 본부 7국 31과, 기반전력사업본부 6부 36팀, 미래전력사업본부 7부(단) 39팀으로 구성된 방위사업청 조직의 특성상 그 대응부서가 다르게 분류된 경우가 포함되어 70.1% 이상의 정확도를 형성하는 것이 제한된 것으로 판단된다.[73]

73) 방위사업청, 개발형직위(감독총괄담당관) 안내자료, 검색일 : 2022. 6.11, 출처 : https://www.mpm.go.kr/flexer/index.jsp

제5장 결 론

제1절 연구의 요약 및 의의

본 연구의 목적은 방위사업과 관련된 디지털 정체성을 관리하기 위해 인터넷 기사 중 어떠한 내용의 기사들을 관리하여야 하는지에 대한 분류 및 기사 대응에 적합한 부서를 분류하기 위한 모델을 개발을 개발하는 것이다. 이를 위해 분석 데이터 수집, 극성분석을 이용한 대응문서 분류모델 개발, BERT를 활용한 자동문서 분류모델 개발을 수행하는 세 단계를 통해 연구를 진행하였다.

첫 번째 단계에서는 방위사업과 관련된 기사를 수집하기 위해, 방위사업청에서 수집한 '16년~'22년 방위사업청 관련기사 2,153건의 데이터와 이를 교정한 2,150건의 데이터를 활용하였다. 추후 대응문서 및 대응부서를 분류하였을때 대변인실에서 분류한 기준을 활용하여야 실용성을 증명할 수 있기 때문이다. 이때, 과거 대비 부서의 통·폐합, 수행업무 등은 현재로 최신화하는 작업을 수행하였다.

두 번째 단계에서는 기사가 대응이 필요한지 판단할 수 있도록 극성분석(Polarity Analysis) 기법을 활용하여 기사를 모두 읽지 않더라도 기사의 성격이 긍정인지 부정인지 극성을 분류할 수 있는 모델이 개발되었다. 디지털 정체성을 관리하기 위한 본 연구의 목적상 기사가 긍정과 중립의 성격을 가질 때보다는, 부정적인 기사에 대해서 적극적으로 반응할 필요가 있기 때문이다. 한글의 해석 다양성 및 활용한 감성사전의 부족을 고

려 영문화 후 VADER를 통해 극성분석을 수행하여 2,150개의 기사 중 약 27.6%(595건)의 기사가 부정적인 내용으로 분류되었다. 따라서 디지털 정체성 관리를 위해서는 부정적인 내용의 595건의 기사에 집중하는 것이 적절하다고 판단되었다.

세 번째 단계에서는 딥 러닝 기반 언어모델인 BERT를 이용하여 방위사업청 대변인실에서 어떠한 부서의 내용인지 분류해둔 자료를 학습시켜 새로운 기사에 대하여 자동으로 부서를 분류할 수 있는 모델이 개발되었고, 성능을 확인하기 위한 실험은 4가지 실험이 수행되었다.

첫 번째 실험은 실무 데이터를 그대로 활용한 실험이었다. 학습결과, 최종 51.4%의 정확도를 보였다. 이는 웹상의 기사를 그대로 복사해온 내용을 대변인실에서 판단하여 관련 부서를 나눈 데이터를 교정 없이 활용한 결과이다.

두 번째 실험은 데이터 균등화를 위한 라벨링 및 노이즈 제거를 거친 후 추진하였고, 그 결과 60.7%의 예측 정확도를 보였다. 따라서 대변인실에서 데이터를 어떻게 관리하느냐에 따라서 추후 인공지능 학습에 주는 영향을 확인할 수 있었다.

세 번째 실험은 데이터셋에 대한 어그멘테이션 후 추진하였다. 7년간의 데이터라 하더라도 사실 데이터가 많은 양이라고 할 수 없었고, 대변인실에서 수집할 정도로 의미 있는 기사가 생성되어야 하는 상황을 고려하여 기다려서 얻을 수 있는 시험환경이 아니었기 때문이다. 결론적으로 데이터를 증강하여 실험한 결과 60.3%의 예측 정확도를 보였다.

마지막 실험은 라벨링의 효과를 확인하기 위해서 교정된 2,150건의 데

이터를 라벨링 분류를 최소화한 5개로 분류하였다. 그러나 현실과 동떨어지지 않기 위해 실제 방위사업청 조직도를 반영하여 청본부(1, 2), 기반전력사업본부, 미래전력사업본부, 출연기관으로 나누어 실험한 결과 70.7%의 정확도를 확인할 수 있었다.

