커뮤니케이션 이론
2023년 19권 1호 84~123쪽
https://doi.org/10.20879/ct.2023.19.1.084

자연어 생성기반 뉴스 보도 패턴 일반화 및 뉴스 구성에 따른 분류 가능성

소규모 LSTM 생성 데이터를 통한 내용 및 표현 형식 기반 뉴스 유형화 원리 고찰*

윤호영 이화여자대학교 커뮤니케이션·미디어학부 조교수**

안도현 제주대학교 언론홍보학과 부교수***

본 논문은 인공지능 자연어 생성 모델을 통해 보도된 기사들의 일반화된 보도 내용 데이터를 만들고, 이를 활용하여 이후 지도학습기반 뉴스 클러스터링 방식을 제안하는 연구이다. 보다 구체적으로는 뉴스를 수집하고 문장 기반 패턴 분석 방법을 활용하여, 이미 작성된 기사에서 쓰이는 단어와 주요 문장의 패턴이 추론된 자연어 생성 기사 문장을 만들어낸다. 생성된 문장은 보도된 기사의 기본적인 보도 내용 및 관행을 보여주는 보도된 내용들의 일반화된 특질을 보여주는 것으로 본다. 그 다음 생성된 문장과 수집된 데이터 문장간의 내용 특질 유사성을 레벤슈타인 거리와 ROUGE 지표로 비교하여 컴퓨터가 만들어낸 문장과 실제 기사 문장 간의 내용과 표현상의 괴리를 측정함으로써, 보도된 뉴스를 빠르게 유형화하는 방법을 제안한다. 본 글에서는 이러한 방법이 적용되는 과정을 소규모 데이터로 감염병 백신 보도를 주제로 시연하고, 해당 방법이 가지는 의의와 향후 연구 가능성을 논의한다.

KEYWORDS 자연어 생성, LSTM, 레벤슈타인 거리, ROUGE, 뉴스 유형화

---

* 이 논문은 2021년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2021S1A3A2A02090597).

** hoyoungyoon@ewha.ac.kr

*** 교신저자, dohyun@socialbrain.kr

1. 서론

기사 유형화는 언론 연구에서 더욱 중요해지고 있다. 디지털 기술의 확산으로 실시간 뉴스 생산량이 급증하고 있고, 그로 인해 대량으로 생산되는 언론 보도를 빠르게 구분하여 연관성이 있는 기사를 묶어냄으로써 독자가 원하는 다양한 정보를 빠르게 제공해야 할 필요성이 높아졌기 때문이다. 또한 흔히 말하는 가짜뉴스의 범람으로 인해 이를 탐지하는 방법에 대한 고민이 늘어가는 것 역시 기사 구분의 중요성을 보여준다(Lazer et al., 2018; Shu, Sliva, Wang, Tang, & Liu, 2017). 물론, 뉴스간 연관성을 어떻게 볼 것인가에 따라 기사를 묶어내는 방식이 달라진다. 예를 들어, 연관성 평가 기준이 실시간 검색 결과에 대한 뉴스 전달이라면, 짧게 검색하는 내용 즉 현재 시점에서 사용자가 원하는 정보가 무엇인지를 판단해야 하는 것이 중요하다. 그렇지 않고 연관성 평가 기준이 유사한 정치적 지향의 기사를 얼마나 잘 묶어내는가라면 사용자의 성향을 파악하여 사용자에 따라 다르게 전달하는 다양성 측면이 더욱 중요할 수 있다. 기사가 제공하는 정보량과 독자에게 잘 읽히는 가독성을 기반으로 뉴스 품질을 정의하는 네이버의 알고리즘이 이와 유사한 기준을 적용한다(유봉석·최재호·최창렬, 2020).

포털 사이트의 예와 마찬가지로, 최근에는 컴퓨터를 이용한 전산화된 방법을 활용하여 뉴스 특질을 파악하는 경우가 많아지고 있다. 단순하게 단어의 빈도를 계산하는 방식부터 통계모형을 이용한 기계학습 등이 쓰이고 있고, 토픽 모델링(Topic Modeling)을 이용하여 비지도학습기반 군집화(Clustering) 방법으로 연구가 많이 진행되는 추세다. 최근에는 지도학습 텍스트 분류 방법을 통해 특정 미디어만이 가지고 있는 보도 내용을 파악한 후 이를 다시 비지도학습을 통해 분석하거나 (윤호영, 2022) 내용 항목을 비지도학습 기반 토픽모델링으로 구분한 후 이를 다시 딥러닝 모형으로 뉴스 내용을 분류하는 등 비지도학습과 지도학습을

혼용하는 연구도 확산되고 있다 (정재철·이종혁, 2022).

이 중 비지도학습과 지도학습을 혼용하는 연구의 경우 판별을 위한 뉴스 내용 라벨링에서는 수작업을 수행하지만 다른 부분에서는 상당 부분 자동화하는 방식을 활용하고 있다. 예를 들어, 비지도학습을 통해 주요 내용(토픽)을 군집화하여 구분하고 이에 대한 라벨링에 해당하는 코딩을 수행한 후 이를 다시 지도 학습에 쓰일 수 있는 학습 데이터로 구축하는 비지도학습-지도학습 연계 방식 (예. 정재철·이종혁, 2022) 또는 뉴스 보도가 보도된 언론사를 정확히 분류할 수 있는가를 지도학습을 통해 분류한 후, 해당 미디어만의 보도기사들의 내용을 비지도학습을 통해 보여주는 방식을 통해 (예. 윤호영, 2022) 대용량 데이터를 활용한 뉴스 분류 작업과 보도 경향을 분석하는 방식을 보여주고 있다.

수용자에게 정보를 전달하는 소비자 만족도가 검색된 '뉴스 품질'이라고 평가하는 포털 사이트의 뉴스 특질 평가 방식 역시 사전 항목을 설정하고 이에 따라 지도 학습 뉴스 분류기를 만드는 지도학습 기반 방식과 그 이후 뉴스 주제에 따른 클러스터링 이후 제시하는 과정에서 사용되는 비지도학습이 혼용된다. 네이버와 같은 포털사이트가 뉴스를 분류하고 제시하는 방식인 뉴스 클러스터링을 통한 뉴스 분류, 추천을 위한 개인화 작업, 네이버가 정의하는 '뉴스 품질' 평가, 뉴스 순위 배열의 순서 정하기 등의 과정에서 보면 '뉴스 품질' 평가 단계에 지도학습이 들어가 있다 (네이버 검색, 2021a). 비지도학습과 지도학습의 혼용은 가짜뉴스 탐지에도 동일하게 적용되는데, 뉴스 내용이 가지고 있는 언어학적 특징이나 단어 특징을 통해 가짜뉴스를 탐지하거나, 소셜 미디어내에서 전파되는 방식을 탐지하여 사용자를 특정하는 등의 연결망 기반 탐지 등을 활용하는 방식에서도 문장이나 단어 임베딩 과정에서 비지도학습이 지도학습 기반 방식과 함께 활용되기도 한다 (Shu et al., 2017).

뉴스 분류와 관련하여 저널리즘 원칙에 따른 전통적인 방법의 뉴스 기사 내용 분석 방식은 사전에 뉴스 내용 평가를 위한 항목을 설정하고

해당 항목에 따라 뉴스 기사를 수작업으로 평가하는 방식이 활용되어 왔다 (예. 김경모·박재영·배정근·이나연·이재경, 2018; 박재영·이완수, 2010; 유수정·이건호, 2020). 뉴스 내용 분석과 관련된 예로서 뉴스 품질 평가 방법을 보면, 이 과정에서 가장 중요한 것은 사전에 기준을 설정하는 것이다. 지도 학습 역시 학습 데이터에 대한 라벨링 분류가 중요하다. 그리고 전통적인 방법과 지도학습 모두 공히 수작업에 의해 분석 대상이 되는 기사들을 평가하는 것으로부터 출발한다. 그러나 수작업에 의한 내용분석 및 평가는 정밀한 분석이 가능하다는 장점이 있지만, 대량의 기사를 처리할 수 없는 한계가 있다. 따라서 수작업에 의한 내용 분석만으로는 유통되고 있는 대용량의 기사를 분류하고 평가하기 어려운 문제점이 있다. 또한, 자동화된 텍스트 분석이 도입된 주요 이유 중 하나로 지적되고도 있는데, 사전에 항목 설정에 대한 합의가 이루어지지 않으면 (DiMaggio, 2015) 뉴스 내용 분석이 어려울 뿐만 아니라, 어떤 항목을 설정하느냐에 따라 뉴스 분류 자체가 달라지는 특징이 있다. 이와 더불어서, 뉴스 내용 구분에 대한 절대적인 기준에 대한 합의 및 실제 코딩 과정에서 코더간의 일치성을 확인하는 과정에서 소요되는 자원과 시간의 문제 역시 기존의 수작업을 대용량 데이터 처리에 적용하기 어려운 점으로 작용한다.

본 연구는 이러한 어려움을 해결하기 위한 방안 중 하나로 자연어 생성을 통해 보도된 기사의 일반화된 데이터를 생성하여 만들어내고, 해당 데이터와 실제 보도된 뉴스 내용과의 유사성 또는 상이성을 측정함으로써 뉴스를 빠르게 유형화하는 방식을 제안한다. 이는 뉴스 데이터의 유형을 빠르게 구분해냄으로써, 이후 지도학습과 결합한 뉴스 분류에 활용할 수 있는 사전작업의 의미를 가진다. 예를 들어, 어떤 특정 주제에 대해 보도된 뉴스를 1) 가장 일반적인 보도 내용과 표현 방식으로 정형화된 뉴스 기사 2) 정형화된 기사와 동일한 내용을 다루고 있으나 정치적 지향이 달라 내용 구성상의 표현 차이가 있는 기사 3) 정형화된 기사와 표현은

유사한데 정작 내용은 다른 기사 4) 일반적인 보도 내용과도 내용이 유사하지 않고 해당 주제를 다루는 구성마저도 다른 기사 등으로 구분할 수 있다면, 해당 기사들이 담고 있는 내용을 향후에 판단하여 기사의 분류에 활용할 수 있을 것으로 기대할 수 있다.

보다 구체적인 이해를 위해 실제 보도된 내용으로 설명하면 다음과 같다. 전기료 인상과 관련하여 "서울 공덕동에 사는 A씨는 21일 전기요금 고지서를 보고 깜짝 놀랐다. 이달 전기 요금이 19만 4420원이나 나왔기 때문이다"라고 표현한 것이 가장 일반화된 보도 기사 첫 줄이라고 생성된 것으로 가정할 경우, "청주 사직동에 거주하는 박 모씨는 21일 19만 3천 820원의 전기 요금이 찍힌 고지서를 받고 기절 초풍했다"라는 기사 첫줄을 가진 기사와 내용과 표현 형식면에서 유사성을 비교할 수 있다.[1] 이 경우 가장 일반화된 표현과 내용과 표현에서 유사한 기사들은 단순히 유사 기사가 아니라, 어뷰징 기사로 판단할 수 있는데, 이후 이들 기사들에 대한 라벨링을 어뷰징으로 하는 지도 학습을 수행할 수 있다는 것이다.

