딥러닝 및 토픽모델링 기법을 활용한 소셜 미디어의 자살 경향 문헌 판별 및 분석*

Examining Suicide Tendency Social Media Texts by Deep Learning and Topic Modeling Techniques

고 영 수 (Young Soo Ko)**
이 주 희 (Ju Hee Lee)***
송 민 (Min Song)****

초 록

자살은 전 세계 사망 원인 중 4위이며 사회, 경제적 손실이 큰 난제이다. 본 연구는 자살 예방을 위하여 소셜미디어에 나타난 자살 관련 말뭉치를 구축하고 이를 통해 자살 경향 문헌을 분류할 수 있는 딥러닝 자동분류 모델을 만들고자 하였다. 또한, 자살 요인을 분석하기 위해 주제를 자동으로 추출하는 분석 기법인 토픽모델링을 활용하여 자살 관련 말뭉치를 세부 주제로 분류하고자 하였다. 이를 위해 소셜미디어 중 하나인 네이버 지식iN에 나타난 자살 관련 문헌 2,011개를 수집한 후 자살예방교육 매뉴얼을 기준으로 자살 경향 문헌 및 비경향 문헌 여부를 주석 처리하였으며, 이 데이터를 딥러닝 모델(LSTM, BERT, ELECTRA)로 학습시켜 자동분류 모델을 만들었다. 또한, 토픽모델링 기법의 하나인 LDA 기법으로 주제별 문헌을 분류하여 자살 요인을 발견하였고 이를 심층적으로 분석하기 위해 주제별로 동시출현 단어 분석 및 네트워크 시각화를 진행하였다.

ABSTRACT

This study aims to create a deep learning-based classification model to classify suicide tendency by suicide corpus constructed for the present study. Also, to analyze suicide factors, the study classified suicide tendency corpus into detailed topics by using topic modeling, an analysis technique that automatically extracts topics. For this purpose, 2,011 documents of the suicide-related corpus collected from social media naver knowledge iN were directly annotated into suicide-tendency documents or non-suicide-tendency documents based on suicide prevention education manual issued by the Central Suicide Prevention Center, and we also conducted the deep learning model(LSTM, BERT, ELECTRA) performance evaluation based on the classification model, using annotated corpus data. In addition, one of the topic modeling techniques, LDA identified suicide factors by classifying thematic literature, and co-word analysis and visualization were conducted to analyze the factors in-depth.

키워드: 자살, 소셜미디어, 단어 동시출현, 딥러닝, 토픽모델링
Suicide, Social media, Word Co-Occurrence, Deep-learning, Topic Modeling

* 본 연구는 정부의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2018S1A3A2075114).
** 연세대학교 문헌정보학과 석사과정(kosue@yonsei.ac.kr) (제1저자)
*** 연세대학교 문헌정보학과 석사과정(juhee5795@yonsei.ac.kr) (공동저자)
**** 연세대학교 문헌정보학과 교수(min.song@yonsei.ac.kr) (교신저자)
논문접수일자 : 2021년 8월 16일 논문심사일자 : 2021년 8월 17일 게재확정일자 : 2021년 9월 6일
한국비블리아학회지, 32(3): 247-264, 2021. http://dx.doi.org/10.14699/kbiblia.2021.32.3.247

1. 서 론

세계보건기구에 따르면 전 세계적으로 한 해에 70만명 이상이 자살로 인해 사망하고 있다. 이는 40초마다 한 명씩 자살하는 것을 의미하며 자살은 전 세계 사망 원인의 4위를 차지한다(World Health Organization, 2021). 대한민국의 10만 명당 자살 사망률은 OECD 평균인 11.3명보다 2배 이상 높은 26.9명으로 2003년부터 2020년까지 2017년을 제외하고 OECD 국가 중 1위를 기록하였다. 1년에 1만3,799명이 자살을 하고 있으며(통계청, 2020) 이는 큰 사회적 비용으로 이어지는데 자살과 우울증으로 매년 발생하는 사회, 경제적 손실이 총 6조 원 이상일 것으로 추정된다(정상혁, 2005).

2019년 코로나바이러스 질병(COVID-19) 대유행으로 인해 자살에 대한 문제는 더 심각해지고 있다. 미국 질병통제예방센터(CDC)에 따르면 코로나 19 팬데믹 기간 동안 미국 10대 소녀들의 자살 시도가 51% 증가했다(News18, 2021). 한국에서는 보건복지부와 한국트라우마스트레스학회에서 실시한 '2021년 1분기 코로나19 국민정신건강 실태조사'에 따르면 자살을 생각하는 비율이 2018년 4.7%에서 2020년 5월 10.1% 2021년 3월 16.3%로 많이 증가하는 것으로 나타나 코로나19 장기화로 인하여 정신건강이 악화되고 자살 시도가 많이 증가하고 있음을 알 수 있다(보건복지부, 2021).

자살은 죽음을 스스로 원해서 치명적인 행동을 통해 사망에 이르는 것이다(Kaplan et al., 2007). 세계보건기구에 따르면 최소 20번의 자살 시도가 있으면 한 번의 자살이 발생하게 되고(World Health Organization, 2021) 자살 생각과 계획 후에 자살 시도가 이루어지기 때문에 자살 시도는 자살의 주요 위험요소이다(NIMH, 2021). 따라서 자살 생각을 하는 사람을 조기에 발견하여 자살 시도를 줄이는 것이 매우 중요하며 이를 통해 자살을 예방할 수 있다. 자살 시도를 막는 가장 큰 방법은 대화로 알려져 자살 징후를 보이는 사람들을 파악하고 적절한 대화를 취하는 것이 중요하지만 사람들의 65.8%는 자살에 관한 이야기를 꺼리는 경향을 보여 대응이 쉽지는 않다(안용민, 2019).

소셜미디어가 발달하면서 익명성을 기반으로 누구나 SNS상에 글을 남길 수 있고 자살을 생각하는 사람들도 자기 생각을 적극적으로 표현하고 있다(이범오, 2020). 이러한 이유로 소셜미디어를 활용한 정신건강 관련 연구가 늘어나고 있다. 특히, 소셜미디어 중 하나인 '네이버 지식iN'은 이용자가 익명으로 질문을 남기고 누구든지 그 질문에 응답해줄 수 있는 시스템으로 이용자들이 솔직하고 자세하게 내용을 서술하는 경향이 있어 이를 활용한 자살 연구에 적합하다(이수빈 외, 2021). 이에 본 연구는 '네이버 지식 iN'을 이용하여 자살 관련 징후를 보이는 말뭉치를 구축하고, 이를 딥러닝 모델에 학습시켜 자동분류 모델을 구축함으로써 자살 징후를 보이는 사람을 빠르게 파악하는 것에 도움을 주고자 한다.

또한, 자살을 근본적으로 해결하기 위해서는 자살의 주요 요인을 파악하는 것이 필요하다. 보건복지부가 진행한 2018 자살실태조사에 따르면 자살 시도 원인은 우울, 불안 등의 정신과적 원인이 35.1%로 가장 높게 나타났고, 두 번째로 대인관계 문제가 30.3%로 높게 나타났다. 대인관계 문제가 자살 시도의 원인으로 높은 비

율을 차지하는 것은 한국 사회의 큰 특징이다(안용민, 2019). 대인관계 문제의 구체적인 대상으로는 가족 47.3%, 연인/배우자 42.2%, 친구 6.2% 순서로 나타났다. 소셜미디어에서는 더욱 솔직하고 적극적인 표현이 가능하기에 소셜미디어의 글에서 자살 요인을 파악하는 것은 자살을 생각하는 사람들의 속마음을 깊이 아는 방법이다. 따라서, 본 연구는 소셜미디어에 나타난 자살 관련 문헌을 분석하여 자살 요인을 파악하고자 한다.