이는 분류모델에서 일반적으로 정확도와 조화평균의 평균이 50% 이상이면 의미가 있다고 판단하므로 가치있는 모델이 개발되었다고 판단되며, 특히 문서분류 기준이 모호할 때 정확도가 50%에 그치는 선행연구의 모델에 비해 뛰어난 성능을 보여주었다. 또한 실무 데이터를 사용함으로서 현재의 인터넷 기사 분류기준을 수행하는 업무를 기준으로 명확하게 할 필요가 있다는 결론을 도출할 수 있었다.

제2절 연구의 한계 및 후속연구 제언

연구 중 식별된 제한사항은 다음과 같다. 첫째는 데이터 수량의 제한이다. 학습을 위해서는 균일한 분포의 충분한 데이터를 활용하는 것이 바람직한데, 애초에 제공된 데이터의 수량이 라벨링 된 내용을 고려할 때 충분하지 않다는 것이다. 따라서 어그멘테이션 기법을 적용하였으나 실제로 생성된 기사를 사용하는 것이 더 의미 있는 결과가 도출되었을 것으로 판단된다.

둘째는 극성분석을 위한 한국어 감성사전이 부족하다는 것이다. 물론 영문번역을 통해 극복할 방안이 있었으나, 한글 극성분석을 위한 전용사전이 있다면 연구가 더욱 한글에 맞추어 효과적으로 진행되었을 것이라

판단된다.

셋째는 기사의 대응부서가 키워드 등에 따라 명확하게 나누어져 있지 않다는 것이다.

따라서 향후 효율적인 기사 대응을 위한 모델을 향상시키기 위해서는 지속적인 데이터의 관리가 필요하다. 데이터의 수량뿐만 아니라 학습을 위해 적절하게 정제된 기사뭉치를 유지하고, 부서의 변경 등에 따라 수행 업무가 달라지면 그에 맞추어 다시 라벨링하는 노력이 요구된다.

또한, 방위사업을 위한 한국어 감성사전 구축이 필요하다. 이는 추후 본 연구뿐만 아니라 방위사업과 연관된 다양한 자연어 처리를 위해 요구되는 것이며, 연구를 진행하면서 BERT 모델 등의 활용을 통해 구축이 가능할 것으로 사료된다.

마지막으로, 대응부서를 명확한 기준을 가지고 분류하려는 평소의 노력이 필요하다. 이는 첫째와 함께 지속적으로 관리되어야 하는 것으로, 학습 모델의 발전과 함께 요구되는 사항이다. 한 사안에 대하여 인적요인을 고려하여 현재와 같이 다양한 부서의 업무로 분류될 수 있다. 그러나 추후 인공지능 모델개발을 위해 의미 있는 데이터를 축적하기 위해서는 기사별 대응부서를 분류할 때 명확한 기준을 가지고 데이터를 관리하여, 설정된 기준에 따라 대응하여야 할 대응부서를 구분할 수 있는 자세가 요구된다고 판단된다.

참고문헌

[1] 장상훈.(2019). 국방부 디지털 정체성 관리를 위한 실시간 인터넷 문서 자동 검색 및 분석 프로그램 발명에 관한 연구. 선진국방연구, 2(3), 1-21.

[2] 윤호영.(2011). 한국 인터넷의 특징: 소통기반 정보축적 및 유통 문화. 한국사회학, 45(5), 61-104.

[3] 한창진(Changjin Han), 조민수(minsu Cho), and 이중식(Joonseek Lee). "인터넷에서의 설화(舌禍)뉴스 생산의 확산에 대한 연구." 한국HCI학회 학술대회 2010.1(2010), 681-685.

[4] 한국 언론진흥재단. “2014 언론수용자 의식조사”, 검색일 : 2022.10. 8. 출처 : https://www.slideshare.net/girujang/2014-45645888

[5] 정이상, and 이석용. "인터넷 쇼핑몰의 기업 이미지와 품질특성과 만족도, 충성도의 구조관계에 관한 실증적 연구." 경영과 정보연구 28.4(2009), 175-197.

