

사용자 기반 Text Cloud Visualization



202155512 금 비
202055567 원윤서

지도교수 조준수

목 차

1. 연구 동기
2. 텍스트 클라우드에 관한 이전 연구
3. 언어모델 대상 실험 평가
 - 3.1 데이터 구성
 - 3.2 사용자 감정 분류
 - 3.2.1 Llama3 미세 조정 모델 감정 분류
 - 3.2.2 Klue 미세 조정 모델 감정 분류
 - 3.2.3 평가 결과
 - 3.3 공감적 응답 생성
 - 3.3.1 Llama3 미세 조정 모델 응답 생성
 - 3.3.2 평가 결과
 - 3.4 MBTI 분류
 - 3.4.1 Llama3 미세 조정 모델 MBTI 분류
 - 3.4.2 Klue 미세 조정 모델 MBTI 분류
 - 3.4.3 평가 결과
 - 3.5 요약 생성
 - 3.5.1 Llama3 미세 조정 요약 생성
 - 3.5.2 평가 결과
 - 3.6 미세 조정 모델과 비조정 모델 성능 비교
 - 3.6.1 두 모델의 멀티턴 대화 학습
 - 3.6.2 평가 결과
 - 3.7 뉴스 요약본 기반 군집화

-
- 3.7.1 평가 결과
 - 3.8 같은 뉴스 topic 내에서의 코사인 유사도
 - 3.9 뉴스 Topic 추천 및 워드클라우드 생성
 - 3.9.1 감정 문장 군집화
 - 3.9.2 감정과 뉴스 Topic의 코사인 유사도
 - 3.9.3 뉴스 문장에서의 키워드 추출 및 텍스트 클라우드 생성
 - 4. 결론 및 추후 연구 과제
 - 5. 참고 문헌
 - 6. 부록

1. 연구 동기

현대 사회는 빅데이터 시대에 접어들면서 방대한 양의 텍스트 데이터가 생성되고 있다. 이러한 데이터는 소셜 미디어 포스트, 온라인 리뷰, 뉴스 기사, 이메일, 연구 논문 등 다양한 형태로 존재하며, 매일 엄청난 속도로 증가하고 있다. 이러한 방대한 텍스트 데이터를 효과적으로 처리하고 의미 있는 정보를 추출하는 것은 매우 중요한 연구 주제이다.

방대한 양의 텍스트 데이터를 처리하고 시각화하는 효과적인 방법의 하나인 텍스트 클라우드 시각화가 있다. 그림 1은 텍스트 클라우드의 예를 보여준다. 텍스트 클라우드는 텍스트 데이터의 단어를 시각적으로 표현하여 텍스트 데이터의 문장 구조와 주요 패턴을 쉽게 이해할 수 있게 해준다. 예를 들어, 단어의 빈도나 중요도에 따라 글자 크기를 다르게 표시하여, 중요한 키워드나 주제가 한눈에 들어오도록 한다. 이는 복잡하고 방대한 양의 텍스트 데이터를 간단하고 직관적으로 받아들일 수 있는 효과적인 도구이다.



그림 1. 텍스트 클라우드의 예

특히, 사용자 기반 텍스트 클라우드 시각화는 사용자가 특정 관심사나 분석 목표에 따라 텍스트 클라우드를 사용자화할 수 있는 기능을 제공한다. 예를 들어, 특정 시간대별로 변화하는 트렌드를 시각화하거나, 특정 주제나 키워드와 관련된 텍스트만을 필터링하여 클라우드를 생성할 수 있다. 이를 통해 사용자는 자신이 원하는 정보를 더 정확하고 빠르게 얻을 수 있다.

또한, 사용자 기반 텍스트 클라우드 시각화는 교육, 연구 등 다양한 분야에

서 활용될 수 있다. 예를 들어, 교육 분야에서는 학생들의 피드백을 시각화하여 교육 방식의 개선점을 찾을 수 있으며, 연구 분야에서는 학술 논문의 주제를 한눈에 파악하여 연구 동향을 쉽게 파악할 수 있다. 이러한 시각화 도구는 복잡한 정보를 쉽게 이해하고 활용할 수 있도록 한다.