연구 질문은 다음과 같다.

1) 자살 경향 말뭉치에 기반한 자살 경향 문헌 판별 딥러닝 모델 중 가장 적합한 모델은 무엇인가?
2) 자살 경향 말뭉치를 통해 파악된 자살의 주요 요인과 관련 어휘들은 무엇인가?

2. 선행연구

2.1 소셜미디어를 활용한 자살 분석

소셜미디어는 익명성과 비대면성을 기반으로 많은 사용자를 확보하고 있다. 질환을 앓는 사람들 역시 소셜미디어 공간에서 자유롭게 질환이나 증상을 이야기하는 경향을 보인다. 이를 통해 잠재적인 환자들의 조기 파악 및 진단도 가능하며, 자세한 증상도 확인할 수 있다.

관련하여 소셜미디어를 이용한 국가 자살자 수를 예측하는 연구 및 자살 암시 징후를 포착하거나 자살 관련 트윗을 판별하는 연구들이 진행되었다. 한국 소셜데이터를 활용하여 자살자 수를 예측하는 프로그램을 개발하여 79%의 정확성을 보인 연구(Won et al., 2013), 2011년부터 시행된 미국 뒤르켐 프로젝트를 통해 소셜미디어에서 드러나는 언어적 자살 암시 징후를 포착하여 자살을 예방하고자 한 연구(Thompson, Bryan, & Poulin, 2014), 소셜미디어 중 하나인 트위터에서 자살 관련 트윗 데이터 세트를 수집한 후 CNN 알고리즘을 통해 자살 관련 트윗을 판별하여 78%의 정확도를 보인 연구(Du et al., 2018)가 그 예시이다.

본 연구는 기존 연구들의 자살 문헌 판별 정확도를 더 높이려는 방법으로 3가지 계열(LSTM, BERT, ELECTRA)의 5가지 딥러닝 모델을 활용하여 실험을 진행하여 가장 적합한 모델을 찾고자 하였고 이를 활용하여 더 높은 정확도의 자살 경향 문헌 판별 모델을 구축하고자 하였다.

2.2 자연어 딥러닝 연구

자연어처리는 자연어의 의미를 분석해서 컴퓨터가 처리하도록 하는 일을 뜻하는데, 이 분야에서 딥러닝 모델은 많은 발전을 이루어내고 있다. 딥러닝을 통해 단어 혹은 구가 무슨 의미를 지니는지 학습하고, 이해하는 것이다. 딥러닝 모델 성능 개선 시 적합한 고품질의 데이터를 사용하여 말뭉치를 구축하는 것이 공통된 중요사항이다(박찬준 외, 2021). 최근 다양한 딥러닝 모델을 활용한 연구들이 많이 진행되었다.

2.2.1 LSTM 모델

LSTM 모델을 활용하여 진행한 연구로는 구축한 전통문화 말뭉치를 기반으로 학습데이터로

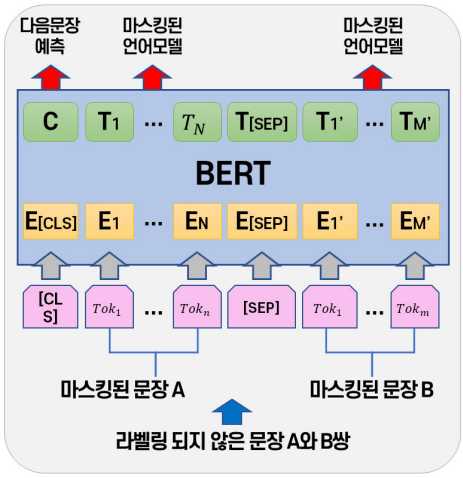
활용하여 다중작업학습 기법을 적용하고 개체명 인식 모델에 대해 성능 비교 분석을 진행한 연구(김경민 외, 2018)가 있다.

LSTM(Long-Short Term Memory)은 BERT 이전부터 가장 많이 쓰인 딥러닝 모델 중 하나로, RNN(Recurrent Neural Networks)에 기반하되 메모리 셀의 은닉층을 직접 연결하여 RNN의 기울기 소실 문제를 해결한 모델이다. 즉, 타임 스텝이 길어져도 이전 정보를 보존할 수 있도록 한 것이다. 김경민 외(2018)는 이 연구에서 LSTM의 양방향 은닉층을 결합하여 더 많은 문맥 정보를 얻는 BiLSTM(Bidirectional LSTM)에 CRF, CNN을 결합하여 사용한 실험이 가장 좋은 성능을 보임을 확인하였다.

2.2.2 BERT 모델

BERT 활용한 자연어 딥러닝 연구로는 한국어 학습자 말뭉치를 딥러닝 언어모델인 BERT를 활용하여 분석한 연구(이진, 정진경, 김한샘, 2021), 공황장애 관련 코퍼스를 구축하고 딥러닝 모델학습을 통해 자동분류를 시도한 연구(이수빈 외, 2021) 등이 있다.

BERT(Bidirectional Encoder Representation from Transformers)는 트랜스포머(Vaswani et al., 2017)의 self-attention을 인코더에 적용하여 양방향 심층 학습이 가능하도록 설계한 언어모델이다(Devlin et al., 2018). BERT는 〈그림 1〉과 같이 다음 문장 예측(Next Sentence Prediction)과 문장에서 가려진 단어 예측(Masked Language Model)을 통해 학습한다. 다음 문장 예측 학습기법은 추출되는 문장 쌍이 논리적으로 연결되는 관계에 놓인 것인지 판단하여 가중치를 업데이트하고, 가려진 단어 예측은 주어진 문장의 일부 단어를 마스크 토큰으로 활용하여 올바르게 예측하도록 가중치를 업데이트한다. 이를 통해 양방향 예측이 가능하고, 더 긴 의존관계를 포착할 수 있도록 하는 것이다.



〈그림 1〉 BERT 모델

BERT는 사전학습에서 양질의 대규모 텍스트가 필수적이기에, 최근 다양한 데이터로 사전학습을 시키는 것이 중요한 연구영역으로 주목된다. 한글의 경우 SKTBrain에서 공개한 KoBERT는 한글 위키피디아를 기반으로 학습하였고(Jeon, 2021), 구글에서 발표한 multilingual BERT(Devlin, 2021)는 다국어 버전으로 104개 언어의 위키피디아 코퍼스를 모두 사용하여 사전 학습하였기에, 다양한 언어에서 좋은 성능을 보였다.

이러한 이점에 기댄 자연어처리 연구들은 좋은 성능을 보였다. 이진, 정진경, 김한샘(2021)은 BERT계열인 KoBERT를 활용한 한국어 문장 분류 정확도를 91%로 높게 확인하였다. 이수빈 외(2021)는 KoBERT, BERT-multilingual, KcBERT를 비교하였고 그 중 KcBERT의 정확

도 및 정확률과 재현율이 가장 높은 성능을 보였음을 확인하였다.