[6] 나익주, 조지 레이코프, 커뮤니케이션북스, 2017년

[7] 김인식, 김자미.(2021). 유튜브 알고리즘과 확증편향.한국컴퓨터교육학회 학술발표대회논문집, 25(1(A)), 71-74.

[8] 황창호(Hwang Changho). "정부역량에 대한 국민인식이 정부성과인식에 미치는 영향 : 정부의 내·외부역량을 중심으로." 지방정부연구 23.4 (2020), 167-189.

[9] 이미나, 박천일 and 왕상한.(2021). 국내 주요 기업의 유튜브 분석 : 홍보활동과 현황. 광고PR실학연구, 14(1), 33-54.

[10] 미디어 오늘, “첨예한 갈등 떠들썩했지만 조용히 사라진 언론중재법 개정안”, 검색일 : 2023. 6. 4., 출처 : http://www.mediatoday.co.kr/news/articleView.html?idxno=304915

[11] 연합뉴스, “모포털기 사라지나...군, ‘평시 솜이불·전시 침낭’ 대체 추진”, 검색일 : 2023. 5.24, 출처 : https://www.yna.co.kr/view/AKR20210711023951504

[12] IT동아, “사이버 렉카 ”또 떴다!..유튜브만 믿으면 되나?”, 검색일 : 2023. 5.24, 출처 : https://it.donga.com/101778/

[13] 최윤성(Younsung Choi), 권오걸(Oh-geol Kwon), and 원동호(Dongho Won). "인터넷 쿠키로 인한 프라이버시 침해와 잊혀질 권리에 관한 연구." 인터넷정보학회논문지 17.2 (2016), 77-85.

[14] 조성태.(2012). 스마트기기의 이용량 증가에 따른 인터넷 포털뉴스 편집환경 연구 - 국내 포털뉴스 사이트를 중심으로 -. 한국디자인포럼, 35, 441-450.

[15] 박재현(Jae-hyun Park) ,and 최호규(Ho-gyu Choi). "인터넷 불매운동에 대한 소비자 의식과 불매운동이 기업의 이미지와 매출에 미치는 영향." 기업경영리뷰 1.2 (2010), 161-180.

[16] 이용성.(2008), 인터넷 자료에 근거한 언론보도의 문제점과 개선방안-인터넷 자료 근거한 오보의 발생구조를 중심으로, 언론중재, 107, 41-49.

[17] 한겨레, “언론중재법 공개 반대...”법이 언론개혁 공감대 훼손“”, 검색일 : 2023. 5.24, 출처 : https://www.hani.co.kr/arti/politics/assembly/1009024.html

[18] 이투데이, "'언론중재법' 처리 앞두고 ‘팽팽’...의견 어떻게 다른가", 검색일 : 2023. 5.24, 출처 : https://www.etoday.co.kr/news/view/2056078

[19] 김태호. "방위사업청 온라인(On-line) 홍보 활성화를 위한 연구." 국내석사학위논문 광운대학교 대학원, 2009. 서울

[20] SPN서울평양뉴스, “북 미사일 70발 쏘자...국민 10명 중 7명 ‘3축 체계 강화’ 지지”, 검색일 : 2023. 6. 4, 출처 : https://www.spnews.co.kr/news/articleView. html?idxno=60426

[21] 최기일, 채우석.(2018). 방위사업 비리 관련 처벌 현황 진단 및 분석 연구. 한국방위산업학회지, 25(4), 13-31.

[22] 장상훈, 전자대변인시스템, 제10-1812933호, 출원 : 2017. 12. 6., 등록 : 2017.12.20.

[23] 구글 알리미 설정, https://www.google.co.kr/alerts

[24] YTN 뉴스 알림 설정 : https://www.ytn.co.kr/info/webpush.php

[25] 카카오톡 뉴스봇 설정 : https://pf.kakao.com/_WlSxbu

[26] 경향신문, “가짜뉴스 SNS 전파 속도 ‘진짜’보다 최고 20배 빨라”, 검색일 : 2023. 6. 4, 출처 : https://www.khan.co.kr/economy/economy-general/article/201803090400005