본 연구에서는 단순한 텍스트 클라우드를 넘어서, 사용자 맞춤 형식의 텍스트 클라우드를 시각화하고자 한다. 그림 2는 텍스트 그래프의 예를 보여준다. 사용자 맞춤 형식의 텍스트 클라우드와 그래프는 현재 사용자의 기분, 사용자의 MBTI를 반영한 새로운 기법의 시각화이다. 실제 사용자의 기분, MBTI를 얻기 위해 Meta의 생성형 AI인 Llama3 모델[1]을 이용한다. 이때, 사용자는 Llama3 모델과 대화를 하게 되며, 사용자가 입력한 문장과 입력한 문장을 바탕으로 감정과 MBTI를 추론하게 된다. 앞선 세 가지의 Embedded Vector를 통해, 코사인 유사도를 이용하여 사용자 맞춤 뉴스를 추천한다. 또한, 추천한 뉴스에서 추출한 주요 단어들을 통해 텍스트 클라우드를 시각화한다.

2. 텍스트 클라우드에 관한 이전 연구

실제 사람들은 자신의 감정이나 선호에 따라 보고 싶은 것만 보는 경향이 있다는 연구 결과들이 있다. 이를 선택적 지각이라고 부른다. 예를 들어, 사람들이 긍정적인 정보를 더 많이 선택적으로 처리하는 현상[2]을 다룬 연구에서는, 노인들이 더 긍정적인 이미지나 사건에 집중하고 이를 기억하려는 경향이 강하다는 것이다. 또한, 직장 내에서 긍정적인 태도를 가진 직원들이 조직의 문화를 더 긍정적으로 평가하는 경향[3]이 있다는 것이다. 텍스트 클라우드 시각화에 관한 많은 연구가 진행됐는데, 기존의 접근법은 사용자의 감정이나 선호 요소를 고려하지 못하는 한계를 지니고 있다.

텍스트 클라우드에서 어떤 텍스트를 선택해야 하는가에 관한 연구가 있다 [4]. 이 연구에서는 텍스트 클라우드의 구조적 특성을 파악하는 지표를 제시하고, 기존에 여러 텍스트 선택 알고리즘을 평가했다. 또한, 텍스트 클라우드의 시각화 문제를 해결하기 위한 알고리즘을 제안하고, 공간 최적화와 관련된 텍스트 그룹화와 관련된 연구[5]가 있다. 하지만 이러한 연구들은 사용자의 선호도와 특성을 전혀 반영하지 못하는 한계가 있다.

3. 언어모델 대상 실험 평가

LLM으로는 미세 조정(fine-tuning)한 Llama3 모델을 활용하였다. 사용자의 입력을 연속적으로 받기 위해, 공감형 대화를 학습했다. 또한, 사용자의 현재 감정 분류와 MBTI에 대한 데이터를 학습했다. 뉴스 기사의 긴 문장을 요약하기 위한 데이터를 학습했으며, 사용자와 상호작용을 위한 멀티턴(multi-turn) 대화도 학습했다. Embedded Vector를 추출하기 위해서 서울대학교 컴퓨터 언어학 자연어처리 연구실의 KR-BERT 모델[6]을 이용했다. 보다 정확한 감정 분류를 위해 BERT 기반 Klue 모델[7] 또한 이용했다.

3.1 데이터 구성

실험에는 AI Hub에서 구축한 일반 감정 대화, 공감형 대화, Hugging Face에서 구축한 네이버 뉴스 요약, 멀티턴 대화, 그리고 MBTI 관련 네이버 카페에서 크롤링한 MBTI 라벨이 붙어 있는 대화를 이용했다.

일반 감정 대화는 표 1과 같으며 각 발화에는 '행복, 분노, 중립, 슬픔, 혐오, 놀람 공포'의 7가지 감정이 태깅되어 있다.

	행복	분노	중립	슬픔	혐오	놀람	공포
감정(%)	23.47	16.84	16.79	14.70	11.98	9.06	7.15

표 1. 감정 대화 데이터

공감형 대화 데이터는 표 2와 같이, 일반화자, 공감화자로 구분되어 있다. 약 8천개의 대화 쌍으로 이루어져 있다.

	발화
일반화자	“엄마, 요즘에 먹고 있는 영양제 효과가 좋은가 봐요!”
공감화자	“그래? 어떤 효과라도 나타난거니?”