이와 같은 방법의 적용은 생성 모형이 가지는 장점을 활용하는 것인데, 현재 보도되고 있는 뉴스의 일반화된 보도 패턴과 내용을 대표적으로 보여주는 데이터를 생성하고, 이를 기반으로 보도되는 기사들을 빠르게 유형화하여 향후 수행될 수 있는 - 예를 들어, 뉴스 품질 평가와 같은 - 뉴스 분류 과정에서 라벨링된 데이터를 구축하는데 상당한 도움을 줄 수 있다. 전통적인 방식으로 기계학습모형을 만들기 위해서는 상당량의 기사를 사전에 설정된 기준에 따라 분류한 양질의 학습자료를 사람의 수작업으로 만들어야 할 필요가 있는데, 이 작업은 매일매일 수행하기도 어렵고, 빠르게 변화하는 보도 관행과 뉴스 내용을 따라가기 어렵다. 결국 기

---

1) 실제 보도된 기사로 처음 인용문 기사는 https://www.mk.co.kr/news/society/8444299에서 확인할 수 있으며, 두 번째 인용문 기사는 다음의 세 개 기사에서 동일하게 쓰였다. https://news.sbs.co.kr/news/endPage.do?news_id=N1004899384, https://www.yna.co.kr/view/AKR20180821121400064, https://www.joongang.co.kr/article/22904086

사 유형화에 기반한 뉴스 분류를 효율적으로 자동화하는 방법을 도입하기 위해서는 기계학습모형 구축에 필요한 학습 자료를 자동으로 생성할 수 있어야 하고 수작업을 최대한 줄이는 방식으로 방법론적인 개발이 이루어져야 할 것으로 본다. 그 방법의 일환으로 본 연구에서는 자연어 생성과 이후 생성된 자연어를 보도된 기사와 비교하는 방식을 제안하는 것이다. 이후에 서술할 것이나 생성 방법을 활용한 일반화된 데이터의 생성 및 보강은 기존의 지도학습 방법에서 나타날 수 있는 데이터 불균형 등 기술적인 문제를 보완하는 여러 가지 장점 역시 가지고 있다.

최근 인공지능을 활용한 자연어처리 기법이 크게 발전하면서 사람이 하는 말을 그대로 컴퓨터 코드로 만들어주거나 동영상을 편집하는 수준으로까지 기술이 발전하고 있다. GPT-3 라 불리는 인공지능은 방대한 양의 언어를 학습하여 자연어를 단순히 분석하는 것에 그치지 않고, 모든 종류의 글을 쓸 수 있는 수준으로 발전하였고 (Floridi & Chiriatti, 2020), 약 3천억개가 넘는 데이터로 사전 학습을 수행하고 약 1800억개에 달하는 딥러닝 파라미터를 통해서 자연어 처리에 있어서 극도로 강력한 모습을 보이는 혁신 기계가 출현하였다. 본 논문을 작성하는 과정에 ChatGPT가 공개되면서 많은 논의가 뒤따르고 있으며, 학계에서도 자연어 처리에 기반한 인공지능의 활용이 어느덧 텍스트 분석에서 선택할 수 있는 여러 가지 선택지 가운데 하나인 시대가 되었다. 그런 의미에서 본 연구에서 시도하는 방식이 가지는 나름의 의의가 있을 것으로 생각된다.

본 연구는 심층학습 기반 LSTM 자연어 생성을 감염병 백신 부작용 보도를 주제로 시연하여 어떠한 한계가 있을지, 어떠한 방식이 가능할지 가늠해 보는 실험 연구이다. 자연어 생성은 생성을 위한 학습 텍스트 자료에 나타난 '패턴 인식'을 하여 확률모형에 따라 단어들을 연속으로 연결지어 자연어를 만들어내는 방법이다. 본 연구에서는 감염병 백신 부작용을 다루지만, 다른 주제를 선택할 수도 있을 것으로 생각한다.

최근 자연어 처리와 관련된 패턴 인식은 주로 '비교'와 '분류'를 위한

방식으로 발전해 왔다(Manning & Schutze, 2000; Young et al., 2018). 특히, 언어학이라고 하는 특정 분야가 인공지능이 본격적으로 발전하기 이전부터 크게 발전한 상태였기 때문에, 규칙 기반 언어 패턴 인식이 존재하고 있었고, 이러한 규칙을 어떻게 인공 지능이 이해하도록 할 것인가가 1세대 인공지능의 지식기반 온톨로지시스템 방식과 더불어 잘 어울렸었다. 그러나, 최근 경직된 온톨로지 기반 규칙이 아니라 실제 활용되는 방식에 대한 통계적인 이해와 이를 신경망을 통해 최적화를 이루는 방식이 개발되면서 규칙에 없는 패턴도 인식하게 되고 이러한 패턴이 신조어에 대한 판독이라든가, 언어 규칙을 모르는 상태에서도 대용량 데이터를 통한 기계 번역이 가능해지는 등의 방식으로까지 발전하고 있다(Young et al., 2018)

본 글에서도 이러한 방식을 적용하여 자연어 생성 방법을 제안하는데, 여기서 제안하는 방법은 생성된 자연어를 바로 지도학습에 적용하기 위한 것이 아니라, 지도학습 데이터 라벨링을 위한 데이터 유형화의 기준을 설정하는 데이터를 만들자는 것이다. 그 과정을 짧게 설명하면 다음과 같다. 우선 보도된 기사를 학습 자료로 하여 보도된 기사의 일반 특징을 반영하는 자연어 생성을 한다. 그 다음, 해당 특징과 보도된 기사를 비교함으로써, 보도된 기사가 일반적인 보도 기사의 패턴과 얼마나 다른지 판단하는 방식을 제안한다. 이 과정을 통해 비교 대상 특정 기사와 자연어로 생성된 기사 간의 유사성 또는 상이성의 차이를 인식하게 된다. 이후, 유사성이나 상이성이 의미하는 바가 무엇인지 사람이 나중에 판정하게 된다면 자연스럽게 현재 보도된 기사의 일반화된 내용과 표현이 가지는 의미를 유추할 수 있게 될 것으로 본다. 그리고 유사성과 상이성의 수준을 정하여 뉴스 데이터를 구분하는 라벨링을 수행한다면 이 데이터가 향후 지도학습기반을 기반으로 한 특정 주제 관련 뉴스 기사 분류기를 만드는데에도 활용할 수 있다고 보는 것이다. 마치 로봇 저널리즘에서 로봇이 기사를 작성하듯이(김동환·이준환, 2015), 기사에 들어갈 필수적인 특징

을 자연어 생성으로 만들어내어 생성에 사용된 보도 기사들의 일반화된 보도 패턴으로 삼는 것이다. 이 과정을 비유하자면, 삼각 측량을 위한 북극성의 위치를 자연어 생성을 통해 만들어내고 이를 기반으로 기사들을 구분하여 라벨링이 이루어질 경우, 북극성의 존재 때문에 다른 기사들이 카시오페이어와 북두칠성 위치 중 어디에 가까운지 판단하기가 더욱 쉬워진다는 것이다.

2. 자연어 생성 기법을 통한 일반화된 뉴스 보도 패턴 발견 및 유형화

1) 자연어 생성을 통한 텍스트 데이터 구분의 장점

새로운 데이터를 생성하여 학습하는 데 활용하는 경우를 선행 문헌에서는 데이터 보강(Data Augmentation, 이하 DA)으로 이해하고 있다. DA가 기계학습기반 심층학습 분야와 관련하여 가장 활발한 분야는 컴퓨터 비전(Computer Vision, 이하 CV) 분야로 다양한 형태의 이미지를 인식하는 분류기의 일반성을 획득하기 위한 목적으로 많이 활용되어 왔다(Shorten & Khoshgoftaar, 2019). 예를 들어 이미지에서 라벨링 대상물의 위치를 변화시키거나, 이미지를 뒤집거나 회전, 색상을 바꾸거나, 일부를 자르거나, 일부러 잡음(noise) 데이터를 집어 넣거나 등의 방식으로 다양한 가능성을 탐색하여 분류기가 이미지의 일반화된 특질을 인식하여 분류 성능을 높일 수 있도록 하는 것이다. 이 과정에서 데이터가 가지는 불균형성을 해결하는데 DA 방법이 매우 적절한 대안이 되는 것으로 알려졌다 (Buda, Maki, & Mazurowski, 2018). 특히 분류를 수행해야 할 범주가 이분화된 범주가 아닌 다중 분류일 경우 데이터의 불균형은 데이터가 많은 범주의 분류 성능은 높이지만 반대의 경우는 성능이 떨어지는 분류 범주 성능의 불균형을 초래하기도 하는데 (Buda et al., 2018; Chen, Liaw, & Breiman, 2004; Sun, Wong, &

Kamel, 2009), 이를 극복하는데 도움이 되는 것으로 본다.

텍스트 데이터와 관련된 DA 역시 동일한 이유에서 제안되었는데, 기존 연구들은 텍스트 생성을 통해 DA를 수행할 경우 다음과 같은 장점이 있다고 설명한다. 우선, CV와 마찬가지로, 데이터를 구분하여 판별할 때 데이터 생성을 통해 데이터가 가진 일반성을 획득해내는 일종의 데이터 정규화(regularization)의 의미를 가진다고 본다(Anaby-Tavor et al., 2020; Bayer, Kaufhold, & Reuter, 2022; Hernández-García, & König, 2018; Shorten, Khoshgoftaar, & Furht, 2021). 생성되는 데이터는 확률에 기반하여 글자 또는 문장의 연속적인 관계를 추출해 내는데 이 과정에서 일반화된 패턴이 자연어 생성 과정에 반영되게 된다. 따라서, 극단적인 방식의 잡음이나 데이터상의 편향이 있는 데이터에 비해 가장 일반화된 표현을 생성하게 된다. 이는 기계학습에서 과적합(overfitting)의 문제를 해결하기 위한 방법과 유사한데 분류기의 일반화된 성능을 유지하기 위해 학습 데이터 전체를 학습하여 반영하는 것이 아니라 교차 검증(cross validation)을 통해 일반성을 확보하는 것과 동일한 방식으로 이해할 수 있다. 이를 뉴스 구분과 관련지어 생각해 보면, 저수준의 기사와 고수준 기사를 동시에 학습하는 과정 중에 두 가지를 모두 반영하는 평균을 만들어 내는 것이 아니라, 극단치를 배제하는 노이즈 제거를 통해 모두의 가장 일반적인 특징을 생성할 수 있다는 의미가 된다. 말 그대로 보도된 내용에서 전형적인 패턴을 찾아내어 이를 만들어내는 것이다. 그런데 이러한 생성이 가진 장점은 단순히 일반적인 기사 패턴을 찾아내어 표현하는 것에 그치지 않는다.

