Predicting Daily Emotional Distribution and Mental Health Index via LLM and Further Transfer Learning

서강대학교 인공지능학과 120240328 정가연

Abstract

본 연구는 국민들의 정신건강 상태를 모니터링 하기 위해서 LLM(Large Language Model)과 유튜브 데이터 기반의 일자 별 감정 분포와 정신건강 지수를 예측하는 것을 목적으로 한다. AI HUB의 감정 레이블 데이터를 기반으로 4가지 감정을 예측하는 다중 분류 모델을 구축한다.

다중 분류 모델은 구축 방법은 2단계로 구성된다. 첫 번째 단계는 Basic Transfer Learning으로 Pre-Trained된 BERT 기반의 모델을 불러온 후, 대화 감정 레이블 데이터를 활용해서 다중 감정 분류를 위한 Fine-Tuning을 진행한다. Pre-Trained 모델로 BERT-base, RoBERTa-base, RoBERTa-large, KoELECTRA를 사용하며 가장 좋은 성능을 보인 모델을 최종 선택한다.

두 번째 단계는 Further Transfer Learning으로 Fine-Tuning된 기존 다중 감정 분류 모델에 추론을 위해 사용할 라벨 데이터셋을 활용해서 한 번 더 Fine-Tuning을 진행한다. 기존 다중 감정 분류 모델로는 Basic Transfer Learning에서 가장 좋은 성능을 보인 모델을 활용했으며, 추론에 사용할 라벨 데이터셋은 생성형 언어 모델인 Gemma를 사용해서 구축한다.

모델의 추론 결과를 활용해서 감정 별 분포와 정신 건강 지수 계산 후 도출한다. 도출된 지수를 일자 별, 월 별로 시계열 자료로 표현함으로써 국민들의 정신건강 지수를 모니터링하는데 활용한다.

1.Introduction

Economist Intelligence Unit(EIU)의 '정신건강 및 통합'보고서에 의하면 한국의 아시아·태평양 정신 보건 통합지수(The Asia-Pacific Mental Health Integration Index)는 75.9점으로 OECD 국가 평균(77.3점)보다 낮은 수치를 기록했다. 해당 지수는 정책 환경, 고용 기회, 의료 접근성, 거버넌스, 총 4개의 범주로 구성되었으며 전반적인 정신건강 상태를 나타낸 수치를 알 수 있다. 2020년 기준, 한국의 자살률은 10만 명당 13.9명으로 OECD 국가 평균(11.3명)보다 높으며 우울증 유병률의 경우 세계 1위(36.8%) 수준이다. 이와 같이 한국은 정신건강 지수 측면에서 개선할 여지가 많은 상황이다. 우울과 불안 등 국민들의 정신건강을 위협하는 변수들이 일반화되고 고착화되고 있는 현 상황에서, 국민들의 정신건강을 모니터링함으로써 적절하고 효율적인 대응이 가능하도록 개선 방안을 마련하는 작업이 매우 중요한 시점이다.

한국전자통신연구원(ETRI)에 의하면 'AI'와 'Mental Health' 관련 키워드의 연구 논문은 2000년부터 2019년까지 총 1,607건 출판되며 급증하는 추세이다. 이렇게 정신 질환이나 장애를 예방하고 진단하기 위해서 인공지능(AI)을 활용하는 경우가 많아지고 있다. 관련된 연구로는 보건복지부의 'AI를 활용한 자살위험 예측 기술 및 AI 기반 정신건강 기술플랫폼 개발', 강남세브란스 병원의 'AI 기반의 챗봇을 활용한 공황장애 치료', 국립 보건원 NIH(National Institutes of Health)의 '정신장애 예측 및 치료를 위한 AI 도입 연구' 등이 있다. 특히, 우울증 환자수, 자살률, 정신 질환 관련 내원자 수 등 정신건강과 관련된 지표를 예측하는데 AI는 유용하게 활용될 수 있다.

코로나19가 도래하며 대면 상황의 축소로 트위터, 인스타그램, 유튜브 등 SNS(Social Network Service)의 사용량이 급증하였으며 비대면 온라인 소통이 증가했다. 이런 상황 속에서 SNS 기반의 텍스트 데이터는 NLP(Natural Language Processing)기술의 발전과 함께 예측이나 감정 분류 등에 활발하게 사용되고 있다. SNS에 게시된 텍스트가 긍정적인 의견인지 부정적인 의견인지 분석함으로써 사회에 어떤 정보가 유행하는지, 어떤 문제가 있는지 등을 파악함으로써 사회 문제의 해결과 예방을 위해 활용될 수 있다.

본 논문에서는 감성대화말뭉치 데이터와 유튜브 댓글 데이터를 활용하여 일자 별 감정을 예측함

으로써 감정 분포와 정신건강 지수를 예측하는 방법론을 제안한다. 먼저, AI HUB 감성대화말뭉치 데이터와 감정 분류를 위한 대화 음성 데이터셋을 수집한 후 전저리를 진행함으로써 4가지의 감정 레이블 데이터셋을 구축한다. 모델 구축 첫번째 단계에서는 Basic Transfer Learning 방법론을 활용한다. Pre-Trained된 BERT 기반의 모델을 불러온 후, 감정 레이블 데이터로 Fine-Tuning을 진행한다. Pre-Trained 모델로 BERT-base, RoBERTa-base, RoBERTa-large, KoELECTRA를 사용해서 분류 성능을 확인 후 가장 좋은 성능의 모델을 선택한다. 선택된 모델 기반으로 Optuna Parameter Optimization을 수행함으로써 성능 고도화를 진행한다. 두번째 단계에서는 Further Transfer Learning 방법론을 활용한다. Fine-Tuning된 기존 다중 감정 분류 모델에 추론을 위해 사용할 라벨 데이터셋을 활용해서 한 번 더 Fine-Tuning을 진행한다. 이를 통해서 유튜브 댓글의 특징을 반영한 감정 분류 모델을 구축할 수 있다. 기존 다중 감정 분류 모델로는 Basic Transfer Learning에서 가장 좋은 성능을 보인 모델을 활용했으며, 추론에 사용할 라벨 데이터셋은 생성형 언어 모델인 Gemma를 사용해서 구축한다. 모델의 추론 결과를 활용해서 감정 별 분포와 정신 건강 지수 계산 후 도출한다. 도출된 지수를 일자 별, 월 별로 시계열 자료로 표현함으로써 국민들의 정신건강 지수를 모니터링하는데 활용한다.