2.2.3 ELECTRA 모델

ELECTRA는 BERT와 같은 MLM (Masked Language Modeling) 방식이 사전학습 시 효율성을 위해 많은 양의 컴퓨팅이 필요하다는 단점을 보완하기 위해 등장하였으며 교체된 토큰 감지라는 효율적인 방식으로 사전훈련을 진행한다. 〈그림 2〉와 같이 일부 토큰을 small generator 네트워크에 대체하여 입력을 손상한 후 손상된 토큰의 원래 ID를 예측하는 모델 학습방법을 사용하며 BERT, RoBERTa 등과 비교했을 때 더 좋은 성능을 보였다(Clark, K et al., 2020). KoELECTRA는 이 방식을 활용하여 34GB의 한국어 텍스트(뉴스, 나무위키, 신문, 문어, 구어, 메신저, 웹 등)를 학습 시킨 모델이며 한국어 말뭉치에 활용하여 좋은 성과를 보였다(Park, 2020). ELECTRA를 활용한 자연어 분석 관련 연구는 아직 많이 진행되지 않았다.

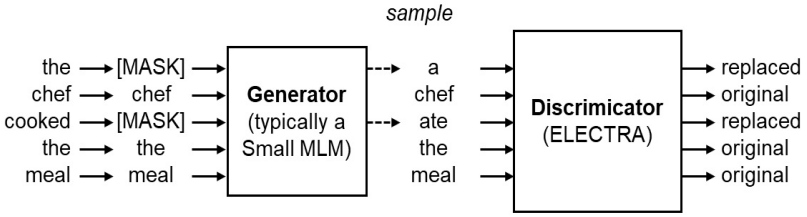
이에 본 연구는 자연어처리 분야에서 두각을 나타낸 딥러닝 모델인 LSTM, BERT와 최근 등장한 딥러닝 모델인 ELECTRA에 자살 경향 말뭉치를 학습 시켜 딥러닝 모델의 성능을 비교하는 실험을 통하여 가장 적합한 딥러닝 모델을 확인 및 활용하고자 하였다.

2.3 토픽모델링 및 동시출현 분석

자살의 주요 요인을 살펴보기 위해서각 문헌의 주제를 발견하는 기법인 토픽모델링을 사용하였다. 토픽모델링은 구조화되지 않은 문헌 집단 내, 같은 맥락에서 나타날 가능성이 있는 단어들을 그룹화하여 주제를 추론하는 방법이다(Steyvers & Griffiths, 2007). 이는 대량의 데이터를 분석하고 소셜네트워크 등 다양한 유형의 데이터에 적용해 일정한 패턴을 찾을 때 유용하게 쓰일 수 있다. LDA(Latent Dirichlet al.location)는 토픽모델링 기법 중 가장 많이 활용되는 문헌 생성 모델로, 문헌 단위에서 각 주제의 분포로 문헌을 표현하며, 이러한 각 주제로 용어분포의 추정 확률을 표현할 수 있다(송민, 2017).

정신질환 분석에 토픽모델링을 활용한 연구들로는 조현병에 대한 사회적 인식변화를 LDA 토픽모델링으로 살펴본 연구(김현지 외, 2019), 우울증 환자의 언어를 분석하고 좋은 결과를 제공하기 위해 토픽모델링을 활용한 연구(Resnik et al., 2015), LDA 토픽모델링으로 우울증에 대한 텍스트를 분석하여 심리적으로 관련된 주제를 산출한 연구(Resnik, Garron, & Resnik, 2013) 등이 있다.

살펴본 토픽별 요인의 깊이 있는 파악을 위해



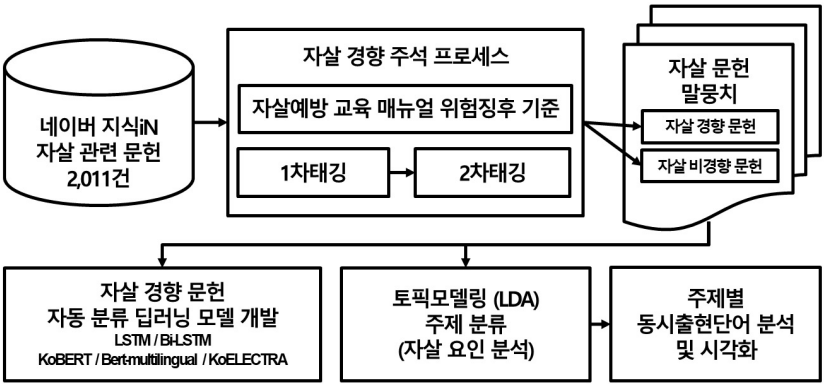
〈그림 2〉 ELECTRA 모델

서 동시출현 단어 분석을 진행하였다. 동시출현 단어 분석은 여러 단어가 주어진 문헌 혹은 텍스트 내에서 단어의 쌍이 함께 출현하는 횟수를 기반으로 단어의 연관성과 문헌 집단의 특성을 파악하는 방법이다. 토픽별 동시출현하는 단어의 횟수를 구하면, 한 토픽에서 어느 단어들이 서로 밀집하게 연결되어 있는지 더 깊이 해석할 수 있다. 서하림, 송민(2019)은 소셜미디어에서 토픽모델링 및 동시출현 단어 분석으로 우울 경향 이용자들의 주제적, 어휘적 사용의 특성을 살펴보았다. 김나리, 이남주(2021) 역시 토픽모델링과 동시출현 단어 분석을 통해 환자안전 관련 사회적 이슈가 어떻게 변화하였는지를 살펴보았다.

위 선행연구들을 통해 토픽모델링 및 동시출현 단어 분석을 통해 더 깊은 의미를 찾아볼 수 있음을 살펴보았다. 본 연구는 정신질환 분석에 활용된 토픽모델링 기법을 사용하여 자살 경향 문헌의 주요 요인을 보다 심층적으로 분석하고자 하였다. 또한, 토픽모델링에 동시출현 분석을 적용하여 토픽별 밀접하게 관련된 단어들이 무엇인지를 살펴보고자 하였다. 기존의 대다수 자살 관련 연구가 자살 경향 문헌 판별과 예측에 주를 두었다면 본 연구를 통해 자살의 주요 요인까지 살펴볼 수 있다.

3. 연구 방법

본 연구의 진행 절차는 〈그림 3〉과 같다. 먼저 네이버 지식iN에서 자살을 언급하고 있는 글을 자살 관련 문헌으로 수집하고 이를 자살예방 교육 매뉴얼에 나타난 위험 징후 기준으로 2차에 걸쳐 수기 분류하여 자살 관련 문헌과 자살 비관련 문헌으로 나누었다. 이렇게 구축된 자살 문헌 말뭉치를 5가지의 딥러닝 모델(LSTM, Bi-LSTM, KoBERT, Bert-multilingual, KoELECTRA)로 학습하여 자살 경향 문헌 자동분류 모델을 개발하고 성능을 비교하였다. 다음으로 자살 문헌 말뭉치 중 자살 경향 문헌을 토픽모델링 LDA 기법을 활용하여 주제별로 분류하였고 각 주제의 문헌들에서 동시출현한 단어들을 분석하고 이를 네트워크로 시각화하여 자살 요인을 심층적으로 파악하고자 하였다.