텍스트의 생성이 데이터의 다양성 확보를 통한 데이터 불균형성 해소와 라벨링 데이터 편향성을 줄이는 방안도 된다는 것이다 (Feng et al., 2021; Shorten et al., 2021). 생성모델의 경우 단순히 있는 그대로 원래 데이터를 보충하는 것이 아니라 새로운 형태의 데이터를 만들어 내어 보충함으로써, 데이터의 다양성을 확보하여 불균형성을 줄이는데

기여할 수 있다고 본다. 이는 텍스트 데이터뿐만 아니라 CV에서도 DA가 가지는 일반적인 장점이다. 예를 들어, 품질 평가에서 낮은 수준에 있는 기사만이 범람할 경우 높은 수준에 있는 것으로 평가받는 기사들이 가지고 있는 특질을 중심으로 새로운 텍스트 데이터를 생성한다면 향후 데이터 분류 과정에서 분류기가 가질 수 있는 편향성을 해소하는데 도움을 줄 수 있게 된다. 물론 이를 위해서는 보도된 기사를 먼저 사전에 군집으로 묶어내고 그러한 군집에 대한 평가를 진행해야 하는 문제가 있으나, 생성없이 진행할 경우에는 기존 데이터에서 부족한 데이터를 다른 데이터를 통해 찾아서 보충해야 하는 문제가 발생한다. 그런데, 문제는 다른 데이터에서 가져와서 보충하는 것이 과연 적절한가에 대한 의문이 남는다. DA는 이러한 문제를 줄이는데 기여할 수 있다는 것이다 .

이와 관련하여, 선행 문헌은 지도학습과 관련하여 새로운 데이터에 대한 라벨링에 들어가는 비용과 시간이 상당한데, 이를 기존 훈련 데이터로 쓸 수 있는 데이터로 생성함으로써 이와 같은 비용과 시간을 줄이는 효율성을 발휘할 수 있다는 점 역시 자연어 생성의 장점이라 언급한다 (Bayer et al., 2022). 데이터의 불균형성을 해소하기 위해 새로운 데이터를 발굴하고 또한 발굴한 데이터에 대한 라벨링이 다시 한번 진행되어야 하는데 이러한 복잡다단한 과정이 아닌 현재 데이터로부터 다양성을 지닌 데이터를 생성시키게 되면 복잡한 과정을 거치지 않아도 되는 장점을 가진다는 것이다. 이는 데이터가 부족하여 프라이버시를 보호해야 하는 경우에도 활용될 수 있다.

생성에 한정하지 않고, 텍스트와 관련된 DA를 보면 CV와 마찬가지로 특정한 단어만 가려서 학습을 수행한 후 해당 단어 위치에 맞는 단어를 찾는 등 단어 수준 DA에서부터 시작하여 구(phrase)와 문장(sentence) 및 전체 문서 수준 DA까지 다양한 수준이 존재한다. 최근에는 생성모델이 트랜스포머(tansfomer)을 통해 많이 이루어지고 있는데, 이러한 DA가 제기되는 배경에는 무엇보다 라벨이 없는 대용량 데이

터를 어떻게 하면 효율적으로 라벨링 할 것인가에 대한 고민에서 출발하고 있다(Shorten et al., 2021). 그리고 그 과정에서 DA와 자기지도학습(self-supervised learnning), 전이학습(transfer learning) 등의 방법을 활용하여 해결하고자 하고 있다. 최근에는 사전 훈련(pre-trained)을 통한 BERT나 GPT-2를 이용한 텍스트 생성을 진행하기도 하는데, 이를 적용하는 과정에서 적게는 몇 천건에 이르는 데이터를 수작업으로 분류해야 하는 문제가 있다.

그렇다면 보도된 기사를 기반으로 자연어 생성을 통해 보도된 뉴스의 일반화된 특질을 파악하여 구분하는 방식이 가지는 장점은 어디에 있을까? 뉴스 기사와 관련된 다양한 방식의 활용이 가능하겠으나 본 연구에서는 뉴스 내용과 표현 형식에 따라 뉴스를 유형화하는 방식에 활용하는 것을 제안한다.

2) 지도학습 기반 라벨링 및 학습 데이터 구축 수작업의 효율성 문제

뉴스를 분류하는 지도 학습을 수행할 경우 가장 중요한 작업은 뉴스 라벨링을 위한 분류 기준을 설정하고 그러한 기준에 따라 객관적인 측정을 수행하는 것이다. 다시 말해 전산화된 지도학습 기반 뉴스 분류는 뉴스가 가진 자질(feature) 중에서 중요하게 생각하는 자질을 선택하고, 이러한 자질에 따라 데이터를 분류할 수 있는 학습 데이터를 만듬으로써, 지도학습을 수행하게 된다. 항목이 저널리즘 기준인가 아니면 소비자 만족도인가 여부를 떠나서 이와 같은 지도학습 기반 뉴스 분류는 네이버와 같은 포털 사이트에서 평가하는 소위 네이버의 '뉴스 품질' 방법과 유사하다. 기계학습을 기반으로 뉴스 기사를 분류하게 랭킹을 부여하는 네이버는 뉴스 추천을 위해 뉴스 기사 품질 점수를 부여하는 항목으로 시의성, 제목 및 본문 길이, 언론사의 중요 기사 표기 여부, 멀티미디어 사용 여부, 실명 기자 여부, 언론사의 영향력 등을 고려한다고 밝힌 바 있다 (네이버 검색, 2021b). 이러한 기준에 따라 뉴스 자질을 선택하고 기사가 해당

자질에 부합하지에 대한 점수를 만들어 해당 점수에 따라 사용자에게 보여주는 식이다. 예를 들어, 연관성 자질을 중시하는 경우, 만약 뉴스 독자가 '코로나 확진자 수'라는 검색을 하였다면 여기서 가장 중요한 것은 정보 내용으로 전국 확진자 숫자 내용이 담겨 있는 뉴스가 독자가 궁금해하는 사항을 담고 있다고 보고 뉴스 내에서 해당 숫자와 관련된 내용이 있는지를 확인한 후 그 내용이 있으면 네이버의 정의에 따른 품질 좋은 뉴스로 분류하여 사용자에게 보여주는 식이다 (예. 네이버 검색, 2019). 또는 클릭유도성으로 제목에 있는 내용이 본문에 없다면 연관성이 떨어지는 사용자가 원하는 정보가 없는 기만성 뉴스로 분류할 수 있다 (이선민, 2020).

그런데 지도 학습 기반 뉴스 분류 방식은 데이터 수집과 관련하여 상당한 어려움을 노정하게 된다. 예를 들어 감염병 보도를 중심으로 살펴보면 가장 먼저 감염병 보도를 수집해야 하는데 감염병 보도에는 예방적 조치, 증상, 현재 상황, 향후 조치, 정부 대응, 백신 등 매우 다양한 영역이 존재하고 있으므로 해당 영역에 따라 뉴스 분류 특질에 대한 탐지가 개별적으로 진행되고 사람마다 판단이 일치해야 한다 (예. 허용강·차수연·서필교·김소영·백혜진, 2015). 따라서, 영역을 선별하여 감염병 보도 데이터를 골고루 수집해야 학습이 적절하게 이루어지게 되는데, 이를 위해서는 각 영역에 맞는 기사들을 선별하여 수집할 필요가 있다. 또한 각 영역의 데이터를 모았다고 하더라도 해당 특질이 많은 기사와 적은 기사가 골고루 분포해야 학습이 이루어질 수 있다. 과도하게 특정한 범주가 많다든지의 데이터 자체가 가지고 있는 불균형이 나타날 경우 기계학습은 분류와 관련된 학습 예측의 성능을 상당히 떨어뜨리게 된다(Chen et al., 2004; Sun et al., 2009). 이는 통계적 기계학습 뿐만 아니라 인공 신경망을 활용하는 방법에서도 동일하게 나타나게 되는 것으로 알려져 있다(Sun et al., 2009; Wu & Chang, 2003). 특히 학습 데이터의 불균형은 데이터 크기 문제가 아니라 불균형 분포에 따른 상대적인 문제이

므로 분류 범주 구분에 따른 데이터가 모두 적정량이 존재해야 한다는 점이 전제되어야 하는 점이 지도 학습 기반 뉴스 분류의 한계로서 존재한다. 다시 말해, 특정 범주의 데이터가 1만개이고, 다른 범주가 2000개일 경우, 다른 범주를 대략 3000개 정도 늘려서 5000개 정도로 데이터의 크기를 키우는 것이 중요한 것이 아니라 상대적으로 5:1 또는 2:1이 되는 데이터를 1:1로 만들어야 하기에 8000개의 데이터를 새로 구축하여 1:1로 만들어야 하는 문제라는 것이다.

정리하면, 지도학습 기반으로 뉴스를 분류하기 위해서는 1) 분류에 사용될 뉴스 자질이 어떤 것인지 기준을 정의해야 하고 2) 해당 자질이 적용되는 코딩이 수작업으로 진행되어야 하며 3) 코더간 일치하는 지 여부도 평가해야 한다. 그리고 데이터의 불균형성 등의 문제와 더불어 4) 기사 내용상에서도 다양성이 존재해야 분류 알고리즘이 정당한 평가를 받을 수 있다. 따라서, 이는 단순히 수집된 기사의 일부를 표본 추출하여 분석하는 것 이상의 전체 수집된 기사에 대한 내용과 표현 등 뉴스 구성에 대한 전반적인 지식을 사전에 필요로 한다는 것을 의미한다. 만약 단순히 정보 전달의 목적을 위해 어떠한 내용을 담고 있는가만 판단한다면 정보 여부만을 판단하면 되지만 그것이 아니라 특정한 특질을 골고루 가지고 있는 뉴스를 판별해야 한다는 것은 그 이상의 조건을 만족시켜야 하는 것이다. 수작업 코딩 역시 매우 훌륭하게 이루어져서 명확히 오류없이 진행되었다고 하더라도 위의 항목에 맞도록 뉴스 기사들이 풍부하게 존재해야 한다. 결과적으로 뉴스 분류를 위한 수많은 기준과 그에 따른 판정 등에 있어서 방대하게 진행되어야 하는 수작업은 대용량 뉴스 데이터가 매우 많이 발행되는 시대에 지속가능하기 어려우며, 방법론적인 효율성을 추구해야 할 필요성이 있다고 본다.

3. 자연어 생성 방식의 뉴스 유형화 데이터 구축 제안

이와 같은 상황에서 데이터를 수작업으로 분류하지 않고 다량의 기사들을 구분하고 그 구분을 기준으로 기사들을 향후 분류하는 판별법을 도입하는 방법이 대안이 될 수 있다. 사전에 특정한 기준이 없는 상태에서 이미 보도된 기사들을 구분하고(Clustering) 그 구분된 항목을 이후 지도학습으로 평가하는 방식을 말한다. 그러나 사전 지도 학습에 의한 것이 아니라 단순히 클러스터링을 통한 내용 분류만을 진행한다면 토픽 모델링 등 어떤 내용이 알아보는 것으로 충분하다. 그러나 동일 내용이 어떤 방식으로 보도되었는가를 살피는 과정에서 보도된 뉴스를 구분하고자 한다면 보다 정밀한 방식을 시도할 필요가 있다. 본 글에서는 마치 로봇 저널리즘처럼, 기존에 제시된 기사들을 기반으로 새로운 자연어 생성 기사를 만들어내는 데이터를 구축하여 이를 일반화된 보도 패턴으로 설정하고, 해당 패턴을 기준점 삼아 뉴스 데이터를 구분하여, 이후 지도학습에 있어서 시간과 자원을 줄일 수 있는 삼각 측정 방식을 제안하고자 한다.