논문의 구성은 다음과 같다. 1장 서론에 이어서 2장에서는 본 논문과 관련된 연구 들을 소개한다. 3장에서는 본 논문에서 제안하는 연구 방법, 4장에서는 데이터 수집 및 처리와 함께 제안 방법론들의 실험 환경 및 결과를 소개한다. 마지막 5장에서는 결론과 향후 연구 방향을 기술한다.

2.Related Work

2.1 BERT

BERT(Bidirectional Encoder Representations from Transformers)는 자연어 처리(NLP) 분야에서 혁신적인 모델로, 다양한 NLP 작업에서 탁월한 성능을 보인다. BERT는 트랜스포머 모델의 양방향성을 이용하여 문맥을 고려한 단어의 의미를 학습한다[1]. 특히, BERT는 사전 학습(pre-training)과 미세 조정(fine-tuning)의 두 단계를 거치는데, 이는 다양한 NLP 작업에 맞춤형으로 적용할 수 있다. 예를 들어, Devlin et al.(2019)은 BERT가 질의 응답(QA), 감정 분석, 이름 인식 엔티티(NER) 등 여러 벤치마크 데이터셋에서 기존의 최첨단 모델들을 능가한다고 보고하였다[2]. BERT의 성공 이후, 여러 변형 모델들이 제안되었는데, 그 중 RoBERTa, ELECTRA, ALBERT 등이 두드러진다. RoBERTa는 BERT의 사전 학습을 더욱 최적화하여, 더 긴 학습 시간과 큰 배치 크기를 사용하여 성능을 향상시킨 모델이다. Liu et al.(2019)은 RoBERTa가 다양한 벤치마크에서 BERT를 능가한다고 보고하였다[3]. ELECTRA는 Generative Adversarial Network(GAN) 개념을 도입하여 더 효율적인 사전 학습을 가능하게 한다. Clark et al.(2020)은 ELECTRA가 같은 컴퓨팅 자원으로 BERT보다 더 나은 성능을 발휘한다고 밝혔다[4]. ALBERT는 Parameter 공유와 Encoder-Decoder 구조를 통해 모델의 효율성을 높인 모델로, 더 적은 Parameter로도 BERT와 동등하거나 더 나은 성능을 보여준다[5].

2.2 Gemma

Gemma는 대규모 언어 모델(LLM) 중 하나로, 다양한 텍스트 생성 및 이해 작업에서 뛰어난 성능을 보여주고 있다. Gemma는 대규모 코퍼스를 사용하여 사전 학습되었으며, 트랜스포머 아키텍처를 기반으로 하며 Multi Encoder-Decoder 구조를 채택하고 있다. 이 모델은 입력된 텍스트를 처리하여 내재된 의미를 파악하고, 이를 기반으로 새로운 텍스트를 생성한다. 특히 문맥을 파악하는 능력이 뛰어나며, 문장 생성, 요약, 번역 등 다양한 NLP 응용 분야에 적용될 수 있다[6]. 예를 들어, 최근 연구에서는 Gemma를 사용한 텍스트 생성 모델이 인간이 작성한 텍스트와 유사한 품질의 문장을 생성할 수 있음을 보여주었다[7].

2.3 Transfer Learning

전이 학습(Transfer Learning)은 한 분야에서 학습한 지식을 다른 관련 분야로 전이하여 모델의 성능을 향상시키는 방법이다. 전이 학습은 주로 두 단계로 이루어진다: 사전 학습(pre-training)과 미세 조정(fine-tuning)이다. 첫 번째 단계인 사전 학습에서는 대규모 데이터셋을 사용하여 모델을 학습시켜 언어의 일반적인 패턴과 구조를 이해하도록 한다. 두 번째 단계인 미세 조정에서는 사전 학습된 모델을 특정 작업에 맞게 조정하여 성능을 최적화한다[8].

전이 학습은 사전 학습된 대규모 모델을 특정 작업에 맞게 미세 조정하는 과정에서 널리 사용된다. 이미지 분류 분야에서는 ImageNet 데이터셋으로 사전 학습된 ResNet, VGG 등을 활용해서 카테고리의 이미지를 분류하며, 자연어 처리 분야에서는 BERT, GPT-3와 같은 사전 학습된 언어 모델을 활용해서 질의 응답, 번역, 요약, 감정 분석 등의 작업을 수행한다. Ruder et al.(2019)은 전이 학습을 통해 다양한 언어 모델이 기존의 개별 학습 방법보다 높은 성능을 발휘할 수 있음을 입증하였다[9]. 이는 특히 데이터가 제한된 상황에서 모델의 일반화 능력을 크게 향상시킨다.