〈그림 3〉 연구 모형

3.1 데이터 수집

본 연구는 소셜데이터에 나타난 자살 관련 문헌을 찾기 위해 소셜미디어 중 하나인 네이버 지식iN에서 데이터를 수집하였다. 이용자가 자신의 심리적인 상태와 이유, 도움이 필요한지, 왜 이 글을 쓰는지 등 구체적인 상황을 지식iN에서 질문의 형태로 남기고 있어 이를 통한 분석 및 활용이 쉽다. 또한, 소셜미디어의 특성상 자신의 정서와 감정을 가감 없이 드러낸다는 점에서 지인과의 대화 혹은 오프라인에서의 대화보다 언어적 자살 위험 징후 감지가 더 쉽다. 네이버에서 "자살" 단어의 검색을 허용되지 않는 점을 감안하여 "죽고"와 "ス ト 살" 단어를 검색하였고 2018년 1월부터 2020년 12월까지의 데이터를 수집하였다. 파이썬 프로그램을 활용한 데이터 수집 코드 실행결과 총 2,011건의 데이터를 수집하였다.

3.2 데이터 주석

수집된 데이터를 중앙자살예방센터에서 발간한 자살예방교육 매뉴얼의 자살 위험 징후를 기반하여 자살 경향 문헌과 자살 비경향 문헌으로 분류하였다(중앙자살예방센터, 2012). 6개의 자살 위험 징후 중 1개 이상 발견된 데이터를 자살 경향 문헌으로 간주하여 분류하였고 추가 분석을 위해 이용자가 연령대를 밝혔을 때 연령을 데이터에 추가로 주석 처리하였다. 주석의 신뢰도를 높이기 위해 2명의 연구원이 2회에 걸쳐 중복으로 검토하였다. 그 결과 1,496건, 약 74%를 자살 경향 문헌으로 분류하였다. 〈표 1〉은 분류 시 사용한 자살 위험 징후이며 〈표 2〉는 자살 경향 및 비경향 문헌 분류 예시이다.

〈표 1〉 자살예방교육 매뉴얼 내 자살 위험 징후

자살 위험 징후	
1	자살이나 살인, 죽음에 대한 말을 자주 한다.
2	자기 비하적인 말을 한다.
3	사후세계를 동경하는 말을 한다.
4	신체적 불편함을 호소한다.
5	자살하는 방법에 대해 질문한다.
6	자살한 사람들에 관한 이야기를 꺼낸다.

3.3 자살 경향 문헌 자동분류 모델

자살 예방에 도움을 주려는 연구의 목적에 따라 자살 경향 문헌을 자동으로 판별할 수 있는 딥러닝 자동분류 모델 개발을 진행하였다. 앞서 구축한 말뭉치를 딥러닝 모델학습을 위한 훈련데이터로 사용하였고 가장 적합한 모델을 찾기 위해 LSTM

〈표 2〉 자살 경향 및 비경향 분류 예시

문헌	위험 징후	문헌 분류
저좀도와주세요죽고싶습니다 안녕하세요 24살이구 미필이구요 진짜하루하루죽고싶습니다 부모님한테도손벌릴수도없구 현재채무카카오뱅크 2080만원햇살론1070햇살론17 700수협새희망홀씨2 1800주택청약통장담보1100가량연봉은3700정두되구요 ... 등급은한7등급나오고요	1번	자살 경향 문헌
죽고 싶다는 가사가 있는 노래와 위로해주는 가사가있는 노래좀 알려주세요!!	-	자살 비경향 문헌

기반 2개의 모델(LSTM, Bi-LSTM)과 BERT 기반 2개의 모델(KoBERT, Bert-multilingual) 및 ELECTRA 기반의 모델(KoELECTRA)로 파인튜닝 하였고 평가 결과를 도출하였다. 최종 학습된 모델은 특정 문헌을 입력할 경우 자살 경향을 보이는지 판단하는 분류 모델(classification model)로 활용할 수 있다.

모델 실험을 위해 전체 데이터를 학습데이터 80%와 테스트데이터 20%로 분류하였다. 학습데이터를 활용하여 모델을 학습시켰으며 테스트데이터는 학습된 모델의 성능 평가에 사용하였다. 하이퍼 파라미터는 BERT의 원논문(Devlin et al., 2018)에서 사용한 기준을 그대로 사용하였다. 성능 평가를 위한 기준으로는 정확도(accuracy), 정확률(precision), 재현율(recall), F-1 score 지표를 사용하였다. 정확도는 모델을 통해 예측된 데이터 중 정답이 얼마나 되는지를 확인하는 지표이다. 정확률은 모델이 정답이라고 분류한 것 중 실제 정답인 것의 비율이며 재현율은 실제 정답인 것 중에서 모델이 정답으로 예측한 것의 비율을 의미한다. 데이터가 불균형 구조일 때 모델의 성능을 정확하게 평가하기 어려운데 이를 보완하기 위해 정확률과 재현율의 조화 평균인 F-1 score을 사용한다. 본 연구에서는 LSTM, BERT, ELECTRA의 5가지 모델의 정확도, 정확률, 재현율, F-1 score를 비교하여 가장 성능이 높은 모델을 찾고자 하였다.

3.4 토픽모델링 자살 요인 분석

연구 질문 2의 자살의 주요 요인과 관련 어휘를 심층적으로 분석하기 위해 토픽모델링 기법의 하나인 LDA 기법을 활용하여 자살 경향 문헌을 주제별로 분류하였다. 이때 사용한 분석 도구는 텍스트 마이닝 통합모듈인 treform(Song, 2021)이다. 자살경향 문헌으로 분류된 1,496개의 문헌 중 전처리를 통해 의미를 찾지 못한 단어를 제외한 1,433개의 문헌이 토픽모델링에 사용되었다. 적정 토픽의 개수를 정하기 위해 2-20까지의 토픽 수로 실험을 하였고 그 결과 perplexity 값이 가장 낮은 9로 정하여 진행하였으며 각 토픽의 대표 단어 수는 10개로 설정하였다. 이후 동시출현 단어 분석 시 주제가 같은 경우에는 묶어서 함께 분석하였다.

3.5 토픽별 동시출현 단어 분석 및 시각화

토픽모델링의 결과는 대표적인 단어로만 나타나 심층적인 내용을 알기 어렵다. 더 자세한 자살 요인을 알기 위해서는 같은 토픽으로 분류된 문헌이 어떤 단어들로 이루어졌는지와 동시에 출현하는 단어는 무엇인지에 대한 분석이 필요하다. 따라서 주제별로 분류된 문헌을 각각 동시출현 단어 분석 및 네트워크 시각화를 진행하였다. 동시출현 분석 역시 텍스트 마이닝 통합모듈인 treform(Song, 2021)을 사용하였고 단어 두 쌍의 출현 빈도를 구한 뒤 graphml 파일로 만들었으며 이를 시각화 도구인 Gephi를 이용하여 네트워크로 표현하였다.

4. 연구 결과

4.1 기초데이터 (연령대) 분석

자살 경향 문헌의 특성을 파악하기 위해 연령

대가 문헌에 표현된 경우 주석처리 시 연령대를 표기하였다. 각 연령별로 분석한 결과는 〈표 3〉과 같다. 자살 경향 문헌 1496건 중 연령대를 표현한 문헌은 386건으로 전체의 25.8%를 차지한다. 이 중 10대의 문헌 수는 323건으로 83.7%를 차지한다. 소셜미디어에 자살 관련 글을 남길 때 자신의 연령대를 밝히는 인원 대부분이 10대임을 확인할 수 있다. 향후 연령대를 표기한 문헌을 추가로 수집 및 분석하면 연령대별 자살 관련 연구로 소셜미디어 말뭉치를 활용할 수 있음을 확인할 수 있다.