표 2. 공감형 대화 쌍의 예

MBTI 데이터는 표3과 같이 사용자의 발화, 사용자 글에 대한 MBTI가 라벨링 되어 있다.

발화	MBTI
“나만의 욕실에서 테이블 놓고 한잔하면 기분 좋을 거 같거든요 ㅎㅎ”	ENFJ
“저도 카톡 같은 걸 하면 제가 대화를 이끌어나가지는 못하는 것 같아요”	INTP

표 3. MBTI 데이터셋의 예

MBTI는 총 16개의 클래스로 나누어져 있는데, 데이터의 불균형 때문에 전

처리를 하였다. 그림 3은 전처리 전의 MBTI 라벨 클래스 분포, 그림 4는 전처리 후의 MBTI 라벨 클래스 분포를 보여준다.

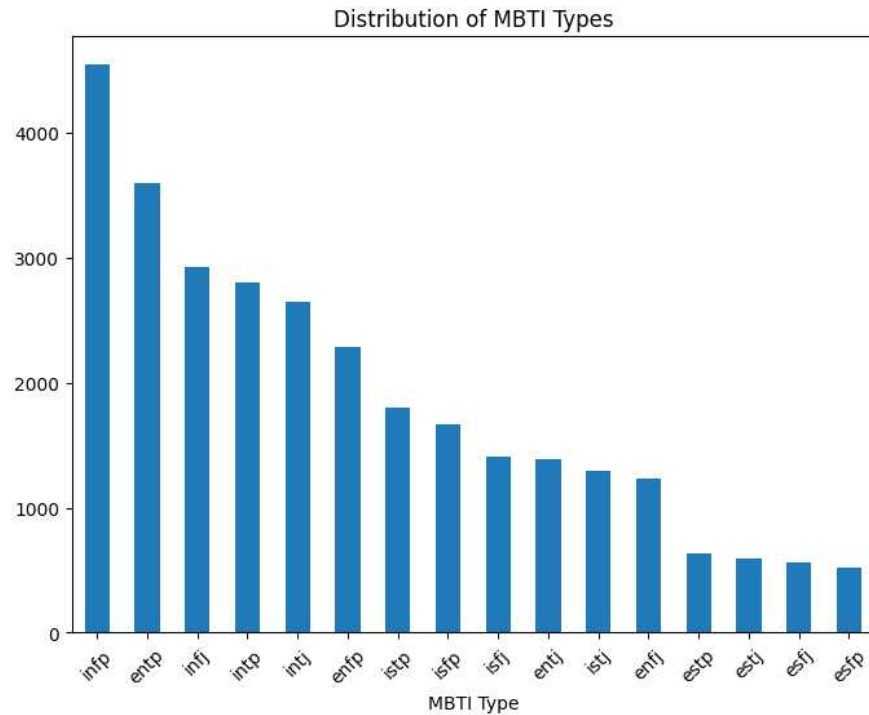


그림 2. 전처리 전의 MBTI 데이터

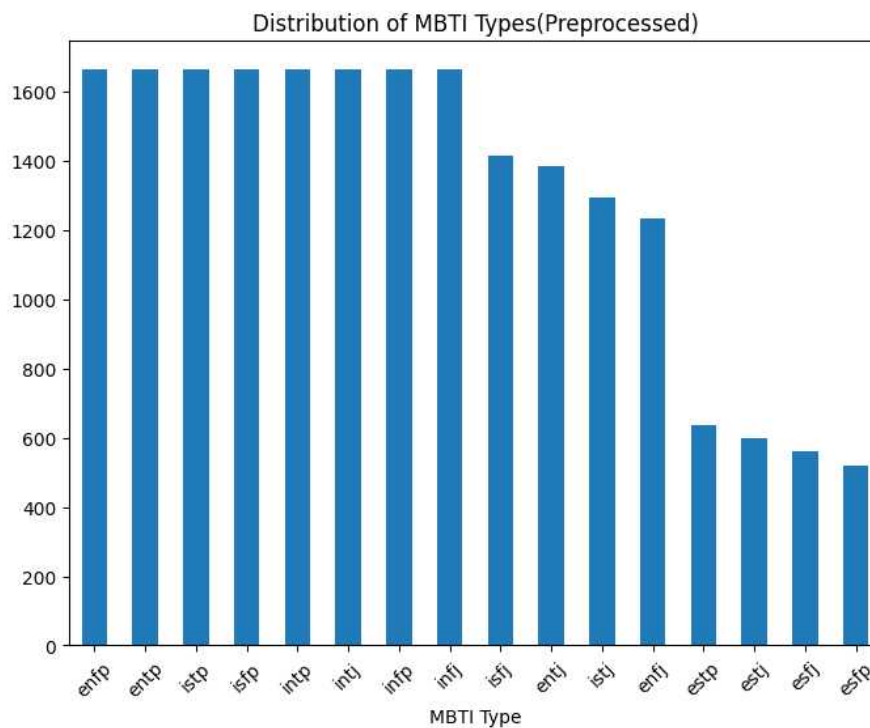


그림 3. 전처리 후의 MBTI 데이터

네이버 뉴스 요약 데이터는 뉴스 원본, 요약본으로 구분되어 있으며, 약 4만 개의 원본 요약 쌍으로 이루어져 있다.

멀티턴 대화는 User와 openai의 GPT의 대화로 이루어져 있으며, 표 5와 같다.

	발화
User	“코딩이 뭐야?”
GPT	“코딩은 인간의 의도된 명령을 컴퓨터가 이해할 수 있도록 번역하는...”
User	“그럼 코딩이 왜 중요해?”
GPT	“코딩은 컴퓨터, 태블릿, 스마트폰 등 컴퓨터를 사용하는 모든 곳에 필수...”

표 4. 멀티턴 대화 데이터셋의 예

3.2 사용자 감정 분류

감정 대화에 포함된 사용자 발화의 감정 레이블을 사용하여 미세 조정한 Llama3와 미세 조정한 Klue의 감정 분류 성능을 비교하였다. 평가 지표로 각 감정 분류별 정확도 (precision), 재현율 (recall) 및 이 두 지표에 기반한 F1-score를 사용하였다.

3.2.1 Llama3 미세 조정 모델 감정 분류

Prompt를 이용하여 Llama3로 감정 분류를 수행하였다. 7가지 감정 레이블을 프롬프트에 포함시키고 그 중 하나로 발화의 감정을 분류할 것을 요구하였다. (부록 표 13 참고) 훈련에서 하이퍼파라미터는 gpu 메모리를 고려하여 batch 크기를 2, learning rate 5e-5, epoch을 1로 설정하였고, optimizer은 AdamW를 사용하였다.