2.4 Sentiment Classification via Text Data

텍스트 데이터를 통한 감정 분류는 NLP의 중요한 응용 분야 중 하나이다. 감정 분류는 텍스트에서 긍정적, 부정적 감정을 자동으로 인식하는 작업을 포함한다. 감정 분류는 특히 소셜 미디어 분석, 고객 리뷰 분석, 여론 분석 등 다양한 실제 세계 응용 분야에서 중요한 역할을 한다.

BERT와 같은 대규모 언어 모델은 이 작업에서 매우 유용하다. BERT는 문맥을 이해하는 능력이 뛰어나기 때문에 텍스트 내의 미묘한 감정적 뉘앙스를 포착할 수 있다. 예를 들어, Liu et al.(2019)은 BERT를 사용한 감정 분석 모델이 기존의 모델보다 우수한 성능을 보이며, 특히 문맥을 이해하는 능력이 감정 분류의 정확도를 크게 향상시킨다고 보고하였다[10]. 또한, 전이 학습을 활용한 감정 분류는 특히 유용하다. 사전 학습된 모델을 감정 분석 작업에 맞게 미세 조정함으로써, 모델은 제한된 데이터로도 높은 성능을 발휘할 수 있다. 예를 들어, 트윗 데이터를 사용한 연구에서는 사전 학습 모델을 미세 조정하여 트윗의 감정을 정확하게 분류할 수 있음을 보여주었다. 이 연구는 감정 분석이 사회적 이벤트의 변화를 실시간으로 모니터링하는 데 유용하다는 것을 입증하였다[11].

3.Proposed Method

3.1 Construct Sentiment Classification Model

감정 분류 모델 구축은 1단계 : Basic Transfer Learning, 2단계 : Hyper Parmeter Tuning, 3단계 Generate Sample Youtube Label Dataset, 4단계: Further Transfer Learning으로 구성된다.

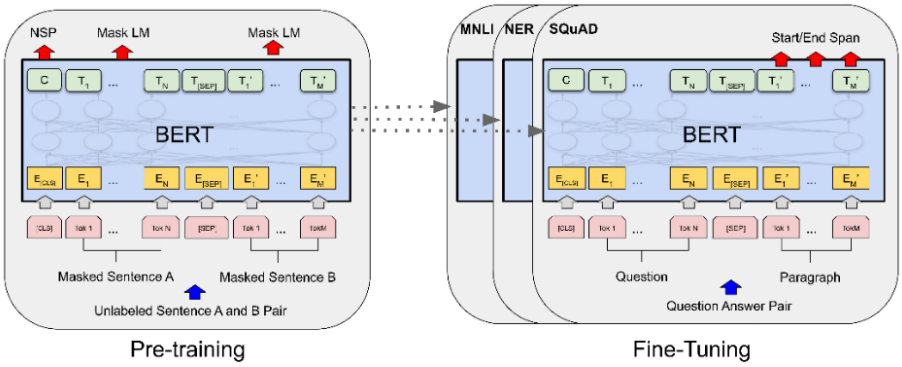
3.1.1 Basic Transfer Learning

첫 번째 단계는 Basic Transfer Learning이다. 기본적인 Transfer Learning으로 Pre-Train된 언어 모델을 불러와 학습한 뒤 목표 테스크에 맞게 Fine-Tuning을 수행한다. AI 모델을 위한 오픈소스 플랫폼인 Huggingface에서 Pre-training 모델로 KLUE-BERT[12], KLUE-RoBERTa-base[13], KLUE-RoBERTa-large[14], KoELECTRA[15] 모델을 불러와서 사용하였다. 3가지 모델 모두 BERT을 기반으로 한 한국어 대규모 언어 모델이다.

KLUE-BERT는 한국어 자연어 처리 작업에 최적화된 BERT 기반 모델로, BERT와 동일하게 Masked Language Modeling을 통해 사전 학습을 수행하되 한국어 위키피디아, 뉴스, 국민청원 등의 대규모 한국어 말뭉치를 사용하였다. KLUE-RoBERTa는 BERT보다 더 큰 배치 사이즈와 더 긴 텍스트 시퀀스를 사용하여 훈련함으로써 성능을 극대화한 모델이다. KoELECTRA는 한국어 기반의 ELECTRA모델로 텍스트를 생성하는 Generator와 생성된 텍스트가 원래의 텍스트인지 아닌지를 판별하는 Discriminator라는 두 가지 네트워크로 구성된다. 이를 통해 MLL보다 더 효율적으로 사전 학습을 할 수 있게 한다.

구체적으로 BERT는 대규모 텍스트 데이터로 사전 학습하여 문맥 이해 능력을 갖추게 하는 Pre-Training과정과 BERT 모델을 통해 특정 작업에 맞게 데이터를 추가로 학습시키는 식으로 Fine-

Tuning을 수행한다. 모델의 아키텍처는 다음과 같다.



[그림1]. (좌) Pre-Training (우) Fine-Tuning

수집된 한국어 말뭉치 데이터는 BERT 모델에 입력할 수 있는 형태로 변환하기 위해서 토큰화 과정을 거친다. BERT의 토큰화는 Word Piece 토큰화 방법을 사용한다. 변환된 데이터를 사용하여 BERT 모델을 Pre-Training한다. 사전 훈련 과정에서는 MLM(Masked Language Model)과 NSP(Next Sentence Prediction) 작업을 수행한다. MLM 작업에서는 입력 문장의 일부 토큰을 마스킹하고, 모델이 이를 예측하도록 한다. 이를 통해 모델은 문맥을 이해하고 단어 간 관계를 학습한다. NSP 작업에서는 두 문장을 입력으로 받고, 그 두 문장이 순서대로 이어져 있는지 여부를 예측한다. 이를 통해 모델은 문장 간 관계를 학습한다.