1) 자연어 생성 방식을 통한 일반화된 보도 패턴 데이터 만들기

자연어 생성 방식을 통해 일반화된 기사 특질을 가진 기사를 만들기 위해서는 가장 먼저 분석 대상이자 자연어 생성을 통해 만들고자 하는 기사 예제 보도를 수집해야 한다. 전수 조사가 가장 좋으나 이런 점이 불가능할 때에는 기사를 수집할 때 기간, 주제, 보도 기관 등에 대한 대략적인 모수 개념을 활용하여 자연어 생성에 기초가 되는 기사를 계층적 샘플링 방법으로 모으는 방법이 있다. 수집된 기사는 특정 기간, 주제 또는 보도 기관 수집 방법에 따라 해당 분류의 표준적인 기사 형태로 취급할 수 있어야 한다. 이러한 기사 수집은 전수 수집도 가능하나 표본 추출 기법에서 몬테카를로 마코프 체인(Monte Carlo Markov Chain) 기법 등 다양한 방식을 동원할 수 있다(Van Ravenwaaji, Cassey, & Brown,

2016).

그 다음 수집된 기사에서 자연어를 생성하는 방식으로 새로운 기사를 만드는 것이다. 이 경우 자연어 생성이 특정한 기준에 따라 만들어지는 것이 아니라 기존에 존재하는 데이터를 기반으로 학습하여 만들어지는 것이므로, 생성된 기사는 수집된 기사들의 일반적인 표현을 담은 대표적인 형태의 내용을 담게 된다. 즉, 자연어 생성을 통해 표준적으로 작성된 기사의 기준점을 잡는 개념이 된다. 최근 화제가 되고 있는 ChatGPT를 생각하면 될 것이다.

기존 연구들은 자연어 생성을 통한 뉴스 기사 작성이 독자들에게 매우 자연스럽게 여길 수 있을 정도로 기술이 발전하였음을 밝힌 바 있다. 예를 들어, 기계가 작성한 뉴스를 사람들이 기계에 의한 작성인지 인지하는지 그리고 신뢰하는지 여부를 측정한 결과 기자가 작성한 기사와 사람이 작성한 뉴스간에 신뢰성 차이가 없음을 밝힌 바 있다 (Van der Kaa & Krahmer, 2014). 또한 자연어 생성을 통해 일부러 뉴스내 편향이 있는 가짜 뉴스를 만든 결과 사람들이 자연어 생성을 통한 뉴스가 부자연스럽다고 느끼지 못하고 자연스럽다고 느꼈으며, 편향도 감지해 내는 결과 역시 보고되었다(Gupta, Ngyuen, Yamagishi, & Echizen, 2020). 따라서, 자연어 생성을 통한 뉴스 기사 자체는 문제가 되지 않는다고 볼 수 있다.

기사가 생성된 후, 실제 기사와 어느 정도 유사한지 비교 평가하는 것이 그 다음 과정이다. 여기서는 단순히 생성된 기사에 기반하여 볼 때, 실제 보도된 기사가 생성된 기사와 얼마나 다른지에 대한 유사성만 평가한다. 여기서 유사성 지표는 하나만 쓰지 않고 다양한 지표를 활용할 수 있다. 그 다음 유사성 여부에 따라 평가 대상 기사들을 나열한 후, 유사성 지표에 따라 기사들을 분류한 후, 해당 내용에 대해 이후 평가하는 방식을 취할 수 있다.

이를 개념적으로 설명하면 A라는 기사와 비교할 때 B 기사뿐만 아

니라 C 기사 역시 A기사와 유사성이 매우 낮을 수 있는데, 이러한 낮은 유사성이 반드시 B와 C 사이의 유사성이 높게 존재한다고 볼 수 있는 증거는 아니라는 뜻이다. B와 C간에 차이가 극명함에도 불구하고 A와 B 사이의 유사성, A와 C 사이의 유사성이 비슷한 정도로 낮을 수 있으므로 B와 C 가 어느 정도 유사한지 역시 측정해야 한다는 것이다. 따라서 하나의 유사성을 보는 것이 아니라 다양한 각도로 유사성을 측정하는 방식이 필요하게 된다.

결과적으로 수집된 데이터에 기반하여 자연어 생성 기사를 만들면 해당 학습 데이터가 가지고 있는 일반화된 대표적인 보도 기사를 만들 수 있고, 이를 기준삼아 뉴스를 유형화하여 구분한다면 뉴스 데이터에 대한 라벨링을 쉽게 부여할 수 있기에 학습 효율성을 가지고 대용량 데이터 학습을 그대로 적용할 수 있게 된다. 또한 사전 데이터 분류 작업이 필요없기 때문에 한번 만들어진 모듈은 범용성이 높은 장점마저 지니게 된다. 데이터 기반 자연어 생성이기에 본 고에서 예를 들고자 하는 감염병 관련 백신뿐만 아니라 보건 관련 다른 사례에 적용할 수 있으며, 더 나아가 다양한 방식으로 활용될 여지가 있다.

2) 자연어 생성 방법 두 가지 방법

그렇다면 자연어 생성을 어떻게 할 것인가? DA와 관련된 자연어 생성 방법은 첫 번째 방법은 단어기반 접근법으로 진행할 수 있다. 이 방법은 작성된 글의 단어를 기반으로 크게 빈도 기반 계산, 기계학습을 통한 통계적 확률기반의 2가지 방향으로 특정한 문장 생성을 진행하는 방법을 생각할 수 있다.

먼저 빈도 기반 계산법은 모든 기사에서 보편적으로 통용되어 많이 나오는 단어가 있다는 가정하에 해당 단어를 생성되는 기사에 반드시 포함시켜야 할 특질로 보는 것이다. N-gram 방식과 Word2vec의 방식으로 볼 수 있다. N-gram 방식은 문자열을 쪼개서 특정한 길이 내 나타나

는 단어들로 텍스트 내용을 파악하는 방법이고, Word2vec은 단어들의 어울림을 기반으로 단어의 배열로서 파악하는 방법이다(Yadav & Borgohain, 2014). 예를 들어, 어떤 기사가 다양한 내용을 담고 있다면 '백신', '효능', '평가', '기대이익', '부작용' 등의 단어들이 모두 어울리겠지만 단순히 사실만 보도한다면 '백신', '효능', '%' 등만이 짧게 서술될 수 있다고 볼 수 있다. 그러나, 단어 기반 접근법은 맥락과 관련된 단어의 활용하면서 전체 문장을 고려한 경우보다 성능이 떨어지는 한계가 있다.

두 번째 방법은 문장 기반 접근법으로 문장 내 단어들간의 연속성을 고려하는 것이다(Sequence Model). 이는 문장을 기반으로 한 패턴을 파악하여 보도된 기사의 실제 서술 방식을 본따서 새로운 보도 내용을 생성하는 것으로 이때의 목표는 정확히 실제 보도 기사의 패턴을 복사하는 것이 아니라 데이터로 활용된 기사내에서 일반적으로 나타나는 패턴을 찾아내어 그와 유사하게 재현하는 것이다. 이 방법에 쉽게 동원할 수 있는 방법으로 재귀 신경망(Recurrent Neural Network)이라 불리는 RNN 방식과 장단기 기억 신경망(Long Short Term Memory)라 불리는 LSTM 방식이 있다(Iqbal & Qureshi, 2020). 단어 기반 문장 생성 방법에서는 단어 출현 여부뿐만 아니라 단어들의 등장 순서를 알아보는 것이 첫 번째 단어 기반 접근 방법과 큰 차이점이며 그 과정에서 단어들의 출현 연속성(Sequence)까지 고려하게 된다. 그런데 신경망 방식에서도 미묘한 차이가 있다. RNN은 연속성에 기반하고 있지만 문장 생성에서 바로 앞 단어의 정보만을 활용하는 단점이 있다. 예를 들어, "오늘 밤에는 천둥 번개가 치기 때문에 이러한 상황에서는 우산이 필요하다"라는 문장이 있을 경우 RNN 방식의 문장 생성에서는 우산이 필요한 이유가 '천둥' '번개'와 같은 단어 때문이 아니라 '상황'이라는 단어 다음에 '우산'이라는 단어가 나와서 나타난 것으로 생각하고, 자연어를 생성할 때 '상황'이라는 단어 다음에 '우산'이 나오는 확률이 높은 것으로 판단하여 만들어낼 수 있다. 이러한 정보 소실 문제는 자연어 생성에 있어서 신경망

이 장기간의 기억이 부족하고 바로 전단계의 상황만을 이해하도록 구성되기 때문인데, 이를 해결하기 위하여 장기간의 기억이 필요한 내용을 활용하면서 단기간의 정보를 처리하는 LSTM 방식이 고안된 것이다(허윤석·강상우·서정연, 2020). 직관적으로 설명하면 LSTM은 위의 문장에서 앞의 '천둥'과 '번개' 단어를 기억한 상태에서 '상황'이라는 단어가 불필요하다는 것을 이해한다. 따라서, 뒤에 '우산'이 등장한 이유가 단순히 앞의 '상황' 때문이 아니라 '천둥'이나 '번개' 때문이라는 것을 고려하여 문장을 생성할 수 있다. 이러한 알고리즘적 특성을 고려하면, 자연어 생성을 통한 일반화된 기사 표현은 RNN보다 LSTM 방식이 더 적합하다고 볼 수 있다.

최근에는 대규모 데이터를 사전 훈련하여 전이학습을 활용하여 인코더(encoder)-디코더(decoder)를 활용한 BERT나 GPT-3 등의 모델이 각광을 받았으며, 얼마전에는 ChatGPT 등에 대한 관심이 폭발했는데, 본 연구는 소규모 데이터로 논증하고 있는 내용이 수행가능한가의 여부를 확인하는 수준에 불과하므로 LSTM으로 수행해보고자 한다.