3.2.2 Klue 미세 조정 모델 감정 분류

감정 대화에 포함된 사용자 발화의 감정 레이블을 이용하여 Klue를 미세 조정하고, 감정 분류 성능을 비교하였다. 훈련에서 하이퍼파라미터는 batch 크기를 16, learning rate 1e-5, epoch을 10로 설정하였고, optimizer은 Adam을 사용하였다.

3.2.3 평가 결과

두 분류기의 성능을 비교하기 위하여, 감정 대화 3875개에 대해 두 모델의 분류 결과를 정답과 비교하였다.

Llama3의 분류 정확도는 68.92%였고, Klue 기반 분류기의 정확도는 92.82%였다. Klue 기반 분류기가 Llama3보다 전반적으로 성능이 우수하였으며, 세부 성능은 표 5와 같았다.

	Llama3 미세 조정			Klue 미세 조정		
	Precision	Recall	F1	Precision	Recall	F1
행복	0.7011	0.8208	0.7563	0.9493	0.9514	0.9503
분노	0.6285	0.5450	0.5838	0.8821	0.9089	0.8953
중립	0.6809	0.6025	0.6393	0.9388	0.8980	0.9180
슬픔	0.5382	0.7169	0.6148	0.9265	0.9217	0.9241
혐오	0.8205	0.6688	0.7369	0.9056	0.9171	0.9113
놀람	0.7774	0.8227	0.7994	0.9611	0.9698	0.9655
공포	0.9939	0.5871	0.7382	0.9510	0.9543	0.9527

표 5. Llama3 및 Klue 기반 감정 분류기 성능

3.3 공감적 응답 생성

본 절에서는 공감형 대화쌍 데이터를 이용하여 평가 지표를 통한 성능을 측정하고자 한다. 이때, 감정 인식 없이 공감적 응답을 생성하는 경우에 대해 실험을 진행한다. 프롬프트는 부록 표 14에 정리하였다.