〈표 3〉 연령대 분석 결과

연령대	문헌 수(개)	비율(%)
전체	386	100
10대	**323**	**83.7**
20대	52	13.5
30대	7	1.8
40대	3	0.8
50대	1	0.3

4.2 자살 경향 문헌 자동분류 딥러닝 모델

자살 경향 문헌을 자동으로 분류할 수 있는 딥러닝 모델 성능 결과는 〈표 4〉와 같다. LSTM 계열의 2개 모델과 BERT 계열의 2개 모델, ELECTRA 계열의 모델을 비교했을 때 ELECTRA 계열의 모델의 성능이 가장 높게 나온 것을 확인할 수 있다. KoELECTRA의 성능이 가장 높은 이유는 효율적인 사전학습 방법을 활용하여 뉴스, 나무위키, 신문, 문어 등의 대량의 말뭉치를 사전 학습한 모델이기 때문으로 확인된다. KoELECTRA로 학습된 모델의 성능은 정확도 93.3%, 정확률 91.96%, 재현율 96.36%, f-1 score 94.11%이다.

이를 통해 연구 질문 1의 가장 적합한 자살 경향 문헌 자동 판별 모델은 KoELECTRA임을 알 수 있다.

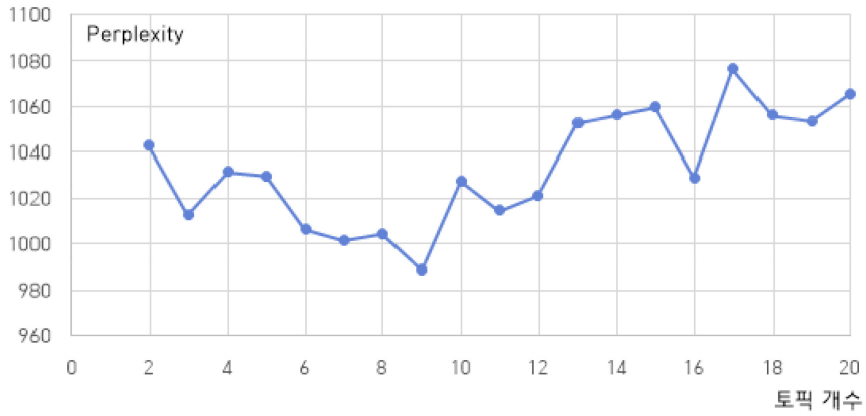
4.3 토픽모델링 결과

토픽모델링을 진행하기 위한 적절한 토픽 개수를 찾기 위해 토픽을 2개에서 20개까지 정했을 때의 토픽별 perplexity의 값을 구하였고 그 결과는 〈그림 4〉와 같다. 토픽 개수를 9로 할 경우의 perplexity 값이 가장 낮아 적합한 토픽 개수를 9로 정하고 실험을 진행하였다.

토픽모델링 결과는 〈표 5〉와 같다. 9개의 토픽 중 가장 많은 비율인 34%를 차지하는 문헌 주제는 "복합적인 이유"로 나타났다. 하나의 주제가 아닌 다양한 주제가 한 문헌에 표현되고

〈표 4〉 자살 경향 문헌 자동분류 딥러닝 모델 성능

딥러닝모델	정확도(%) accuracy	정확률(%) precision	재현율(%) recall	f-1 score(%)
KoBERT	92.83	90.86	95.21	92.98
Bert-multilingual	90.50	88.71	93.30	90.94
KoELECTRA	**93.33**	**91.96**	**96.36**	**94.11**
LSTM	88.38	86.26	88.35	87.29
Bi-LSTM	87.07	83.33	87.38	85.31



〈그림 4〉 토픽모델링 토픽 개수별 Perplexity 결과

〈표 5〉 토픽모델링 결과

Topic 번호	문헌 수(개)	문헌비율(%)	주제	대표단어
7	489	34	복합	가족, 친구, 엄마, 부모, 공부, 마음, 행복, 인생, 방법, 삶, 자살, 위로, 이유, 세상
4	402	28	정신질환	우울, 우울증, 정신, 스트레스, 상담, 병원, 요즘, 애기, 기분, 하루
8	130	9	가족	아빠, 동생, 언니, 오빠, 방, 폰, 할머니, 학년, 애기, 술
1	125	9	학업	학원, 시험, 선생, 성적, 스트레스, 학년, 수학, 키, 중학교, 이번
6	88	6	복합	욕, 글, 아이, 눈물, 미안, 얼굴, 마음, 감사, 오늘, 잘못
5	66	5	경제	빚, 대출, 회사, 도박, 직장, 카드, 연체, 재산, 상태, 취업
2	57	4	신체질환	병원, 여드름, 얼굴, 수술, 피부, 효과, 처방, 치료, 코, 가능
3	47	3	가족	아버지, 남자, 어머니, 사랑, 연락, 누나, 미안, 남친, 결혼, 글
0	29	2	진로	대학, 운동, 수능, 대학교, 말씀, 아침, 자퇴, 음식, 하루, 고등학교

있어서 토픽모델링으로 정확하게 분류하기에는 한계가 있음을 확인하였다. 대표 단어로는 '가족', '친구', '엄마' 등이며 대표적인 문헌은 다음과 같다.

"전 평범한 초등학생 소녀입니다 지금까지 열심히 살아왔는데 다 포기하고 죽고싶어요. 이런 학업문제, 인간관계 모조리 힘들어요. 잘려고 보면 12시 25분전입니다...공부가 자는시간 빼고 거의다 차지해요 근데..친구는 무슨 상관인지 지가 선생님인척 공부하라고 하고 부모님도 공부을 하고있는 와중에도 성공, 공부 관한 이야기을 합니다.어쩌면 이세상에 살이유도 태어날이유도 없는것 같아요 힘들어요 심지어 학교에선 은따입니다..."

두 번째로 많이 나타난 토픽의 주제는 문헌의 28%을 차지한 "정신질환"이다. 대표적인 단어로 '우울', '우울증', '정신' 등이 나타났다. 세 번째로 많은 주제의 비율은 문헌의 9%로 "가족"에

대한 내용이 나타났다. 가족에 관한 내용은 8번째 토픽으로도 나타나 향후 동시출현 분석 및 네트워크 시각화를 할 때 두 토픽의 문헌을 합쳐서 분석하였다. 네 번째로는 9%로 "학업"에 대한 내용이 나타났으며 대표적인 단어로는 '학원', '시험', '선생' 등이다. 소셜미디어에서 이 글을 남기는 연령대가 10대가 많기 때문에 학업에 대한 스트레스와 부담이 많이 등장하는 것으로 확인된다. 다섯 번째로 많이 나타난 토픽은 "복합적인 이유"로 확인된다. 문헌들을 자세히 살펴보면 가족, 친구들에게 상처받고 죽고 싶다는 내용이 많이 등장한다. 대표적인 단어로는 '욕', '글', '아이' 등이다. 여섯 번째로 많이 나타난 토픽은 "경제"이다. 주요 단어는 '빚', '대출', '회사'이며 이 문헌들에서는 경제적인 어려움을 호소하며 죽고 싶다는 내용이 많이 등장한다. 7번째 토픽은 "신체질환"이다. '병원', '여드름', '얼굴' 등이 대표 단어로 나타나며 두 번째로 많이 등장한 "정신질환"과는 다르게 신체적인 고통을 호소하며 죽고 싶다는 내용이 많이 등장한다. 9번째 토픽은 "진로" 이다. '대학', '운동', '수능'이 대표 단어로 확인되며 "학업" 주제와 달리 수능과 대학 입시로 인한 스트레스를 호소하는 내용이 많이 등장한다. 요인을 세부적으로 분석하기 위해서 이 중 문헌의 수가 100개를 넘어가는 상위 4개("복합", "정신질환", "가족", "학업")토픽의 문헌들을 선택하여 동시출현 단어 분석 및 네트워크 시각화를 진행하였다.