본 연구에서는 Huggingface에서 Pre-training 모델로 KLUE-BERT, KLUE-RoBERTa-base, KLUE-RoBERTa-large, KoELECTRA 모델을 불러와서 사용하였다.

Model	Embedding Size	Hidden Size	Layers	Heads
KLUE-BERT-base	768	768	12	12
Monologg-KoELECTRA-Base-v3	Discriminator:768 Generator:768	Discriminator:768 Generator:256	Discriminator:12 Generator:12	Discriminator:12 Generator:4
KLUE-RoBERTa-base	768	768	12	12
KLUE-RoBERTa-large	1,024	1,024	24	16

[표 1]. 모델 별 파라미터 크기

Pre-Training된 모델을 감정 분류 작업에 맞게 Fine-Tuning하기 위해 감정에 대한 라벨링 데이터를 준비한다. 각 문장에는 특정 감정(행복, 분노, 슬픔, 중립)에 대한 라벨이 부여된다. 준비된 데이터를 사용하여 BERT 모델을 Fine-Tuning한다. Fine-Tuning 과정에서는 사전 훈련된 BERT 모델의 가중치를 초기값으로 사용하고, 감정 분류 작업에 맞게 조정한다. 테스트 데이터셋을 사용해서 Fine-Tuning된 모델을 성능을 평가한다. 이를 위해 Accuracy, F1 Score, 감정 별 Accuracy을 평가 지표로 사용한다.

3.1.2 Hyper Parameter Tuning

감정 분류 모델의 성능 고도화를 위해서 하이퍼 파라미터 최적화 과정을 수행한다. 하이퍼 파라미터란 기계학습 모델을 제어하기 위한 설정 값으로, 값을 바꾸면서 모델의 성능 조절한다. 하이퍼 파라미터 최적화란 설정 값을 찾는 과정으로, 여러가지 조합을 시도해보고 어떤 조합이 가장 좋은지 평가를 진행하며 최적의 값을 탐색한다. Grid Search, Random Search, Optuna 기반의 TPES 최적화 방법론을 적용해서 성능 최적화 실험을 수행하였다.

'Gird Search'란 사전에 정의한 하이퍼 파라미터 조합을 모두 시도하여 최적의 조합을 찾는 방법으로 모든 조합을 탐색하기 때문에 정확한 최적값을 찾을 수 있다. 하지만 조합이 많을수록 계산 비용이 매우 커진다는 한계를 가진다. 'Random Search'란 사전에 정의한 하이퍼 파라미터 공간에서 무작위로 조합을 선택하여 탐색하는 방법으로 모든 조합을 탐색하지 않기 때문에 계산 비용이 낮고, 랜덤하게 점들을 찍으므로 좋은 조합을 찾을 가능성이 있다. 하지만 무작위로 선택하므로 최적값을 보장하지 못한다. 'TPES'란 확률적으로 좋은 후보 값을 선택하기 위해 이전에 시도한 후보 값들의 분포를 고려한다. 매 시도마다 새로운 후보 값들의 분포를 추정하고, 이를 평가하여 더 좋은 후보 값을 탐색한다. 3가지 방법을 실험해본 결과 'TPES' 최적화 실험을 진행했을 때 가장 좋은 성능 개선을 보였으므로 본 방법론을 사용해서 감정 분류 모델의 파라미터 최적화를 수행했다. 각 모델 기본 성능 및 파라미터 최적화 적용 성능은 아래 표와 같다.

정확도	F1 Score	감정 별 정확도(%)			
		행복	분노	슬픔	중립
90.99%	91.87%	95.29%	85.63%	86.94%	99.35%

[표 2]. RoBERTa-LARGE

정확도	F1 Score	감정 별 정확도(%)			
		행복	분노	슬픔	중립
91.16%	92.07%	95.00%	85.96%	87.24%	99.87%

[표 3]. RoBERTa-LARGE + Hyper Parameter Optimization

학습 데이터에 대한 Fine-Tuning된 모델의 감정 분류 결과는 다음과 같다. 각 문장 별로 감정이

	Sentence	Predict Sentiment
0	진짜 너무너무 실망이야.	슬픔
1	오랜만에 가족들이랑 동해 바다에 가서 놀다 오기로 했어. 그런데 망했어.	슬픔
2	주변 사람들한테 이벤트에 당첨돼서 선물 받았다고 자랑했어.	행복
3	구경하는 사람들이 너무 많아서 볼 수가 없었어.	중립
4	주식 가격이 많이 떨어져서 손해를 봤어.	슬픔

부여된다.

[그림 2]. 훈련 데이터에 대한 감정 예측 결과

3.1.3 Generate Sample Youtube Label Datasest

Further Transfer Learning을 위해서는 추론을 위한 라벨링 데이터가 필요하다. 이를 위해서 Pre-Training된 구글의 Gemma 모델을 불러와서 감정 분류를 위한 프롬포트 기반의 텍스트 생성 파이프라인을 구현한다. 주어진 유튜브 댓글을 토큰화 한 후, 1,024개의 토큰으로 구성된 텍스트를 생성한다. 생성된 텍스트에서 'Answer:'이후의 부분을 추출하여 '행복', '분노', '슬픔', '중립' 중 하나로 감정을 분류하도록 한다. 생성형 모델 특성 상 추출된 텍스트에 오탈자가 포함되는 경우가 생기므로 rule 기반으로 이를 처리해주었다. 최종적으로 9,000개의 유튜브 댓글 라벨링 데이터셋을 구축하였다.