3.3.1 Llama3 미세 조정 모델 응답 생성

앞서 감정 대화로 학습한 모델을 이용하여 Llama3를 미세 조정하고, 성능을 측정하였다. 훈련에서 하이퍼파라미터는 gpu 메모리를 고려하여 batch 크기를 2, learning rate 5e-5, epoch을 1로 설정하였고, optimizer은 AdamW를 사용하였다.

3.3.2 평가 결과

언어 모델이 생성한 공감적 응답은 데이터에 포함된 시스템 발화와의 비교를 통해 평가되었다. 평가에는 SacreBLEU, METEOR, ROUGE-L, Semantic Textual Similarity (STS)가 사용되었다. SacreBLEU는 생성된 문장과 정답 문장에서 형태소가 얼마나 많이 중첩되는지 수치화한 지표인 BLEU[8]를 사용하기 용이하게 표준화한 것이다. ROUGE[9]은 자동 요약 모델의 성능을 평가하는 지표 중 하나로, ROUGE-L은 시스템 요약본과 정답 요약본 간에 겹치는 가장 길이 부분 문자열의 비율에 기반한다. STS는 의미론적 유사도를 측정하는 지표로, 문장 벡터 간의 코사인 유사도로 계산된다.

모델	SacreBLEU	ROUGE-L	STS
Llama3	0.6507	0.0012	0.3763

표 6. Llama3 응답 평가

Llama3가 생성한 공감적 응답에 대한 평가는 공감형 대화 1716개를 대상으로 진행되었다. 특히, 표 6에서 ROUGE-L 지표가 상당히 좋지 못한 결과를 보

였다. 그 이유는, ROUGE-L은 두 문장 간의 공통 부분(서브시퀀스)의 길이를 기반으로 평가하는데, 두 문장 간에 어순이나 구조가 많이 다를 경우 점수가 낮게 나올 수 있다. 한국어는 문장의 어순이 유연하기 때문에 동일한 의미라도 구조가 달라질 수 있기 때문이다.

또한, 한국어는 조사와 어미 변형이 다양하여 동일한 문장이라도 다양한 형태로 표현될 수 있다. 이러한 특성 때문에 ROUGE-L과 같은 지표가 제대로 된 유사도를 측정하지 못할 수 있다.

3.4 MBTI 분류

네이버 카페 글에서 크롤링한 데이터를 사용하여 미세 조정한 Llama3와 미세 조정한 Klue의 성능을 알아보았다. 평가 지표로는 test-set에서의 accuracy를 사용하였다.

3.4.1 Llama3 미세 조정 모델 MBTI 분류

Prompt를 이용하여 Llama3로 MBTI 분류를 수행하였다. 16가지 MBTI 레이블을 프롬프트에 포함시키고 그 중 하나로 발화의 MBTI를 분류할 것을 요구하였다. (부록 표 15 참고) 훈련에서 하이퍼파라미터는 gpu 메모리를 고려하여 batch 크기를 2, learning rate 5e-5, epoch을 1로 설정하였고, optimizer은 AdamW를 사용하였다.

3.4.2 Klue 미세 조정 모델 MBTI 분류

네이버 카페 글 대화에 포함된 사용자 발화의 MBTI 레이블을 이용하여 Klue를 미세 조정하고, MBTI 분류 성능을 비교하였다. 훈련에서 하이퍼파라미터는 batch 크기를 16, learning rate 1e-5, epoch을 10로 설정하였고, optimizer은 Adam을 사용하였다.

3.4.3 평가 결과

분류기의 성능을 알아보기 위해, MBTI 레이블이 있는 대화 4188개에 대해 두 분류 모델 결과를 정답과 비교하였다.