4.4 토픽별 동시출현 단어 분석 및 시각화

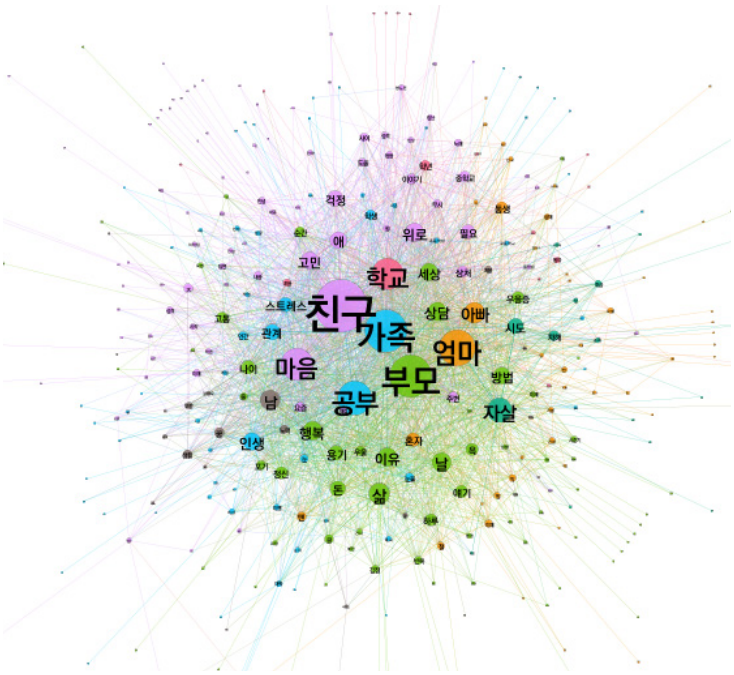
토픽별 문헌 내 동시출현 단어 분석 결과는 〈표 6〉과 같다. "복합" 주제에서는 '친구-가족' 단어 쌍이 47회로 동시출현한 단어 중 가장 많이 등장하였고 "정신질환" 주제에서는 '친구-학교' 단어 쌍이 39회, "가족" 주제에서는 '엄마-아빠'

〈표 6〉 토픽별 문헌 내 동시출현 단어

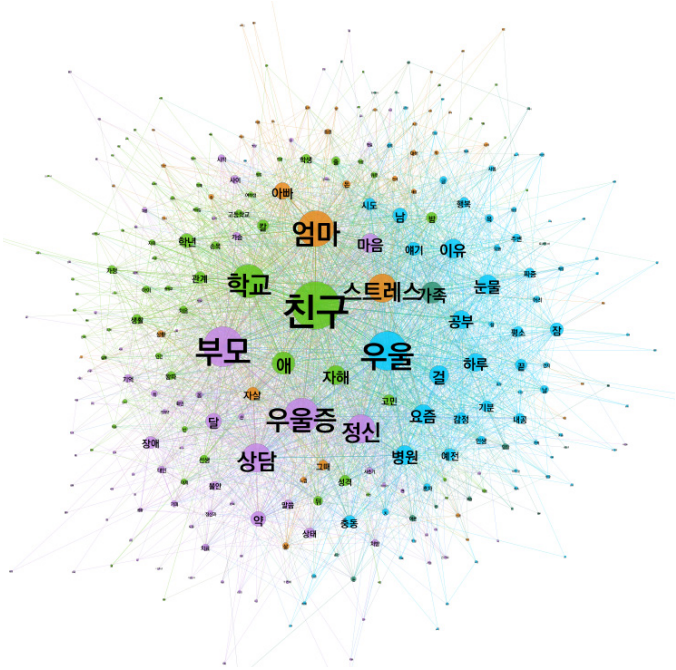
주제	"복합"			"정신질환"			"가족"			"학업"		
순위	단어1	단어2	동시출현 수(개)	단어1	단어2	동시출현 수(개)	단어1	단어2	동시출현 수(개)	단어1	단어2	동시출현 수(개)
1	친구	가족	47	친구	학교	39	엄마	아빠	69	시험	공부	33
2	부모	친구	43	우울	친구	34	엄마	동생	44	엄마	학원	21
3	엄마	아빠	36	엄마	아빠	33	엄마	친구	33	공부	학원	21
4	친구	학교	31	부모	친구	31	아빠	동생	27	엄마	공부	20
5	엄마	친구	27	우울	우울증	31	엄마	언니	25	엄마	시험	17
6	공부	친구	27	친구	엄마	31	엄마	부모	25	학원	수학	16
7	친구	마음	26	애	친구	26	엄마	욕	25	성적	공부	16
8	애	친구	25	관계	친구	26	엄마	학교	24	성적	시험	15
9	부모	가족	23	우울증	친구	25	아빠	학교	23	공부	스트레스	15
10	자살	시도	23	스트레스	친구	25	엄마	방	21	시험	학원	15

단어 쌍이 69회, “학업” 주제에서는 ‘시험-공부’ 단어 쌍이 33회로 가장 많이 등장하였다. 특이한 점은 문헌 집단별로 주제는 각각 다르나 상위에 동시출현한 단어에 친구, 부모, 엄마, 아빠 등 인간관계와 연결된 단어가 많이 등장한다는 점이다. 이는 자살의 세부적인 원인이 인간관계와 밀접하게 연결되어 있음을 의미하며 관계적 원인이 한국 사회의 큰 특징적인 결과로 나타난다는 보건복지부의 조사결과와도 같은 의미를 가진다(안용민, 2019).