3.1.4 Further Transfer Learning

본 과정을 통해 추론 데이터에 특화된 더욱 고도화된 Transfer Learning을 수행한다. 감정 분류 테스크를 위해 Fine-Tuning된 기존 다중 감정 분류 모델에 유튜브 댓글 라벨링 데이터셋과 기존 감정 라벨링 데이터셋을 사용해서 한 번 더 Fine-Tuning을 진행한다. 이를 통해서 유튜브 댓글의 특징을 반영한 감정 분류 모델을 구축할 수 있다.

기존 감정 분류 모델로는 Basic Transfer Learning에서 가장 좋은 성능을 보였던 KLUE-RoBERTa 모델에 TPSE 하이퍼 파라미터 최적화를 적용한 모델을 사용하였다. Further Transfer Learning을 위한 데이터로는 유튜브 댓글 샘플 데이터셋 9,000개와 기존 감정 라벨링 데이터셋 35,305개를 합하여 사용하였다. 이렇게 구축된 모델은 기존 감정 분류 모델보다 더 좋은 성능을 보였다.

3.2 Inference Youtube Data

Further Transfer Learning을 통해 구축된 유튜브 댓글 특화 감정 분류 모델을 활용해서 전체 유튜브 댓글의 감정을 추론하였다. 원본 유튜브 댓글의 오탈자 제거, 조사 및 접미사와 같은 전처리를 수행한 후 댓글 텍스트를 batch(size=100)로 분할하였다. 각 배치에 대해 다음의 과정을 수행한다. 텍스트들을 토큰화 한 후 모델 입력 형태로 준비한다. 모델을 평가 모드로 전환한 후, 입력 데이터를 이용해 추론을 수행한다. 출력 로짓(logits)에 소프트 맥스(softmax)함수를 적용하여 각 감정에 대한 확률을 계산한다. 그 후, 각 텍스트에 대해 가장 높은 확률을 가진 감정을 예측 레이블로 추출한다. 결과적으로, 각 텍스트에 대해 예측된 감정 레이블과 각 감정에 대한 확률 값을 포함하는 결과를 생성하게 된다 이 결과는 'Text', 'Top_Sentiment', 'Prob_Happy', 'Prob_Angry', 'Prob_Sad', 'Prob_Neutral' 등의 열을 가진 데이터프레임 형태로 저장된다. 최종적으로, 모든 배치의 예측 결과를 합쳐 데이터프레임으로 저장하여 유튜브 댓글 데이터의 감정 분석 결과를 통합했다.

Text	Top_Sentiment	Prob_Happy	Prob_Angry	Prob_Sad	Prob_Neutral	Mental_index_score
한동훈 너무 뻔뻔스럽다, 대한민국 법이 너무 웃켜, 법인이 큰소리 치고있고만.	happy	0.623844	0.011767	0.356197	0.008191	0.255880
애들아고맘다""몰났리나서'내가살았다"코로나가나를인기올려주네'그래'그래잘한다'애들...	happy	0.999581	0.000219	0.000033	0.000167	0.999328
핵발전의 축복! 물부족 국가에서 물폭탄 국가로, 중국에 핵발전소 올 연말이면 49기...	happy	0.999403	0.000472	0.000109	0.000016	0.998823

[그림 3]. 유튜브 댓글 별 감정 및 감정 확률 추론 결과

3.3 Predict Sentiment Distribution & Mental Health Index

유튜브 댓글 추론 결과 댓글 별 각 감정의 예측 확률을 알 수 있으므로, 이를 활용해서 일자 별 감정 분포를 도출했다. 일자 별로 각 감정 별 확률을 더한 뒤 해당 일자의 댓글 수로 나눠 줌으로써 일자 별 평균 감정 분포를 계산했다. 일자 별 평균 감정 분포 계산 산식은 다음과 같다.

$$daily\ sentiment\ ratio_d = \frac{\sum_{i=1}^{Number\ of\ Comments_d} Probability\ of\ Sentiment_i}{Number\ of\ Comments_d}$$

월 별 평균 감정 분포 계산 산식은 다음과 같다.

$$monthly\ sentiment\ ratio_m = \frac{\sum_{i=1}^{Number\ of\ Comments_m} Probability\ of\ Sentiment_i}{Number\ of\ Comments_m}$$

각 댓글 별 감정의 예측 확률 값에 가중합을 적용해서 댓글 별 정신 건강 지수를 계산하였다. 긍정적인 감정인 행복에는 +1의 가중치를 부여했고, 부정적인 감정인 분노와 슬픔에는 -1의 가중치, 중립에는 0의 가중치를 부여한 후 합해주었다. 정신건강 지수 도출 산식은 다음과 같다.

$$Mental\ Health\ Index = (Probability\ of\ happy * +1) + (Probability\ of\ angry\ * -1)$$

$$+(Probability\ of\ sad\ * -1) + (Probability\ of\ neutral\ * 0)$$

4. Experiment

4.1 Dataset

4.1.1 AI HUB의 감성대화말뭉치, AI HUB의 감정 분류를 위한 대화 음성 데이터셋

감정 분류 모델을 학습하기 위한 감정 레이블 데이터셋은 다음과 같다. 감성대화말뭉치에서는 '분노', '슬픔', '행복'에 해당하는 문장-감정 쌍을 추출했고, 감정 분류를 위한 데이터셋에서는 '중립'에 해당하는 문장-감정 쌍을 추출했다. 4가지 레이블이 균형을 이루도록 하였고 총 35,305개의 감정 레이블 데이터셋을 구축하였다.