Epoch	Training Loss	Validation Loss	Accuracy
1	2.679700	2.652122	0.127477
2	2.524100	2.523706	0.206254
3	2.377000	2.484131	0.226546
4	2.206900	2.493537	0.232036
5	2.010200	2.544478	0.226068

표 7. Klue 기반 MBTI 분류기 성능

Llama3는 test-set에서 Accuracy 약 11.29%를 기록했다. Accuracy가 너무

좋지 않아, Klue 미세 조정 모델에서 비교 실험을 해보았다. 애초에 실험은 10 epoch까지 진행될 예정이었지만, Klue 모델에서도 Accuracy가 너무 좋지 않아 5 epoch에서 중단하였다. Klue 미세 조정 모델에서의 구체적인 성능은 표 7과 같다.

MBTI 분류 실험이 실패한 원인은 다음과 같다.

- 데이터의 복잡성: MBTI 유형은 사람의 성격을 나타내는 매우 복잡한 요소들을 포함하고 있다. 단순한 대화 데이터만으로 이 복잡한 특성을 완전히 학습하기는 어려울 수 있다. 특히, 짧거나 맥락이 부족한 대화의 경우, 모델이 해당 대화를 특정 MBTI 유형에 맞게 분류하기가 매우 어려울 수 있다.
- 레이블의 주관성: MBTI 레이블은 사람들의 주관적인 판단에 따라 부여될 수 있으며, 이러한 주관성은 모델이 학습하는 데 혼란을 줄 수 있다. 특히, 동일한 대화에 대해 사람마다 다른 MBTI 유형을 레이블링할 가능성이 있다면, 모델이 정확한 패턴을 학습하기 어려워진다.

3.5 요약 생성

네이버 뉴스 요약본 데이터를 이용하여 미세 조정된 Llama3의 성능을 알아보았다. 평가에는 SacreBLEU, METEOR, ROUGE-L, Semantic Textual Similaryity (STS)가 사용되었다.

3.5.1 Llama3 미세 조정 요약 생성

Prompt를 이용하여 Llama3로 요약 생성을 요구하였다 (부록 표 16 참고). 훈련에서 하이퍼라마리터는 batch 크기를 2, learning rate 3e-5, epoch을 1로 설정하였고, optimizer은 AdamW를 사용하였다.

3.5.2 평가 결과

요약 성능을 알아보기 위해, Test Set에 있는 뉴스 원본 데이터 8154개 중 sampling하여 200개에 대해서만 평가하였다.

모델	SacreBLEU	ROUGE-L	STS
Llama3	13.0662	0.4226	0.6804

표 8. Llama3 요약 평가

SacreBLEU의 경우, 20 이상이면 좋은 성능을 나타낸다고 볼 수 있다. 표 8에서 확인할 수 있듯이, 13 정도의 점수가 나왔는데, 이는 비록 20에는 미치지 못하지만 여전히 준수한 성능을 보여준다. 특히, SacreBLEU 점수는 완벽한

일치를 목표로 하는 지표임을 감안할 때, 13이라는 점수는 모델이 참조 요약과 상당한 유사성을 보이며, 충분히 의미 있는 성능을 발휘하고 있음을 시사한다. 더욱 향상된 결과를 얻기 위해서는 추가적인 개선이 필요하겠지만, 현재의 결과도 충분히 긍정적으로 평가할 수 있다.

3.6 미세 조정 모델과 비조정 모델 성능 비교

한국어 학습이 이미 완료된 Llama3 모델과, 한국어 학습이 전혀 안 된 Llama3 모델을 각각 멀티턴(multi-turn) 대화를 학습시켰다. 한국어 학습이 이미 완료된 모델은 서울과학기술대학교의 Llama3 기반 Blossom 모델[10]을 이용했다.

3.6.1 두 모델의 멀티턴 대화 학습

Prompt를 이용하여 Llama3로 멀티턴 대화를 학습했다(부록 표 17 참고). 훈련에서 하이퍼파라미터는 batch 크기를 8, learning rate 3e-5, epoch을 1로 설정하였고, optimizer은 AdamW를 사용하였다.