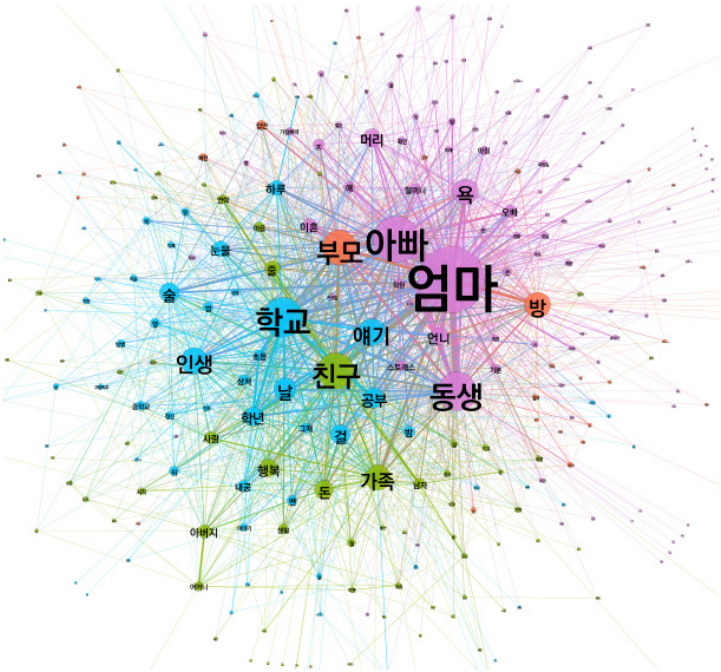
토픽별로 분류된 문헌들에서 각각 동시출현한 단어들을 네트워크로 시각화한 결과는 〈그림 5, 6, 7, 8〉과 같다. “복합” 주제로 분류된 문헌들에서는 하나의 글에 다양한 주제가 등장하기 때문에 2,3,4번째 토픽인 “정신질환”, “가족”, “학업” 관련 단어가 모두 보이는 것을 확인할 수 있다. “정신질환” 주제로 분류된 문헌들에서는 ‘우울’, ‘우울증’ 등 정신질환 이름도 나타나지만 ‘엄마’, ‘친구’, ‘부모’ 등 우울증의 원인이 되는 대상도 함께 등장하는 것을 확인할 수 있다. “가족” 주제로 분류된 문헌들에서는 ‘아빠’, ‘엄마’ 다음으로 ‘동생’ 단어의 출현 빈도가 높게 나오는 것을 확인할 수 있다. 주로 동생과 비교당해서 힘들어하는 내용이 많이 확인된다. “학업” 주제로 분류된 문헌들에서는 ‘공부’, ‘시험’, ‘성적’이 가장 주요한 단어이지만 ‘엄마’, ‘부모’, ‘선생’, ‘친구’ 등 관계적으로 영향을 주는 주체들도 함께 등장함을 알 수 있다. 이로써 주제는 서로 다르게 분류되지만, 주요 단어로 엄마, 부모, 가족, 친구 등 인간관계와 관련된 단어가 공통으로 등장하는 것을 확인할 수 있다.



〈그림 5〉 “복합”으로 분류된 문헌들의 동시출현 단어 네트워크



〈그림 6〉 “정신질환”으로 분류된 문헌들의 동시출현 단어 네트워크



〈그림 7〉 “가족”으로 분류된 문헌들의 동시출현 단어 네트워크

〈그림 8〉 “학업”으로 분류된 문헌들의 동시출현 단어 네트워크

5. 결 론

본 연구는 사회적인 문제인 자살 예방에 도움이 되기 위해 본인의 의견을 솔직하게 공유하는 소셜미디어에서 자살 관련 문헌을 찾고 해당 문헌의 자살 경향 및 비경향여부를 주석처리하여 자살 관련 말뭉치를 구축하였다. 또한, 자살 경향 문헌을 자동 분류할 수 있는 딥러닝 모델을 제작하였고 토픽모델링을 통해 주요 자살 요인을 찾고 이를 세부적으로 분석하였다는 점에서 의의를 갖는다.

자살 경향 말뭉치에 기반한 자살 경향 문헌 판별 딥러닝 모델 중 가장 적합한 모델을 찾기 위한 딥러닝 자동 분류 모델 실험 결과는 다음과 같다. 실험 후 각 모델(LSTM, BERT, ELECTRA)의 성능을 비교하였을 때 ELECTRA의 성능이 93.33%의 정확도로 가장 높았음을 확인하였다. 이를 통해 ELECTRA 계열의 모델 성능이 BERT, LSTM 계열의 모델의 성능보다 더 높음을 알 수 있었고 특히, 한국어 모델인 KoELECTRA로 구축한 자동 분류 모델이 가장 적합한 모델임을 확인하였다. 하루에도 자살 관련된 많은 글이 온라인상에 오르고 있기에 본 연구의 결과인 자살 경향 여부를 자동으로 분류하는 모델을 활용한다면 자살 경향 문헌을 탐지하고 빠르게 대응할 수 있을 것으로 기대한다.

자살 경향 말뭉치를 통해 파악된 자살의 주요 요인을 확인하기 위한 토픽모델링의 결과는 다음과 같다. 대표적인 자살 요인으로는 “정신질

환", "가족", "학업", "경제", "신체질환", "진로"로 확인되었다. 하지만, 가장 많은 문헌이 "복합적인 이유"로 분류되어 자살을 생각하는 이유로 여러 가지 요인이 복합적으로 작용하고 있음을 알 수 있다. 관련 어휘를 찾기 위해 동시출현 단어 분석 및 네트워크 시각화하였을 때 상위 4개의 토픽 문헌에서 모두 인간관계와 관련된 단어가 많이 등장하는 것으로 확인하여 관계적인 문제가 자살의 가장 주요한 요인임을 확인할 수 있었다.

추가로 문헌 주석처리 시 연령을 공개한 경우 별도로 표기하였는데 1,496개의 자살 경형 문헌에서 386개(25.8%)의 글이 연령을 표기하였고 이 중 323개의 글이 10대(초, 중, 고, 학생)로 나타나 이후 연령대별 자살 문헌을 추가로 수집 및 분석한다면 후속 연구로 청소년의 자살 관련 연구가 가능함을 발견하였다.

글의 특성상 한 문헌에 여러 주제가 있어서 이를 토픽모델링으로 정확하게 분류하기 어렵다는 점은 이 연구의 한계로 보인다. 향후 다양한 주제가 있는 문헌을 자동으로 분류할 수 있는 토픽모델링 연구가 진행된다면 더욱 정확한 의미를 찾을 수 있을 것으로 기대한다.

또한, 향후 자살 요인 자동분류 모델 개발과 자살 연령대 자동 분류 모델 개발 등 다양한 연구로 발전시켜 나갈 수 있을 것으로 기대한다.

참 고 문 헌

김경민, 김규경, 조재춘, 임희석 (2018). 한국 전통문화 말뭉치구축 및 Bi-LSTM-CNN-CRF 를 활용한 전통문화 개체명 인식 모델 개발. 한국융합학회논문지, 9(12), 47-52. https://doi.org/10.15207/JKCS.2018.9.12.047

김나리, 이남주 (2021). 토픽모델링과 동시출현 단어 분석을 활용한 환자안전 관련 사회적 이슈의 변화. 한국콘텐츠학회논문지, 21(1), 92-104. https://doi.org/10.5392/JKCA.2021.21.01.092

김현지, 박서정, 송채민, 송민 (2019). 조현병과 정신분열병에 대한 뉴스 프레임 분석을 통해 본 사회적 인식의 변화. 한국문헌정보학회지, 53(4), 285-307.

박찬준, 박기남, 문현석, 어수경, 임희석 (2021). 인공신경망 기계번역에서 말뭉치 간의 균형성을 고려한 성능 향상 연구. 한국융합학회논문지, 12(5), 23-29. https://doi.org/10.15207/JKCS.2021.12.5.023

보건복지부 (2021. 5. 6.). 2021년 1분기 「코로나19 국민 정신건강 실태조사」. 출처: http://www.mohw.go.kr/react/al/sal0301vw.jsp?PAR_MENU_ID=04&MENU_ID=0403&CONT_SEQ=365582&page=1

서하림, 송민 (2019). 소셜미디어를 통한 우울 경향 이용자 담론 주제 분석. 정보관리학회지, 36(4), 207-226. https://doi.org/10.3743/KOSIM.2019.36.4.207

송민 (2017). 텍스트 마이닝. 서울: 청람.