[27] 나재훈. "軍의 이미지 회복 전략에 관한 연구." 국내석사학위논문 고려대학교 대학원, 2008. 서울

[28] 조덕현. "중앙정부부처 홍보조직의 우수성에 관한 연구." 국내석사학위논문 경희대학교 언론정보대학원, 2011. 서울

[29] 김익현. "인터넷의 매체특성이 인터넷신문 기사에 미치는 영향." 국내석사학위논문 연세대학교 언론홍보대학원, 2003. 서울

[30] 진행남. "인쇄신문과 독립 인터넷신문의 기사특성에 관한 비교연구." 국내박사학위논문 경희대학교 대학원, 2002. 서울

[31] WEF, Technology Tipping Points and Societal Impact, 검색일 : 2023. 5.24, 출처 : https://www3.weforum.org/docs/WEF_GAC15_Technological_Tipping_Points_report_2015

[32] 안상훈, 한은영, 장근영, & 김선희. (2013). 초연결 사회에서 디지털 자아의 정체성 연구. 정책연구, 2013(51), 1-167.

[33] 이상민(Sang-min Lee), 박명호(Myung-ho Park), 김병준(Byung-jun Kim), and 박대근(Dae-keun Park). "빅데이터 분석을 통한 인터넷 뉴스 포털에서의 탈세 논란이 기업 가치에 미치는 영향 연구." 인터넷정보학회논문지 22.6 (2021), 51-57.

[34] Adobe Summit, ADOBE DIGITAL MARKETING SUMMIT. 검색일 : 2023. 5.24, 출처 : https://business.adobe.com/summit/adobe-summit.html

[35] 윤인아. “로봇저널리즘의 이해와 전망”, 제4차 산업혁명과 소프트파워 이슈리포트, 2018-18. 정보통신산업진흥원, 1-14

[36] 고성수. "가짜뉴스 규제법안 실효성에 관한 연구." 국내석사학위논문 서울과학기술대학교, 2018. 서울

[37] 강덕찬. "軍 이미지 類型과 形成要因에 대한 硏究." 국내석사학위논문 高麗大學校, 1992. 서울

[38] 서정근. "국내 신문에 반영된 군 이미지와 보도 성향에 관한 연구." 국내석사학위논문 동국대학교 언론정보대학원, 2001. 서울

[39] 김태웅. "청소년의 정보원 이용이 군 이미지, 복무의사, 신뢰도에 미치는 영향에 관한 연구." 국내석사학위논문 서울대학교 대학원, 2008. 서울

[40] 김수진. "초급장교 교육훈련 홍보 보도기사에 나타난 장교상이 군 이미지와 신뢰에 미치는 영향." 국내석사학위논문 서울대학교 대학원, 2009. 서울

[41] 변의혁. "군 홍보가 ROTC 이미지 및 지원의사에 미치는 영향에 관한 연구." 국내석사학위논문 연세대학교 정경대학원, 2013. 서울

[42] 조인상. "군 이미지에 관한 통합적 연구." 국내박사학위논문 大田大學校, 2014. 대전

[43] 김건아. "빅 데이터를 이용한 제품디자인의 감성반응 분석." 국내박사학위논문 부산대학교 대학원, 2016. 부산

[44] Kamps, J., Marx, M., Mokken, R. J., Rijke, M., “Using WordNet to Measure Semantic Orientation of Adjectives,” Proc. of the International Conference on Language Resources and Evaluation, Vol.4, 2004, pp.1115-1118.

[45] Esuli, A., Sebastiani. F., “SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining,” Proc. of the 5th International Conference on Language Resources and Evaluation, 2006pp. 417-422.

[46] 김명규, “인터넷 감성 텍스트에 대한 극성 분류 시스템,” 한국항공대 대학원 컴퓨터공학 박사학위 논문, 2010.

[47] 김승우, 김남규. “오피니언 분류의 감성사전 활용효과에 대한 연구,” 지능정보연구, 제20권 제1호, 2014, pp.133-148.

[48] 박상민, 나철원, 최민성, 이다희, 온병원.(2018).Bi-LSTM 기반의 한국어 감성사전 구축 방안.지능정보연구,24(4), 219-240.

[49] 유혜연. "텍스트 스토리에서 이벤트의 감정과 등장인물의 역할 인식." 국내박사학위논문 성균관대학교 일반대학원, 2022. 서울

[50] 강아미. "VADER와 성향 점수를 이용한 텍스트 분류." 국내석사학위논문 이화여자대학교 대학원, 2021. 서울

[51] Bhavani, A. & Kumar, B. S. (2021). A Review of state Art of Text Classification Algorithms, Proceedings of the 25th International Conference on Computing Methodologies and Communication (ICCMC), 1484-1490.