3) 생성된 문장과 평가 문장간의 유사성 측정

일단 문장을 기반으로 글이 만들어지면 문장의 유사성에 대한 비교를 통해 기존에 보도된 기사와 비교를 해야 때문에 어떤 방식으로 유사성을 측정할 것인가는 중요한 문제라 볼 수 있다. 현재 연구에서는 단순한 몇 개의 짧은 문장 단위의 비교이므로 짧은 문장 단위로 유사성을 판정하는 두 가지 지표를 활용하기로 한다. 첫 번째 유사성 지표는 가장 직관적이고 도식적인 방식으로 측정할 수 있는 레벤슈타인 거리(Levenshtein Distance)이다 (e.g. Heeringa, 2004; Li, Zhang, & Zong, 2016). 레벤슈타인 거리가 짧을수록 유사도가 높은 방식이 된다. 직관적으로 설명해서, 레벤슈타인 거리는 특정한 두 개의 문장이 있을 경우 기준 문장을 기준으로 다른 문장의 몇 군데가 고쳐져야 실제 해당 기준 문장과 동

일한 문장이 되는지를 계산하여 거리로 표현한다. 예를 들어, "홍길동이 아버지를 아버지라 부르지 못한다"라는 표현과 "홍길동이 아버지를 부르지 못한다. 서자라서" 라는 문장이 있을 경우 앞의 문장을 기준으로 "서자라서"라고 쓰인 부분이 "아버지라"라는 표현으로 바뀌면 두 문장은 의미상으로 동일한 문장이 된다. 레벤슈타인 거리 방법은 뉴스 헤드라인이 실제 본문의 인용과 동일한 내용을 담고 있는지 파악하는 연구에서도 쓰인 바가 있듯이(백지수·이승언·한지영·차미영, 2021), 유사성 판단에서 가장 기초적으로 흔히 쓰이는 방법이다. 그러나 레벤슈타인 방법은 유사도 측정이 매우 엄격한 기준에 따라 이루어지기 때문에 문맥상 사람이 읽었을 때 비슷한 표현이라도 유사도가 떨어지는 문제가 발생하는 점은 단점으로 존재할 수 있다.

두 번째 지표로는 ROUGE 점수 지표를 활용하기로 한다. ROUGE는 텍스트를 자동요약이나 기계 번역에서 요약본 또는 번역본과 사람이 만든 참조본을 서로 대조하여 그 성능을 평가하는데 쓰는 매우 일반적인 지표다(Lin, 2004). ROUGE 지표는 ROUGE-N으로 일반화하여 표기하기도 하는데 여기서 N은 N-gram의 길이를 통해 참조본과 요약본에서 일치하는 n-gram 수를 기준으로 평가한다. ROUGE-L은 가장 긴 연속적으로 공통된 표현을 측정한다. 레벤슈타인이 문장 순서를 통해 동일한 표현이 되는지를 거리 개념으로 본다면 ROGUE 지표는 참조본과 요약본의 실제 표현이 겹치는 것을 본다. 완전히 일치할 경우 1 완전히 불일치할 경우 0이 된다. 여기서 ROGUE recall은 참조본의 단어가 얼마나 요약본의 단어들과 겹치는가를 보는 반면, precision은 요약본의 단어가 얼마나 참조본과 겹치는가를 본다. recall과 precision을 함께 봄으로써 유사성 평가를 보다 정밀하게 할 수 있다. 예를 들어, precision이 높은데 반해 recall 지표가 더 낮은 경우, 생성된 요약본이 참조본의 내용을 상대적으로 충실히 반영하지만 참조본의 단어가 요약본에 모두 등장하지 않으므로, 상대적으로 참조본 내용의 일부만이 반영한 상태가 된다. 반대

의 경우에는 요약본이 참조본의 내용을 더 잘 반영하므로 요약본이 오히려 참조본과 다른 표현을 쓰고 있다는 이야기가 된다. 전자의 경우는 요약본이 참조본의 일부가 되며 후자의 경우는 참조본보다 요약본이 오히려 더 많은 이야기를 담게 된다. 어느 방향이던 ROGUE 점수의 절대적인 수치의 높낮이는 유사성이 높아지는 것을 의미한다. 이를 본 연구의 유사성 측정에 적용하기 위해 생성된 문장을 모두에 대한 참조본으로 두고, 원본을 비교 대상으로 취급하면, Precision이 높다는 건 원본이 생성된 내용을 더 반영하지만 생성된 내용이 모두 원본에 있는게 아닌 정도가 더 높다는 의미이다. 생성된 문장과 원본의 불일치 이유가 원본에 다양한 내용이 있기 때문인 셈이다. 본 연구에서 수행하는 자연어 생성의 원리와 비교 방법상 이와 같은 방향이 나타나는 것이 자연스럽다.

이 두 지표를 통해서, 내용 유사성과 실제 진술 유사성을 함께 비교하는 방식으로 살펴볼 수 있다. 지금 제안하는 유사성 지표가 매우 좋은 최선의 방법이라기 보다는 현실적으로 접근 가능한 유효한 방법 중 쉽게 적용 가능한 방법을 적용하는 것으로 이해하면 될 것이다 .

4) 유사성 기반 구분된 기사에 대한 유형 분류 방안

새롭게 만들어진 자연어 생성 문장과 유사성에 따라 구분된 기존 기사들을 식별한 이후에는 실제 해당 기사들의 내용이 어떤 내용을 담고 있는지 평가하는 방식이 동원되어 기사를 유형화할 수 있는데, 이 방식이 가지는 장점은 크게 세 가지가 있을 것으로 예상 된다. 우선, 지표를 통해 나타난 내용을 분석함으로써 빠르게 문서들을 구분할 수 있으며, 이에 기반하여 라벨링에 기반한 지도학습을 진행하기 쉬워진다. 사전에 뉴스 분류 특질을 만들고 그에 기반한 내용 분석이나 분류를 수행한다는 것은 보도 내용이 어떨 것이라는 점을 이미 알고 있는 상태를 전제로 한다. 하지만 모든 보도 내용을 미리 알기는 불가능하므로, 데이터에 기반하여 일반화된 보도 패턴의 내용을 만들고 이를 통해 구분함으로써 이러한 사전 인식 문

제를 우회할 수 있다. 두 번째로는 뉴스 기사 분류 지도학습을 위한 데이터의 불균형성 문제도 해소할 수 있다. 앞서 언급하였듯이 데이터에 기반하여 유사성을 근거로 판단하기 때문에, 주어진 데이터 내에서 가장 일반적인 특징을 찾게 된다. 절대적인 기준에 근거하는 것이 아니라 상대적인 근거에 기반한 분류가 가능해지는 장점을 가지게 된다. 이미 쓰여진 글을 기준으로 평가하는 것이기 때문이다. 또한 자연어 생성에서 '씨앗(seed, 시드) 단어'를 부여하면서 특정 단어로부터 생성이 가능하기 때문에 생성되는 문장의 다양성을 살리는 방향으로의 생성도 가능하다. 이와 연결되어 세 번째 장점으로는 뉴스 유형화를 실제 보도된 데이터 기반으로 확인하여 살펴볼 수 있는 점을 언급할 수 있다. 예를 들어 감염병 보도에서 갖추어야 할 지점들은 알고 있으나 그것이 실제로 보도되는 경우에 대한 자료가 부재한 상황이 있을 수 있는데, 실제 보도 되는 내용으로부터 출발하기 때문에, 이러한 보도된 내용 중심 판단을 할 수 있게 된다.

그러나 이러한 방법에는 하나의 전제가 있는데, 생성된 기사의 수준이 어떤지 살펴보는 것과 계산된 유사성에 따라 배열된 기사들을 검토하여 유사성 차이가 의미하는 바를 확인하여야 한다는 점이다. 예를 들어, 본 고에서 제시하고 있는 두 가지 유사성 지표를 단순하게 표현하면 아래 〈표 1〉과 같이 생각할 수 있다.

표 1. 유사성 지표에 따른 기사 구분

		ROGUE	
		높음(유사)	**낮음(상이)**
Levene D	**낮음(유사)**	A	B
	높음(상이)	C	D

위의 표에서 A의 경우는 원본과 생성본의 내용 유사도가 높으면서 동시에 글이 쓰인 표현의 유사도 역시 높은 수준이므로 A에 속하는 기사의 경우는 가장 일반화된 정형화된 기사 보도 형태를 따르고 있다고 볼

수 있다. 만약 이렇게 일반화된 기사 보도 형태가 기사가 많이 존재하게 된다면, 보도 기사를 천편일률적으로 그대로 보도했다든가 서론에서 제시한 예시처럼 어뷰징이 많다는 뜻이 될 수도 있다. D는 이것에 완전히 반대되는 경우로 동일 주제를 다룬다고 하는데 일반화된 보도 패턴과 내용과 표현에서 완전히 벗어나는 경우다. 극단적인 경우로 완전히 다른 내용을 보도하는 것이 되는데 기사가 담고 있는 정보의 내용이 정형화된 틀을 완전히 벗어나고 표현 자체도 다르므로, 뉴스가 담고 있는 프레임이 일반적이지 않다는 뜻이 된다. B의 경우는 n-gram 방식으로 측정하는 ROGUE 지표의 정의상 다루는 내용의 유사도가 매우 낮은데 표현이 유사하는 경우라는 의미가 되는데 현실에서 찾아보기 어려울 것으로 생각된다. C는 보도된 기사들의 일반적인 내용을 담고 있으면서도 표현이 다르다는 것인데 정치적 지향에 따른 기사 내용 구성상의 표현 차이가 있을 경우라든가 하는 경우에 나타날 수 있다.

정리하면, C에 해당 하는 기사들에 대해서는 내용상의 정밀 검증이 필요한 반면, D의 경우는 주제와 내용이 완전히 다르다는 점에서 정밀 검증이라기 보다는 실제 왜 그렇게 나타나는지에 대한 분석이 이루어질 필요가 있는 등 다양한 각도로 접근할 수 있다. 또는 D에 해당하는 기사만 다시 모아 분석을 다시 수행할 수도 있다. 결과적으로 현재 제시하고 있는 방법은 향후에 지도기계학습을 수행하기 위한 사전 정지작업으로서의 의미를 가지며 이 과정을 통해 보도된 뉴스를 빠르게 유형화함으로써 상당한 자원 투입을 줄이기 위한 방식으로 이해할 수 있다.

4. 실제 데이터를 활용한 자연어 생성 결과

그렇다면 이렇게 제안한 방법을 실제로 활용하여 분석을 하게 되면 어떠한 결과가 나타날 수 있을까? 이번 장에서는 실제 데이터를 활용하여 자

연어 생성을 수행하여 본고에서 제안하고 있는 방법이 적용 가능한지를 탐색해 보고자 한다.

1) 자료와 연구 방법

본 연구에서 실제로 구현할 자료는 한국언론진흥재단의 데이터베이스인 빅카인즈에서 다운로드를 받은 자료이다. 2021년 10월 1일부터 2022년 2월 28일까지 기간으로 설정하여 "백신 부작용"으로 검색하여 나타나고 경향신문과 한겨레 신문의 보도 150건을 활용하였다. 빅카인즈는 최대 200자까지 선도 본문을 제공하고 있는데, 이를 기반으로 하여 자연어 생성을 진행하였다. 한겨레와 경향신문만을 생성한 이유는 우선 보도 내용 자체가 많지 않아서 본 고에서 제안하는 방법이 가질 수 있는 장점과 한계를 뚜렷이 보여줄 수 있을 것으로 생각했기 때문이다.

본 연구는 제안한 방법의 가능성을 보여주기 위한 것이므로, 자연어 생성을 위한 LSTM 모형의 레이어를 1개만 활용하였으며, 해당 배치(batch)의 크기는 256으로 설정하였다. 활성 함수로는 소프트 맥스(softmax)를, 옵티마이저는 RMSprop을 사용하였다. 다양한 결과를 살펴보기 위해서는 LSTM 생성 모형의 하이퍼파라미터인 다양성(diversity)는 0.2, 0.5, 1.0, 1.2로 다양하게 설정하였다. 이 숫자가 커지면 커질수록 자연어 생성에서 확률상으로 나타나는 경우가 적은 수의 말들이 더 많이 만들어져서 다양한 자연어 문장이 만들어진다. 그런데, 문장을 생성해보니 학습 자료가 적어 다양성이 적은 0.2 경우가 가장 적실한 문장을 만들어내는 것으로 판단되었다. 반복학습을 의미하는 에포크(epoch)는 100회로 설정하였다.