데이터셋	감정 레이블	문장 수
감성대화말뭉치	분노, 슬픔, 행복	27,884
감정 분류를 위한 대화 음성 데이터셋	중립	7,421

[표 4]. 학습 데이터 개요

분노 문장 수	슬픔 문장 수	중립 문장 수	행복 문장 수	합계
10,417	10,128	7,421	7,339	35,305

[표 5]. 학습 데이터 구성

4.1.2 유튜브 댓글 데이터셋

유튜브 댓글 데이터는 2020년 1월~ 2022년 12월 동안의 뉴스 영상 별 댓글로 구성된 JSON 파일 형태이다. 약 3백만 개의 댓글 행으로 구성된다. 불용어 제거 및 오탈자 교정 전처리를 수행했다.

	publishedAt	videoDate	videoTitle	text
0	2020-01-14T04:23:46Z	2020-01-06	'백인들의 잔치' 골든글로브 뚫었다...또 새 역사 (2020.01.06뉴스데스크MBC)	사실 기생충이란 영화는 그 많은 요소들에 감독이 내포하는 의미가 대단해서 영화를 정...
1	2020-01-08T15:20:37Z	2020-01-06	'백인들의 잔치' 골든글로브 뚫었다...또 새 역사 (2020.01.06뉴스데스크MBC)	골든글로브건 오스카건 여전히 인종에 대한 편견의 벽을 깨야할 부분이 남아있다고 생각...
2	2020-01-07T19:04:44Z	2020-01-06	'백인들의 잔치' 골든글로브 뚫었다...또 새 역사 (2020.01.06뉴스데스크MBC)	각본상이 once upon a time in hollywood 한테 간건 진짜 에바...
3	2020-01-07T15:01:08Z	2020-01-06	'백인들의 잔치' 골든글로브 뚫었다...또 새 역사 (2020.01.06뉴스데스크MBC)	CJ는 연관 없어????\n투자자가 몇프로이지 얘기는 안나오지?
4	2020-01-07T09:31:57Z	2020-01-06	'백인들의 잔치' 골든글로브 뚫었다...또 새 역사 (2020.01.06뉴스데스크MBC)	번역해주신 분도 같이 조명해주세염 \n칸 영화제 영상올라올때는 관련영상도 본거같았는뎅

[그림 5]. 유튜브 댓글 데이터 개요

4.1.3 유튜브 댓글-감정 매핑 샘플 데이터셋

	publishedAt	videoTitle	clean_text	clean_sentiment
0	2020-08-01T07:32:13Z	2명 사망·이재민 150명..충청·전북에 오늘 또 물폭탄 - [LIVE]MBC 뉴스투...	기상청 날씨하나 맞추는게 어렵니	슬픔
1	2020-07-31T10:41:35Z	2명 사망·이재민 150명..충청·전북에 오늘 또 물폭탄 - [LIVE]MBC 뉴스투...	뭔 뉴스에 광고가 이리도 많나요 징허네 진짜	분노
2	2020-07-31T09:59:54Z	2명 사망·이재민 150명..충청·전북에 오늘 또 물폭탄 - [LIVE]MBC 뉴스투...	잠금만 풀면되지 왜 오래들고 있냐증거물 없애려고 그러는거 아냐	슬픔

[그림 6]. 유튜브 댓글-감정 매핑 샘플 데이터 개요

Gemma 모델을 사용해서 유튜브 댓글을 4가지 감정 중 하나로 추론하도록 prompt tuning을 수행해서 얻은 데이터 셋이다. 약 9천 개로 구성되며 특수문자 및 자음, 모음 제거, 특이값 대체와 같은 전처리를 수행하였다.

4.2 Evaluation Metric

4.2.1 Accuracy

정확도는 전체 예측 중에서 실제 레이블과 일치한 예측의 비율을 의미하며, 모델이 얼마나 자주 올바르게 예측하는지 나타내는 척도로 사용된다. 0부터 1까지의 값을 가지며 높을수록 성능이 좋다고 평가한다.

$$Accuracy = \frac{Number\ of\ Correct\ Prediction}{Total\ Number\ of\ Predictions}$$

4.2.2 F1 Score

모델이 True라고 예측한 것 중에서 실제 True인 비율을 나타내는 Precision과 실제 True 중에서 모델이 True라고 정확하게 예측한 비율을 나타내는 Recall의 조화 평균을 나타낸다. 0부터 1까지의 값을 가지며 높을수록 성능이 좋다고 평가된다.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

4.2.3 Per-Class Accuracy

Confusion Matrix를 통해 각 클래스의 정확도를 계산한다 예측 결과와 실제 레이블 간의 매핑을 행렬 형태로 나타냄으로써, 각 클래스의 대각선 값을 추출하며 이는 해당 클래스의 정확도를 나타낸다.

$$Accuracy_k = \frac{CM_{k,k}}{\sum_j CM_{k,j}}$$

4.3 Sentiment Classification

4.3.1 Basic Transfer Learning Performance

분류 모델 학습을 위해서 훈련 데이터 9 : 평가 데이터 1의 비율로 분할하였고 4개의 감정(행복, 분노, 슬픔, 중립)을 분류하도록 훈련하였다. 분류 모델의 파라미터는 다음과 같다.

Per_device_train_batch_size=4, per_device_eval_batch_size=128, learning_rate-1e-5, weight_decay=0.1, adam_beta1=0.9, adam_beta2=0.98, adam_epsilon=1e-6, num_train_epochs=5, warm_ratio=0.06

각 모델 별 분류 성능은 다음과 같다. 4개의 모델 중 RoBERTa-large의 성능이 가장 좋은 것으로 확인됐다.