[52] 임희석, 고려대학교 자연어처리연구실 (2019). 자연어처리 바이블-핵심이론 · 응용시스템 · 딥러닝. 서울 : 휴먼싸이언스

[53] 김정미, 이주홍.(2017). Word2vec을 활용한 RNN기반의 문서 분류에 관한 연구. 한국지능시스템학회 논문지, 27(6), 560-565.

[54] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L.Zettlemoyer, “Deep contextualized word representations,” in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Jun. 2018, pp. 2227 - 2237.

[55] Khan, A., Baharudin B., Lee L. H. & Khan K. (2010). A Review of Machine Learning Algorithms for Text-Documents Classification. Journal of Advances in Information Technology, 1(1), 4-20

[56] 유소엽, 정옥란.(2019) BERT 모델과 지식 그래프를 활용한 지능형 챗봇. 한국전자거래학회지, 24(3), 87-98

[57] Pan, S. J. & Yang, Q. (2010). A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345-1359

[58] 김판준.(2018). 기계학습에 기초한 국내 학술지 논문의 자동분류에 관한 연구. 한국정보관리학회, 35(2), 37-62.

[59] 박규훤, 정영섭.(2021). KoBERT를 사용한 한국어 일상 주제 분류, 2021년 한국컴퓨터종합학술대회 논문집, 1735-1737.

[60] 박진배.(2020) 사전훈련 된 모델을 통한 한국어 임베딩 성능 비교, 한국국방기술학회 논문지, 2(3), 1-4.

[61] 이수빈(Lee Soobin), 김성덕(Kim Seongdeok), 이주희(Lee Juhee), 고영수(,Ko Youngsoo), and 송민(Song Min). "딥러닝 자동 분류 모델을 위한 공황장애 소셜미디어 코퍼스 구축 및 분석." 정보관리학회지 38.2 (2021), 153-172.

[62] 최윤수.(2018). 기술용어에 대한 분산표현과 딥러닝 모델을 이용한 특허 문헌 자동 분류에 관한 연구. 박사학위논문. 경기대학교 대학원 문헌정보학과.

[63] 심하영, 오수진, 김응모.(2018), “감성분석 기반 호텔 리뷰의 특성별 극성분석 및 유저의 선호도 반영 시스템”, 성균관대학교 2018년 춘계학술발표대회논문집, 제25권 제1호

[64] 박상민, 나철원, 최민성, 이다희 and 온병원. (2018). Bi-LSTM 기반의 한국어 감성사전 구축 방안. 지능정보연구, 24(4), 219-240.

[65] 김영민(Young Min Kim), 정석재(Suk Jae Jeong), and 이석준(Suk Jun Lee). "소셜 미디어 감성분석을 통한 주가 등락 예측에 관한 연구." Entrue Journal of Information Technology 13.3 (2014), 59-69.

[66] 행정안전부, “방위사업청, 사업관리 중심 조직개편으로 제2의 개청!”, 검색일 : 2023. 6. 6, 출처 : https://www.mois.go.kr/frt/bbs/type010/commonSelectBoardArticle.do?bbsId=BBSMSTR_000000000008&nttId=72892

[67] 방위사업청훈령 제591호(2020.4.21.) 방위사업 홍보 규정

[68] Manning D. C., Raghavan P. & Schutze H. Introduction to Information Retrieval. 안동언, 김재훈, 남영준, 박혁로, 이상곤 공역 (2010). 최신 정보검색론. 경기도 : 교보문고

[69] 황상흠, 김도현.(2020). 한국어 기술문서 분석을 위한 BERT 기반의 분류모델. 한국전자거래학회지, 25(1), 203-214

[70] Jason Wei, Kai Zou, Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks(2019)

[71] 김인후. "딥러닝 기반의 BERT 모델을 활용한 학술문헌 자동 분류." 국내석사학위논문 중앙대학교 대학원, 2022. 서울

[72] 김미선. "핵심 키워드 추출 기반의 토픽 모델링을 통한 신문기사 분류모델 제안 : 한국 농업 신문기사 데이터를 중심으로." 국내석사학위논문 충북대학교, 2022. 충청북도

[73] 방위사업청, 개발형직위(감독총괄담당관) 안내자료, 검색일 : 2022. 6.11, 출처 : https://www.mpm.go.kr/flexer/index.jsp