자연어 생성과 관련된 복잡한 모형을 만든 것이 아니고, 또한 매우 큰 데이터를 가지고 학습한 것이 아니라 소규모의 데이터로 학습하여 만든 자연어 문장 생성이기 때문에, 조사의 사용이나 단어의 연결이 매끄럽지 않았다. 따라서 연구자가 원문의 뜻을 살리기 위해 조사를 약간 보정

하였다. 본 논문의 목적이 자연어 생성이 얼마나 자연스럽게 제대로 되는가가 아니라, 기존의 보도들을 학습하여 새롭게 생성한 자연어 문장의 내용, 즉 보도 기사 구분을 위한 기본적인 레퍼런스를 만드는 것이기 때문에 이러한 조정에 무리가 없을 것으로 판단한다. 문장이 생성된 이후에는 만들어진 문장을 가지고 실제 보도내용과의 유사성 평가까지만 진행해 보고자 한다.

2) 생성된 '백신 부작용' 관련 보도의 내용

우선 생성된 문장들이 학습 데이터가 적고 딥러닝 모델이 매우 정교하게 짜여진 것이 아니라서 높은 수준의 문장이 아니었다. 예를 들어, 아래 문장을 보면, 문장마다 그 내용이 끊기면서 서로 상이한 내용들이 별다른 연결고리가 없는 상태에서 연결되는 식이다 .

> 백신 접종 후 부작용 이상반응에 대한 보상 범위도 확대하겠다는 계획을 밝혔다. 방역 당국은 백신 접종 사건에 대해 연이어 나섰다. 부작용 수집 항목에 '월경 장애'를 추가할 방침인 것으로 6일 확인됐다. 질병과 사망 등 신체 이상 반응 이상이 쏟아지고 있다
>
> (Epoch 59, Diversity 0.2)

비록 에포크가 증가하면서 문장의 자연스러움이 향상되고 그에 따라 읽을 수 있는 수준의 말로 향상되는 측면도 있었지만 150건의 100번 에포크 그리고 단순 LSTM 모델에서 어느 정도 예상되는 매우 일반적인 내용들의 연결이었다. 그럼에도 불구하고, 생성된 자연어 보도 내용 중에서 에포크가 어느 정도 진행된 90회 이후 생성된 문장을 무작위로 하나를 선택하여, 자연어 생성의 결과를 활용하기로 하였다. 무작위로 선택한 결과 93번째 에포크에 생성된 0.2 다양성을 가진 문장이 선택되었는데, 이용한 학습 데이터의 기본 서술 단위와 맞추기 위해 94번째 에포크에 생

성된 문장과 함께 연결지었다. 해당 문장들은 '백신 미접종자'라는 무작위 추출 씨앗 단어를 기반으로 만들어 졌다. 이를 통해 완성된 문장은 아래와 같다.

> "위중증 사망에 이른 확진자 중 백신 미접종자가 과반이라는 통계치. 학생들 기준을 단계적 일상회복, 1차 청소년 방역패스를 적용하면서 논란이 뜨거웠다. 현재 미국에선 화이자의 코로나19 백신을 12세 이상까지 접종할 수 있는 가운데 정부가 방역 패스 추진하며 조건부로 안전성을 보고 청소년 방역패스를 적용하면서 발생했다. 이르면 9월로 접종이후 시작될 단계적 일상 회복 지원을 일부 코로나19 방역패스 효력과 함께 적용한다."

위 문장에서 가장 중요 단어가 '청소년'과 '방역 패스'라는 점에 착안하여 해당 두 단어를 가지고 있는 기사 리드를 추출하였더니 자료 중 16개 기사 리드에 이에 해당하였다. 전체 기사 150건으로 학습을 했으므로, 약 10%에 해당하는 기사가 위의 생성된 문장의 내용과 직접적인 관련을 가진 셈이다. 이제 문장이 생성되었으므로, 해당 문장과 학습에 쓰인 원래 기사간의 유사성을 평가하기로 하였다. 16개 기사 리드에 대한 레벤슈타인 거리를 측정하였고 또한 ROGUE 점수 역시 측정하여 〈표 2〉에 제시하였다. ROGUE 점수의 경우는 ROGUE-L의 Recall과 Precision을 표기하였다.

〈표 2〉를 보면 우선 3번 기사가 레벤슈타인 거리가 가장 짧으며 또한 ROGUE 점수 역시 가장 높은 것으로 나타난 것을 알 수 있다. 기사인 3번 기사를 보면, 단어 사용에 있어서 '청소년'과 '방역패스' 뿐만 아니라 생성된 글에 있는 '논란' 및 '지원' 등 주요 단어가 모두 언급 되어 있는 것을 확인할 수 있다. 그러나 일상 회복과 관련된 이야기는 찾아볼 수 없는데 이는 청소년과 방역패스와 관련된 16건 전체 기사에서도 찾아보기

어려운 부분으로 학습된 기사의 다른 부분의 내용이 생성 과정에서 만들어진 것으로 볼 수 있다. 실제로 ROGUE Precision 점수가 Recall 점수보다 모든 기사에서 더 높다(기사 10번은 반올림의 영향). 원본이 생성된 내용을 반영하지만 생성된 내용이 모두 원본에 있지 않다는 것을 반영하는 것이기도 하다.

이번에는 레벤슈타인 거리가 비슷한데 ROGUE-L 점수가 다른 7번과 9번을 보기로 한다. 두 개의 경우는 앞서 표 1의 D에 가까운 것이라 볼 수 있다. 이 둘의 기사 내용은 매우 흥미로운데, 7번의 경우는 김부겸 국무총리가 청소년 대상 방역패스 적용이 불가피함을 이야기하는 내용인 반면, 8번의 경우는 유은혜 교육부 장관이 유연하게 적용 시기를 살피겠다는 이야기이다. 본래 생성된 문장이 방역 패스를 적용했다는 이야기라는 점에서 적용한다는 내용이 담긴 김부겸 국무총리 발언이 ROGUE-L가 더 높게 나온 것을 알 수 있다. 반면에 9번의 경우는 정부가 적용을 고민한다는 이야기로 7번의 내용과 반대의 내용이 된다. 즉, 유사성 2개의 내용으로 보면 ROGUE-L 점수 차이는 내용상의 차이를 보여준다는 의미가 된다.

이러한 결과는 기사 유형 분류를 위해 자연어 문장을 생성할 때 나타날 수 있는 매우 중요한 사실을 확인하여 주고 있다. 즉, 구체적인 상황 - 예를 들어, 가처분 신청이라든가, 교육부 간담회 등 - 이 아닌 일반적인 내용만을 담고 있는 기사가 생성된 기사와 유사했고 구체적 내용이 담길수록 유사성이 떨어져서 생성된 기사와 멀어지는 것을 확인해 준다. 적어도 생성되는 기사는 일반적인 내용으로 만들어지고, 기사의 구체성이 더해지면 이로부터 유사성이 떨어질 수 있다는 점을 보여준 셈이다. 보도된 내용을 가지고 일반적인 기사 보도의 기준점을 잡고자 문장을 생성했는데, 실제 생성된 문장이 그러한 목적을 어느 정도 달성하고 있는 것으로 평가할 수 있다.

표 2. 레벤슈타인 거리 및 ROGUE-L 유사성 점수

No.	기사 리드	문자수 (A)	Lev D (B)	(B)/(A)	Rouge-l	
					Recall	Precision
1	THIS Covery 팀이 정부의 청소년 방역패스 적용 방침을 둘러싼 논란을 짚어봤습니다. 정부는 지난 3일부터 신규 방역패스 의무 적용 시설에 학원과 독서실, 도서관을 추가했는데요, 내년 2월부터 12~18살에게도 방역패스를 적용하기로 했습니다. 청소년 역시 백신 접종을 하지 않으면 이들 시설을 이용할 수 없게 되는 겁니다. 백신 접종 부작용이.	195	182	0.933	0.08	0.10
2	법원이 서울의 12~18세 청소년 방역패스 효력을 정지하면서 방역패스 도입을 통해 청소년 미접종자의 백신 접종을 유도하려던 정부 구상도 차질이 예상된다. 당장 청소년 백신 접종 참여에 영향을 미칠지 주목된다. 서울행정법원 행정4부는 조두형 영남대 의대 교수와 의료계 종교인 등 1000여명이 서울시장을 상대로 낸 집행정지 신청을 14일 일부 인용하면서.	198	186	0.939	0.08	0.10
3	24일까지 학생 백신 접종하면 내년 2월1일 적용 무리 없어 전면 등교도 유지 접종시일 촉박 기말고사 지장 지적도 청소년 방역패스(접종증명 음성확인제) 적용을 둘러싼 논란이 커지고 있는 가운데 정부가 학습권보다 학생들의 안전이 우선한다며, 대부분 학교의 기말고사가 끝나는 오는 24일까지 학생들의 코로나19 백신 접종을 집중 지원하겠다고	188	178	0.947	0.14	0.18
4	서울행정법원이 14일 조두형 영남대 의대 교수 등 1023명이 서울시장 등을 상대로 낸 방역패스(백신접종증명 음성확인제) 집행정지 신청을 일부 받아들였다. 서울시내 3000㎡ 이상 마트 백화점 상점에 적용한 방역패스 조치의 효력을 정지하고, 12~18세 청소년에 대한 다중이용시설 방역패스 적용을 중지시킨 것이다. 이번 결정으로 식당 카페 영화관 PC방 .	199	189	0.950	0.04	0.05
5	법원이 서울지역 상점과 마트 백화점에 대한 방역패스(백신접종증명 음성확인제) 정책에 제동을 걸었다. 서울시가 12~18살에게 확대 적용하려던 청소년 방역패스도 일시 정지된다. 시민 1천여명이 다중이용시설에 대한 방역패스 효력을 중단해 달라며 낸 집행정지 신청을 법원이 일부 받아들이면서 본안 판결이 나올 때까지 해당 시설과 청소년을 대상으로 한 방역패스	197	189	0.959	0.06	0.08