Accuracy	**F1 Score**	**Per-class Accuracy**			
		Happy	Angry	Sad	Neutral
90.29%	91.22%	95.15%	84.25%	86.04%	99.61%

[표 6]. BERT-base

Accuracy	**F1 Score**	**Per-class Accuracy**			
		Happy	Angry	Sad	Neutral
90.26%	91.23%	95.15%	83.97%	86.34%	99.48%

[표 7]. KoELECTRA-base

Accuracy	F1 Score	Per-class Accuracy			
		Happy	Angry	Sad	Neutral
90.06%	90.98%	94.44%	83.49%	87.04%	98.97%

[표 8]. RoBERTa-base

Accuracy	F1 Score	Per-class Accuracy			
		Happy	Angry	Sad	Neutral
91.00%	91.87%	95.29%	85.86%	86.94%	99.35%

[표 9]. RoBERTa-large

4.3.2 Hyper Parameter Tuning Performance

성능이 가장 좋았던 RoBERTa-large 모델을 기반으로 하이퍼 파라미터 최적화를 진행했다. Optuna 라이브러리의 TPES Sampler를 활용하였고 기존 모델보다 성능이 개선된 것을 확인할 수 있다.

Accuracy	F1 Score	Per-class Accuracy			
		Happy	Angry	Sad	Neutral
91.16%	92.07%	95.00%	85.96%	87.24%	99.87%

[표 10]. RoBERTa-large

4.3.3 Further Transfer Learning Performance

하이퍼 파라미터 최적화가 적용된 RoBERTa-large 모델에 유튜브 라벨링 데이터셋을 활용해서 한 번 더 Fine-Tuning 시킨 분류 성능이다. 정확도와 F1 Score 모두 가장 좋은 성능을 보였다.

Accuracy	F1 Score	감정 별 정확도(%)			
		행복	분노	슬픔	중립
91.32%	92.25%	92.84%	89.45%	88.30%	97.99%

[표 11]. RoBERTa-large

4.4 Inference

RAM은 128GB, SSD(nvme)은 2TB, GPU는 Titan Xp으로 구축된 환경에서 유튜브 댓글 데이터를 추론 속도를 측정해보았다. Model은 추론에 사용한 모델, Batch Size는 댓글 수, Inference Count는 추론 횟수, Inference Time은 평균 추론 시간을 의미한다. 감정 분류 테스크에 Fine-Tuning된 RoBERTa-large 모델로 유튜브 댓글을 추론을 했을 때의 속도는 다음과 같다.

Model	Batch Size	Inference Count	Inference Time	Inference Time (1batch)
RoBERTa-base	10	100	0.04초	0.004초
RoBERTa-base	100	100	0.5초	0.005초
RoBERTa-large	10	100	0.12초	0.012초
RoBERTa-large	100	100	1.6 초	0.016초

[표 12]. 유튜브 데이터 감정 추론 속도

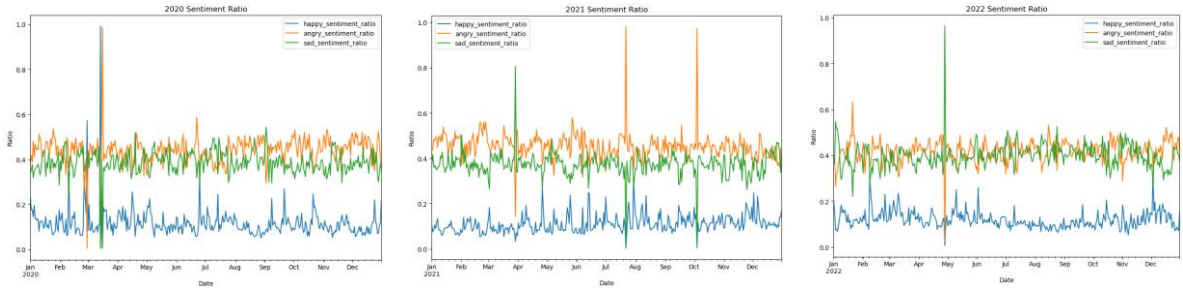
유튜브 댓글 감정 추론 결과 중립 댓글의 비율이 가장 높았고 분노 댓글, 슬픔 댓글, 행복 댓글 순의 비율로 구성되었다.

추론 댓글	추론 댓글 수	추론 댓글 비율
중립 댓글	1,500,286	62%
분노 댓글	632,930	26%
슬픔 댓글	173,058	8%
행복 댓글	92,244	4%

[표 13]. 유튜브 댓글 감정 추론 결과

4.6 Daily Sentiment Distribution & Mental Health Index

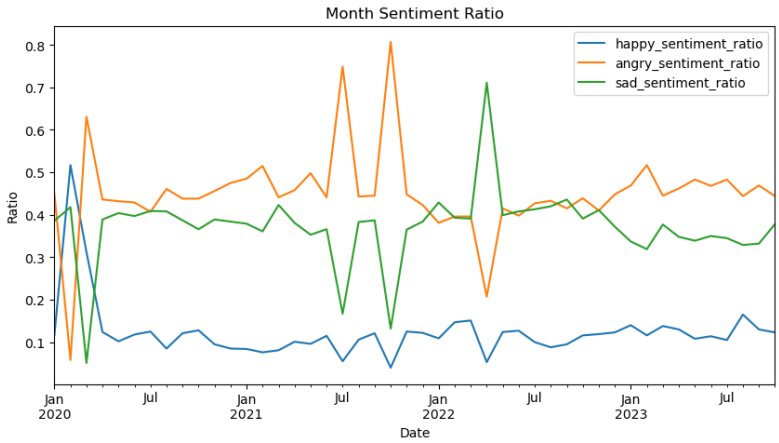
유튜브 댓글 데이터 추론 결과 도출된 일별 감정 분포는 다음과 같다.