안용민 (2019). 2018 자살실태조사. 보건복지부.
출처: http://www.mohw.go.kr/react/jb/sjb030301vw.jsp?PAR_MENU_ID=03&MENU_ID=032901&CONT_SEQ=350956

이범오 (2020). 인터넷 자살 암시글 유형분류에 관한 연구. 한국민간경비학회보, 19, 153-172.

이수빈, 김성덕, 이주희, 고영수, 송민 (2021). 딥러닝 자동 분류 모델을 위한 공황장애 소셜미디어 코퍼스 구축 및 분석. 정보관리학회지, 38(2), 153-172.
https://doi.org/10.3743/KOSIM.2021.38.2.153

이진, 정진경, 김한샘 (2021). 딥러닝 언어모델의 한국어 학습자 말뭉치 원어민성 판단 결과 분석 연구. 언어와 문화, 17(1), 155-177. http://doi.org/10.18842/klaces.2021.17.1.007

정상혁 (2005). 우리나라 자살의 사회・경제적 비용부담에 관한 연구. 국립서울병원 국립정신보건교육연구센터.

중앙자살예방센터 (2012). 자살예방교육 매뉴얼.

통계청 (2020. 9. 22.). 2019년 사망원인통계 결과.
출처: http://kostat.go.kr/portal/korea/kor_nw/3/index.board?bmode=read&aSeq=385220

Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.

Devlin, J. (2021). Bert multilingual. GitHub. Available: https://github.com/google-research/bert/blob/master/multilingual.md

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Du, J., Zhang, Y., Luo, J., Jia, Y., Wei, Q., Tao, C., & Xu, H. (2018). Extracting psychiatric stressors for suicide from social media using deep learning. BMC medical informatics and decision making, 18(2), 77-87. https://doi.org/10.1186/s12911-018-0632-8

Jeon, H. (2021). KoBERT. GitHub. Available: https://github.com/SKTBrain/KoBERT

Kaplan, M. S., Huguet, N., McFarland, B. H., & Newsom, J. T. (2007). Suicide among male veterans: a prospective population-based study. Journal of Epidemiology & Community Health, 61(7), 619-624. http://dx.doi.org/10.1136/jech.2006.054346

NEWS 18 (2021, October 15). Covid-19 Spiked Suicide Attempts in Teenage Girls by 51%: US CDC. Available: https://www.news18.com/news/lifestyle/covid-19-spiked-suicide-attempts-in-teenage-girls-by-51-us-cdc-3848564.html

NIMH (2021, June 6). Suicide. Available: https://www.nimh.nih.gov/health/statistics/suicide

Park, J. (2020). KoELECTRA: Pretrained ELECTRA model for Korean. GitHub. Available: https://github.com/monologg/KoELECTRA

Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V. A., & Boyd-Graber, J. (2015). Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, 99-107.

Resnik, P., Garron, A., & Resnik, R. (2013). Using topic modeling to improve prediction of neuroticism and depression in college students. In Proceedings of the 2013 conference on empirical methods in natural language processing, 1348-1353.

Song, Min (2021, June 6). treform. GitHub. Available: https://github.com/MinSong2/treform

Steyvers, M. & Griffiths, T. (2007). Probabilistic topic models. In Handbook of latent semantic analysis. New Jersey: Psychology Press, 439-460.

Thompson, P., Bryan, C., & Poulin, C. (2014). Predicting military and veteran suicide risk: Cultural aspects. In Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, 1-6. http://dx.doi.org/10.3115/v1/W14-3201

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems, 5998-6008. https://dl.acm.org/doi/10.5555/3295222.3295349

Won, H. H., Myung, W., Song, G. Y., Lee, W. H., Kim, J. W., Carroll, B. J., & Kim, D. K. (2013). Predicting national suicide numbers with social media data. PloS one, 8(4), e61809. https://doi.org/10.1371/journal.pone.0061809

World Health Organization (2021, June 6). Suicide prevention. Available: https://www.who.int/health-topics/suicide#tab=tab_1

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

Ahn, Y. (2019). 2018 National Survey on Suicide. Ministry of Health and Welfare. Available: http://www.mohw.go.kr/react/jb/sjb030301vw.jsp?PAR_MENU_ID=03&MENU_ID=032901&CONT_SEQ=350956

Jung, S. H. (2005). The socioeconomic burden of suicide and depression in South Korea. National Center for Mental Health.

Kim, G. M., Kim, K., Jo, J., & Lim, H. S. (2018). Constructing for Korean traditional culture corpus and development of named entity recognition model using Bi-LSTM-CNN-CRFs. Journal of the Korea Convergence Society, 9(12), 47-52. https://doi.org/10.15207/JKCS.2018.9.12.047

Kim, H. J., Park, S. J., Song, C. M., & Song, M. (2019). Text mining driven content analysis of social perception on schizophrenia before and after the revision of the terminology. Journal of the Korean Society for Library and Information Science, 53(4), 285-307.

Kim, N. & Lee, N. J. (2021). An analysis of changes in social issues related to patient safety using topic modeling and word co-occurrence analysis. The Korea Contents Society, 21(1), 92-104. https://doi.org/10.5392/JKCA.2021.21.01.092

Korea Suicide Prevention Center (2012). Suicide Prevention Training Manual.

Lee, B. O. (2020). A study on the classification of internet suicide suggestions types. Journal of the Korean Society of Civil Security, 19, 153-172.

Lee, J., Jung, J., & Kim, H. (2021). A study on the judgment of nativelikeness of korean learner corpus by deep learning language model. Korean Language And Culture Education Society, 17(1), 155-177. http://doi.org/10.18842/klaces.2021.17.1.007

Lee, S., Kim, S., Lee, J., Ko, Y., & Song, M. (2021). Building and analyzing panic disorder social media corpus for automatic deep learning classification model. Journal of the Korean Society for Information Management, 38(2), 153-172. https://doi.org/10.3743/KOSIM.2021.38.2.153

Ministry of Health and Welfare (2021, May 5). First quarter of 2021 「Corona 19 National Mental Health Survey」. Available: http://www.mohw.go.kr/react/al/sal0301vw.jsp?PAR_MENU_ID=04&MENU_ID=0403&CONT_SEQ=365582&page=1

Park, C., Park, K., Moon, H., Eo, S., & Lim, H. (2021). A study on performance improvement considering the balance between corpus in Neural Machine Translation. Journal of the Korea Convergence Society, 12(5), 23-29. https://doi.org/10.15207/JKCS.2021.12.5.023

Seo, H. & Song, M. (2019). An analysis of the discourse topics of users who exhibit symptoms of depression on social media. Journal of the Korean society for information management, 36(4), 207-226. https://doi.org/10.3743/KOSIM.2019.36.4.207

Song, M. (2017). Textmining. Seoul: Chungram.

Statistics Korea (2020, September 22). The results of statistics on causes of death in 2019. Available: http://kostat.go.kr/portal/korea/kor_nw/3/index.board?bmode=read&aSeq=385220