6	12~17살 백신 접종완료율이 43.8%(20일 0시 기준)에 머무는 가운데, 소아 청소년 방역패스(접종증명 음성확인제)가 위법하다며 행정소송을 제기한 단체들이 주최한 기자회견에서 백신 관련 가짜뉴스가 나와 확산된 것으로 뒤늦게 확인됐다. 전국학부모단체연합, 함께하는사교육연합 등은 지난 17일 서울행정법원 앞에서 기자회견을 열어 정부가 청소년.	194	188	0.969	0.06	0.08
7	김부겸 국무총리가 논란이 일고 있는 청소년 코로나19 백신 접종과 방역패스 추진에 대해 11일 정부가 욕 좀 덜 먹자고 우리 청소년들의 목숨을 담보로 잡을 수 없었다고 밝혔다. 김 총리는 최근 코로나19 확산과 정부 대응을 두고 제기되는 각종 의문들에 상세한 설명을 내놨다. 김 총리는 이날 밤 사회관계망서비스(SNS)에 올린 장문의 글에서 솔직히.	195	189	0.969	0.12	0.13
8	정부가 내년 2월부터 학원 등 다중이용시설에 청소년들의 방역패스 적용을 예고한 가운데 교원단체들이 잇따라 강도높은 비판을 쏟아내고 있다. 전국교직원노동조합은 6일 청소년 방역패스와 관련해 성명을 내고 백신 접종을 강요하는 각종 정책을 중단하라고 주장했다. 전교조는 성명에서 찾아가는 학교 단위 백신 접종, 방역패스 적용 시설 확대 및 청소년	191	186	0.974	0.12	0.14
9	청소년 방역패스(접종증명 음성확인제)에 대한 반발이 여전한 가운데 유은혜 사회부총리 겸 교육부 장관이 적용시기는 물론 범위까지 충분히 논의해 결정하겠다는 뜻을 밝혔다. 감염병 전문가들도 정책의 필요성을 인정하면서도 적용시기 연기나 범위 조정의 필요성을 언급했다. 유은혜 사회부총리 겸 교육부 장관은 13일 저녁 〈한국방송1〉(KBS1) 긴급진단.	194	189	0.974	0.02	0.03
10	정부가 내년 2월부터 식당 카페 학원 등에서 12 18세 청소년에게까지 확대 적용할 예정인 방역패스(접종증명 음성확인)가 위헌이라며 고3 학생 등이 헌법소원심판을 내기로 했다. 고교 3학년생 양대림군(18)은 452명의 청구인을 모아 오는 10일 헌법재판소 앞에서 기자회견을 연 뒤 방역패스에 대한 헌법소원심판을 청구할 예정이라고 밝혔다. 양군 측은.	196	191	0.974	0.02	0.02
11	내달 1일부터 시행될 예정이었던 경기도의 청소년에 대한 방역패스(백신접종증명 음성확인제) 효력이 일시 정지됐다. 수원지법 행정2부(부장판사 양순주)는 17일 백신패스반대국민소송연합 소속 회원 등 경기도민 256명이 경기도지사를 상대로 낸 방역패스 처분 취소 집행정지 신청을 일부 인용했다. 경기도에서 12세 이상 18세 이하를 대상으로 한 청소년 방역 패스.	200	196	0.98	0.10	0.12

12	정은경 질병관리청장이 13일 코로나19 의료대응 역량을 확충하고 3자 접종으로 고령층의 면역도를 대폭 올려놓는 데까지 시간이 필요하다며 사적모임 인원 축소와 다중이용시설 영업시간 제한 등 물리적(사회적) 거리 두기 강화를 검토 중이라고 밝혔다. 유은혜 부총리 겸 교육부 장관은 내년 2월로 예정된 청소년 방역패스 적용을 두고 현장과 소통을 강화하면서.	196	194	0.99	0.06	0.07
13	청소년 방역패스(접종증명 음성확인제)가 2022년 3월부터 시행된다. 학원과 학부모 반발을 고려해 당초 계획보다 한 달 미뤄진 것이지만 현장에서는 여전히 반발의 목소리가 높다. 중앙재난안전대책본부(중대본)와 교육부는 31일 오전 정부서울청사에서 열린 코로나19 정례 브리핑에서 3월1일부터 12~17세를 대상으로 청소년 방역패스제를 시행한다고.	192	191	0.995	0.04	0.06
14	학원이나 독서실에는 방역패스를 적용하면서 종교시설이나 백화점, 놀이공원에는 왜 방역패스 적용 안하나(중학교 3학년 학생) 성장기 청소년과 어른의 백신 투여량이 같을 수 있나. 영국은 청소년에게 2회가 아니라 1회만 접종하는 것으로 안다(학부모) 교육부가 8일 서울 양화중학교에서 소아 청소년 백신 접종과 관련한 학생 학부모 전문가 간담회를	189	189	1	0.10	0.12
15	애들이 학원 안 갈 수 있나요? 강제 (접종)이네요, (접종률 제고에) 학생들 필수시설인 학원을 이용하다니요. (온라인 학부모 커뮤니티 글 일부) 내년 2월1일부터 적용되는 청소년 학원 방역패스(접종증명 음성확인제)가 학습권을 침해한다는 학원 및 일부 학부모들의 반발에 부딪히면서 논란이 커지고 있다. 이번 논란은 근본적으로 높은 사교육	189	192	1.016	0.10	0.13
16	청소년 방역패스 논란 일자 교육부, 간담회 자리 마련 자녀 2명 접종 후 부작용 병원 안전 범위라는 말뿐 학부모도 불안 불만 쏟아내 학원은 방역패스를 적용하면서 백화점, 놀이공원에는 왜 적용 안 하나요.(중학교 3학년생) 교육부가 8일 서울 양화중학교에서 소아 청소년 백신 접종과 관련한 학생 학부모 전문가 간담회를 열었다. 내년.	185	191	1.032	0.10	0.11

만약 보도되는 내용을 기반으로 기사 분류를 진행하고자 한다면, 기사를 생성할 때 다양한 조건을 주어서 기사를 생성하고 이를 기반으로 판단할 수 있을 것으로 보인다. 예를 들어, '정보원'을 중요시하고 위와 같이 '김부겸' 이라든가 '유은혜' 라는 인물 보도가 중요한 것이라고 판단했다면 해당 단어들을 씨앗(seed)으로 하여 자연어 생성을 하고 유사성을 측정하는 방식으로 기사를 분류할 수 있다. 또한 '논란'이라는 표현을 중요시하는 경우라면 이를 바탕으로 생성할 수도 있다. 무엇보다, 하나의 글에 대해 다양한 잣대로 유사성을 중복 측정하여 - 예를 들어, '정보원' 출현 여부와 '논란'이라는 단어 출현 기사 등 - 복합적으로 측정함으로써, 기사의 유형 분류 하는데 활용할 수 있을 것으로 보인다.

특히 한정된 내용의 데이터로 생성했으나, 대규모의 데이터로 생성된 ChatGPT보다 훨씬 더 정확히 청소년 방역 패스와 관련된 사항들을 설명하고 있다. 참고로 ChatGPT를 통해 '방역 패스 청소년 논란'이라는 자연어 생성 씨앗을 주었더니 ChatGPT는 아래와 같은 문장을 생성해 내었다.

> "방역 패스란, 코로나19로 인해 격리된 곳에서 이동할 때 필요한 공식적인 증명서를 말합니다. 청소년들을 대상으로 한 논란은 이들이 방역 패스를 얻기 위해 가짜 증명서를 사용하는 것을 말합니다. 이러한 행위는 범죄이며 청소년들이 이를 할 경우 처벌을 받을 수 있습니다. 그러나 이러한 행위를 하는 청소년들은 일반적으로 정부의 규제가 강해지면서 일어나는 일들이기 때문에, 정부는 이러한 문제를 해결하기 위해 적절한 대책을 수립해야 합니다."
>
> ChatGPT, 2023/1/09일 생성결과

위의 내용이 실제 청소년 방역 패스 논란과 하등 관계가 없는 내용들을 연결지은 것임을 알 수 있다. 결과적으로 관련 없는 모든 사항들을 가

지고 학습하여 문장을 생성한 것보다, 정확히 신문 기사를 가지고 생성한 경우가 정확한 내용을 담고 있다. 이는 자연어 생성을 통해 뉴스 기사의 일반적인 내용을 추정하는 과정은 데이터의 크기의 문제이기보다는 전체 데이터가 어떠한 보도 내용을 공통적으로 가지고 있는 것인가가 중요한데, 그런 측면에서 뉴스 기사를 데이터로 하여 만들어내는 자연어 문장은 공통의 아젠다를 가진 기사에 대해서 매우 정확한 결과를 낼 수 있음을 보여준다.

정리하면, 자연어 문장 생성을 통해 현재 보도된 기사를 일반적으로 요약할 수 있는 기사를 만들어내고 이를 중심으로 보도된 기사들과의 유사성을 비교하는 작업은 충분히 향후 지도학습을 통해 기사를 분류하기 위한 사전 작업으로 적용가능한 것으로 생각할 수 있다. 특히 복합 시드로 기사를 생성하여 뉴스 기사를 비교할 수 있다면 자연어 생성 방법을 제안한 주요 이유인 자원과 시간의 절약에 상당한 효과를 볼 수 있을 것으로 보인다.

5. 토론

지금까지 본 연구는 자연어 생성을 통해 일반화된 보도 패턴을 파악하고 이를 기반으로 보도된 내용과의 유사성을 비교한 후 기사 유형화 분류에 기여하는 방식을 제안하였다. 이 과정에서 자연어 생성을 복합적으로 수행함으로써 다양한 데이터를 구축하는 방식을 제안하였다. 이러한 제안이 실질적으로 가능하다는 점을 보이기 위해 기초적이기는 하지만, 감염병과 관련된 주제로 실제 자연어 생성을 하고, 어떠한 결과가 나오는지 살펴보았다.

그동안 기사 내용 분석에 따른 유형화 방식은 사전에 내용 분석 항목을 나눈 후 그에 따라 내용 분석을 진행하는 방식과 학습 데이터에 대한

평가를 진행한 후 기계학습 방법으로 텍스트를 분류하는 두 가지 방식을 활용하여 왔다. 하지만, 이러한 방법은 뉴스 특질에 대한 사전 지식과 사전 지식에 기반한 코딩 그리고 데이터 규모의 균형성 등 상당히 많은 조건이 전제되어야 가능한 방법이며 실제 적용과정에서 시간과 자원이 많이 들게 된다. 본 연구에서는 이러한 한계를 극복하기 위해서 데이터 기반으로 자연어를 생성하고 이후 지도학습으로 이어나가는 방법을 제안하고 있다. 실제로 일련의 데이터로 분석하여 본 결과, 초보적이기는 하나 이러한 방법이 가능하다는 점을 확인하였으며 감염병 백신 부작용 관련 기사외에 앞으로 다른 경우에서도 적용될 수 있을 것으로 예상된다.

본 연구에서 실험한 감염병 등 건강 정보와 관련하여서 보다 구체적으로 서술하여 보면 이러한 방식이 가진 장점이 상당한 것으로 보인다. 가장 먼저, 기존에 보도되는 방식의 일반화된 보도 패턴을 파악하여 현재의 문제점을 진단하는 데 도움이 될 수 있을 것으로 보인다. 지금까지 건강 관련 보도를 분석하는 내용 분석의 경우 대부분 내용 분석의 형태를 띄고 담론 주제, 프레임 분석, 정보원 분석 등과 관련된 항목에 치중하는 경향이 높았다(예. 임유진·강승미, 2021; 홍주현·차희원, 2018). 따라서 구체적인 표현과 관련된 사항보다는 특정 요소 분포에 대한 판단을 사람이 직접적으로 수행해야 했다. 이 경우 소규모 데이터에 대한 판정은 어렵지 않으나 데이터 크기가 커질수록 분석이 어려워지는 단점이 있다. 뿐만 아니라 구체적인 표현을 중심으로 보지 않기 때문에, 내용의 방향성이나 프레임 등을 지적할 수는 있어도 구체적으로 보도 형태의 방식과 표현에 대해서는 다루지 못하는 한계가 있다. 그러나 본 연구에서 제시하는 방식은 이러한 한계를 개선하는 방안이 될 수 있다는 점이 고무적이라 하겠다.