(a) 2020년 일자 별 감정 분포 (b) 2021년 일자 별 감정 분포 (c) 2022년 일자 별 감정 분포

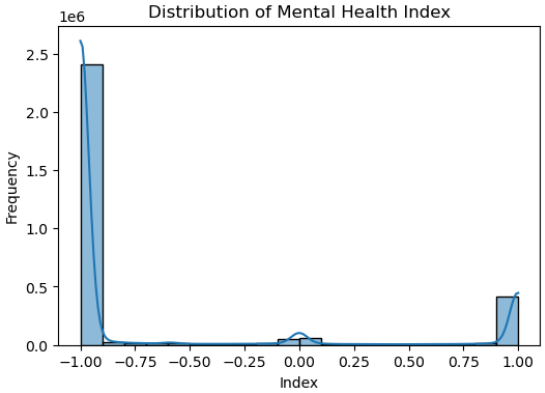
[그림 7]. 일별 감정 분포

월별 감정 분포는 다음과 같다.



[그림 8]. 월별 감정 분포

정신건강 지수의 분포는 다음과 같다. 0점과 -1점에 몰려 있는 것을 확인할 수 있다.



[그림 9]. 정신건강 지수 분포

5. Conclusion

5.1 요약

본 논문에서는 AI HUB 감성대화말뭉치 데이터와 유튜브 댓글 데이터를 활용하여 국민들의 정신건강 지수를 예측하고 모니터링하는 방법론을 제안하였다. 구체적으로, Basic Transfer Learning 방법론을 통해 BERT 기반의 Pre-Training 모델을 Fine-Tuning하고, Optuna Parameter Optimization을 통해 성능을 고도화하였다. 또한, Further Transfer Learning 방법론을 적용하여 유튜브 댓글의 특징을 반영한 감정 분류 모델을 구축하였다. 이를 통해 일자별, 월별 감정 분포와 정신건강 지수를 시계열 자료로 표현하고 모니터링하는 방안을 제시하였다.

본 연구의 주요 성과는 다음과 같다:

다중 감정 분류 모델 구축: Basic Transfer Learning을 통해 다양한 BERT 기반 모델을 Fine-Tuning하여 최적의 감정 분류 성능을 확인하였다.

성능 고도화: Optuna Parameter Optimization을 통해 선택된 모델의 성능을 극대화하였다.

Further Transfer Learning: 유튜브 댓글 데이터를 반영한 맞춤형 감정 분류 모델을 개발하였다.

정신건강 지수 예측 및 모니터링: 감정 분포를 바탕으로 정신건강 지수를 계산하고, 이를 시계열 자료로 시각화 하여 국민들의 정신건강 상태를 모니터링하는 시스템을 제안하였다.

5.2 향후 연구 방향 및 개선 방안

향후 연구에서는 다음과 같은 개선 방안을 통해 본 연구를 확장하고자 한다:

데이터 다양화 및 확대: 다양한 소스의 데이터를 수집하여 데이터셋을 확대하고, 감정 레이블의 다양성을 높여 모델의 일반화 성능을 향상시킨다. 예를 들어, 트위터, 페이스북 등 다른 SNS 플랫폼의 데이터를 추가적으로 수집하여 감정 분석의 범위를 확장할 수 있다.

실시간 모니터링 시스템 구축: 실시간으로 데이터를 수집하고 분석할 수 있는 시스템을 개발하여, 국민들의 정신건강 상태를 실시간으로 모니터링하고 즉각적인 대응이 가능하도록 한다. 이를 통

해 정신건강 문제를 조기에 발견하고 대응할 수 있다.

정밀한 정신건강 지수 개발: 현재의 감정 분포 기반 정신건강 지수 외에도, 더 정밀한 지수를 개발하기 위해 다양한 심리학적 요인과 사회적 변수를 고려한 모델을 구축한다. 예를 들어, 경제 상황, 사회적 고립도 등의 변수를 포함하여 정신건강 지수의 예측 정확도를 높인다.

본 논문에서 제안한 방법론과 향후 연구 방향을 통해, 국민들의 정신건강 상태를 보다 정확하게 파악하고, 이를 기반으로 효과적인 정신건강 관리 및 정책 수립이 가능할 것으로 기대된다. 지속적인 연구와 개선을 통해, AI 기반의 정신건강 모니터링 시스템이 사회적 문제 해결에 기여할 수 있기를 바란다.

6. References

[1] Economist Intelligence Unit(EIU), 아시아태평양 지역 정신건강통합지수.

[2] 송근혜, 김문구, 박안선, ETRI. Promising Services Based on AI for Mental Health.

[3] Adyan Marendra Ramadhani; Hong Soon Goo. Twitter Sentiment Analysis using Deep Learning Methods. IEEE, 2017.

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. [arXiv preprint arXiv:1810.04805].

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[7] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

[8] 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.

[9] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. arXiv preprint arXiv:1909.11942.

[10] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165.

[11] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 9.

[12] Ruder, S., Peters, M. E., Swayamdipta, S., & Wolf, T. (2019). Transfer Learning in Natural Language Processing. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials (pp. 15-18).

[13] Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. arXiv preprint arXiv:1801.06146.

[10] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

[14] Muhammad, A., Wiratunga, N., & Lothian, R. (2016). Contextual sentiment analysis for social media genres. Knowledge-Based Systems, 108, 92-101.

[15] klue/bert, https://huggingface.co/klue/roberta-base

[16] klue/roberta-base https://huggingface.co/klue/roberta-base

[17] klue/roberta-large https://huggingface.co/klue/roberta-large

[18] monolog/koelectra-base-v3-discriminator https://huggingface.co/monologg/koelectra-base-v3-discriminator