두 번째로, 건강 정보와 관련된 정보의 정확성 판정과 관련하여 앞으로 가짜 정보나 오정보를 탐지하는 방식으로 발전해 나갈 수 있는 가능성이 있다. 최근 상당히 많은 연구들이 가짜 뉴스를 자동으로 탐지하기 위

한 자연어 처리 방법을 고안하고 있다(Oshikawa, Qian, & Wang, 2018). 본 연구 방법을 응용하면 가짜 건강 정보, 잘못된 건강 정보의 내용과 실제 정보간의 비교로까지 발전할 수 있을 것으로 생각된다. 실제로 가짜 뉴스 탐지에서 LSTM 방법을 활용한 방법이 연구되고 있는 만큼(Bahad, Saxena, & Kamal, 2019), 뉴스 구분 및 유형화 뿐만 아니라 잘못된 정보, 허위 정보 등의 탐지와 관련해서도 유용한 기술이 축적될 수 있을 것으로 생각된다. 예를 들어, 잘못된 정보와 실제 정보를 비교하는 방법, 잘못된 정보의 표준적인 표현을 패턴화하여 생성하는 방법 등 매우 다양하게 활용될 수 있다.

본 연구에서 자연어 생성을 시연하여 보여주고 제안하고 있지만 여러 한계에서 무척이나 자유롭지 않다. 우선 본 연구는 실제 생성하여 보여주기 위한 학습 데이터 수가 많지 않아 학습의 효율성이 떨어지고 그로 말미암아 본 연구에서 시연한 내용의 일반성을 향후 더 검증해야 할 한계를 가진다. 본 연구에서 학습 데이터 숫자를 줄인 이유는 적은 수로 학습하여 어떠한 모습으로 설정한 내용이 나타나는지 향후 어떠한 문제점이 예상되는지 살펴보기 위한 방법이기는 하였으나, 향후 대용량의 데이터와 트랜스포머에 기반한 발전된 알고리즘을 적용하여 더욱 자세히 살펴볼 필요가 있다. 두 번째 한계는 실제 지도기계 학습까지는 시연하지 못하고, 바로 그 전 단계까지 수행해야 하는 일의 과정과 방식 그리고 예상되는 결과까지만 보여주고 있다는 점이다. 실제로 실행 과정에서 나타날 수 있는 지점들, 예상하지 못하는 결과들까지 서술하고 있지 못하는 단점을 크다. 그리고, 본 연구에서 제시하고 있는 예시가 매우 제한된 기간동안 보도된 특정한 주제를 다루고 있다는 점 역시 한계로 지적될 수 있다. 보다 일반적인 주제로 시간 간격이 넓어지는 기사 보도가 있을 경우 현재 제안한 방법이 가능할 것인가에 대해서는 여전히 경험적으로 검증해야 할 필요가 있을 것으로 생각된다. 마지막으로 본 연구는 다양한 연구 방법 중 하나의 제안에 불과하다는 점을 상기할 필요가 있다.

참고문헌

김경모·박재영·배정근·이나연·이재경 (2018). 〈기사의 품질: 한국일간지와 해외 유력지 비교 연구〉. 서울: 이화여자대학교 출판문화원.

김동환·이준환 (2015). 로봇 저널리즘: 알고리즘을 통한 스포츠 기사 자동 생성에 관한 연구. 〈한국언론학보〉, 59권 5호, 64-95.

네이버 검색 (2019). 개선되는 네이버 뉴스 검색 모델을 소개합니다. URL: https://blog.naver.com/naver_search/221453678530

네이버 검색 (2021a). 네이버 뉴스 추천 알고리즘에 대해. URL: https://blog.naver.com/PostView.naver?blogId=naver_search&logNo=222439351406

네이버 검색 (2021b). 뉴스 검색 '관련도순' 배열 결과에 대해 설명드립니다. URL: https://blog.naver.com/PostView.naver?blogId=naver_search&logNo=222446236905

박재영·이완수 (2010). 〈뉴스평가지수의 개발과 적용〉. 서울: 한국언론재단.

백지수·이승언·한지영·차미영 (2021). 기계학습 기반 국내 뉴스 헤드라인의 정확성 검증 연구. 〈제33회 한글 및 한국어 정보처리 학술대회 논문집〉, 281-286.

유봉석·최재호·최창렬 (2020). 네이버 뉴스 알고리즘 이렇다. 〈관훈저널〉, 157호, 48-56.

유수정·이건호 (2020). 방송뉴스의 단독 보도 품질 연구. 〈한국방송학보〉, 34권 3호, 174-210.

윤호영 (2022). 〈뉴스공장〉에 대한 언론사 보도 분석: 신문사별 보도 경향 분석. 〈정치커뮤니케이션 연구〉, 67권, 75-113.

이선민 (2020). 클릭유도성 뉴스가 일반적인 뉴스의 이용에 미치는 효과. 〈한국언론정보학보〉, 102권, 160-188.

임유진·강승미 (2021). 비만에 대한 국내 미디어 뉴스 내용 분석 연구: 건강 신념 모델 (Health Belief Model) 의 적용. 〈홍보학 연구〉, 25권 2호, 135-159.

정재철·이종혁 (2022). 한미동맹 보도에 대한 의제 도출과 '동맹-자주'관점의 비교 분석: BERT 모델 기반 딥러닝 모형의 활용. 〈사이버커뮤니케이션학보〉, 39권 4호, 205-263.

허용강·차수연·서필교·김소영·백혜진 (2015). 감염병 보도 지침에 따른 에볼라 바이러스 언론보도 내용분석: 국내 주요 일간지를 중심으로. 〈헬스커뮤니케이션연구〉, 12권, 75-113.

허윤석·강상우·서정연 (2020). 목적지향 대화시스템에서 LSTM 언어모델 기반의 한국어 자연어 생성. 〈한국차세대컴퓨팅학회 논문지〉, 16권 3호, 35-50.

홍주현·차희원 (2018). 정부의 위기 커뮤니케이션 연구: 의약품 부작용 관련 언론 보도에 나타난 주요 주제, 정보원, 위기 책임 귀인, 프레임 분석 및 네트워크 분석을 중심으로. 〈한국콘텐츠학회논문지〉, 18권 4호, 575-585.

Anaby-Tavor, A., Carmeli, B., Goldbraich, E., Kantor, A., Kour, G., Shlomov, S., Tepper, N., & Zwerdling, N. (2020, April). Do not have enough data? Deep learning to the rescue!. In *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(5), 7383-7390.

Bahad, P., Saxena, P., & Kamal, R. (2019). Fake news detection using bi-directional LSTM-recurrent neural network. *Procedia Computer Science, 165*, 74-82.

Bayer, M., Kaufhold, M. A., & Reuter, C. (2022). A survey on data augmentation for text classification. *ACM Computing Surveys, 55*(7), 1-39.

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks, 106*, 249-259.

Chen, C., Liaw, A., & Breiman, L. (2004). *Using random forest to learn imbalanced data.* (unpublished manuscript).

University of California, Berkeley, 110(1-12), 24.

DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). *Big Data & Society, 2*(2), 2053951715602908.

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. (2021). A survey of data augmentation approaches for NLP. *arXiv preprint* arXiv:2105.03075.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines, 30*(4), 681-694.

Gupta, S., Nguyen, H. H., Yamagishi, J., & Echizen, I. (2020). Viable threat on news reading: Generating biased news using natural language models. *arXiv preprint* arXiv:2010.02150.

Heeringa, W. J. (2004). *Measuring dialect pronunciation differences using levenshtein distance*. Doctoral dissertation, University Library Groningen.

Hernández-García, A., & König, P. (2018). Data augmentation instead of explicit regularization. *arXiv preprint* arXiv:1806.03852.

Iqbal, T., & Qureshi, S. (2020). The survey: Text generation models in deep learning. *Journal of King Saud University-Computer and Information Sciences, 34*(6), 2515-2528.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S., Sustein, C., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science, 359*(6380), 1094-1096.

Li, X., Zhang, J., & Zong, C. (2016). One sentence one model for neural machine translation. *arXiv preprint* arXiv:1609.06490.

Lin, C. Y. (2004, July). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out* (pp. 74-81), Barcelona, Spain.

Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, US: MIT press.

Oshikawa, R., Qian, J., & Wang, W. Y. (2018). A survey on natural language processing for fake news detection. *arXiv preprint* arXiv:1811.00770.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data, 6*(60), 1-48.

Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text data augmentation for deep learning. *Journal of Big Data, 8*(101), 1-34.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter, 19*(1), 22-36.

Sun, Y., Wong, A. K., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence, 23*(04), 687-719.

Van der Kaa, H., & Krahmer, E. (2014, October). Journalist versus news consumer: The perceived credibility of machine written news. In *Proceedings of the Computation+ Journalism Conference, 24*, Columbia University, New York (p. 25).

Van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple

introduction to Markov Chain Monte-Carlo sampling. *Psychonomic Bulletin & Review, 25*(1), 143-154.

Wu, G., & Chang, E. Y. (2003, August). Class-boundary alignment for imbalanced dataset learning. In *ICML 2003 Workshop on Learning from Imbalanced Data Sets II,* Washington, DC (pp. 49-56).

Yadav, A. K., & Borgohain, S. K. (2014, May). Sentence generation from a bag of words using N-gram model. In *2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies*, IEEE (pp. 1771-1776).

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Computational Intelligene Magazine, 13*(3), 55-75.

투 고 일 자: 2023년 01월 12일
심 사 일 자: 2023년 02월 04일
게재확정일자: 2023년 03월 09일

Abstract

A Study on the Discovery of General News Reporting Pattern

A Small Sample LSTM Experiment with Natural Language Generation

Ho Young Yoon
Assistant Professor, Division of Communication & Media, Ewha Womans University
Dohyun Ahn
Associate Professor, Dept. of Journalism & Mass Communication, Jeju National University

This paper proposes using natural language generation model with LSTM neural network for clustering news reporting pattern. More specifically, it suggests to collect news articles and to utilize a sentence-based natural language generation that can infer the general patterns of news reporting and then to compare the similarity of content features between the generated sentences and the collected data sentences. Levenstein distance and ROUGE-L metric are used for the comparison. These two metrics are to measure the content and expression similarity etween the computer-generated sentences and the actual article sentences. In doing so, we propose a rapid news clustering method that can be used for supervised learning in the later analysis stage. In this article, we demonstrate the application of these methods using small-scale data on infectious disease vaccine coverage and discuss the potential benefits of this method.

KEYWORDS Natural Language Generation, LSTM, Levenstein Distance, ROGUE, News Clustering