3.6.2 평가 결과

평가 방법은 학습된 모델에 특정한 문장 10개를 넣고, 모델별 Perplexity를 구했다. Perplexity는 언어 모델의 성능을 평가하는 대표적인 지표로, 모델이 주어진 문장에서 다음 단어를 얼마나 잘 예측하는지를 수치화한 값이다. 이 값은 모델이 다음 단어의 가능성을 얼마나 정확하게 추정하는지를 기반으로 하며, Perplexity 값이 낮을수록 모델의 예측 성능이 우수하다는 것을 의미한다. Perplexity 값이 높을수록 모델이 다음 단어를 예측하는 데 어려움을 겪고 있음을 나타낸다. 이는 모델이 텍스트의 자연스러움이나 일관성을 얼마나 잘 학습했는지를 평가하는 데 유용하다.

Index	문장
1	"오늘은 비가 많이 올 것 같아요. 우산을 꼭 챙기세요."
2	"저는 어제 친구들과 함께 영화를 보러 갔어요."
3	"안녕하세요, 저는 서울에서 온 김민수라고 합니다."
4	"한국의 전통 음식 중 하나는 김치입니다. 매운 맛이 특징이에요."
5	"학생들은 매일 아침 8시에 학교에 가야 해요."
6	"최근 경제 상황이 어려워지면서 물가가 많이 올랐습니다."
7	"인공지능 기술은 날이 갈수록 발전하고 있으며, 다양한 산업에서 활용되고 있습니다."
8	"조선 시대는 한국의 역사에서 중요한 역할을 했습니다."
9	"규칙적인 운동과 균형 잡힌 식단은 건강을 유지하는 데 매우 중요합니다."
10	"다음 달에 제주도로 여행을 가려고 해요. 거기는 자연 경치가 정말 아름답다고 하더군요."

표 9. Perplexity 측정을 위한 문장들

일반적으로 Perplexity 값이 1에서 10 사이인 경우, 모델이 문장을 매우 정확하게 예측하고 있다는 뜻이며, 10에서 50 사이인 경우, 모델이 일반적인 대화나 텍스트 생성 작업에서 좋은 성능을 나타낸다. 50에서 100 사이인 경우, 모델이 텍스트를 예측할 수 있지만, 불확실성이 크다는 의미이다.

표 9는 Perplexity를 측정하기 위해 입력한 문장 10개이다.

Index	Perplexity	
	Blossom	Untrained
1	11.78	7.85
2	22.86	14.13
3	34.13	25.72
4	15.44	12.19
5	35.80	21.25
6	17.01	11.38
7	11.53	6.81
8	12.40	18.29
9	6.23	5.18
10	21.04	10.37

표 10. 모델별 Perplexity

표 10은 Blossom 모델과 학습이 안 된 Llama3 모델(Untrained)의 Perplexity 비교 결과이다. 흥미로운 점은 학습이 안 된 Llama3 모델의 문장이 훨씬 자연스럽고 성능이 좋았다. 이는 예상과는 다른 결과로, 한 가지 가능성은 Blossom 모델이 특정한 데이터셋에 과도하게 적합되어 일반화 성능이 떨어졌을 수 있다는 것이다. 반면, 학습이 안 된 Llama3 모델은 특정한 패턴이나 데이터에 종속되지 않아, 오히려 더 다양한 문장을 자연스럽게 처리할 수 있을 가능성이 있다.

이러한 결과는 모델의 학습 과정에서 데이터 선택이나 학습 방식에 따라 성능이 달라질 수 있음을 시사하며, 보다 정교한 평가 및 분석이 필요하다.

3.7 뉴스 요약본 기반 군집화

이 절에서는 3.5절의 뉴스 요약본을 기반으로 KR-BERT 모델을 이용하여 임베딩 벡터를 구하여 군집화 실험을 하고, 뉴스 주제별 유사도를 측정하였다. 표 11은 뉴스 주제별 수집한 데이터를 보여준다.

	economy	politics	interna.	science	social	culture
Article 수	340	325	352	356	338	307

표 11. 뉴스 주제별 Article 수

총 6개의 Article을 임베딩 벡터로 변환 후, 군집화하여 시각화하고자 한다.

요약본의 임베딩 벡터는 768차원으로 구성된다. 이 임베딩 벡터를 PCA를 이용하여 2차원으로 변환 후 시각화한다. 군집화는 k-means 알고리즘을 사용했다.

3.7.1 평가 결과

표 7에서의 모든 데이터를 이용하여 군집화를 진행했다. 결과는 그림 5와 같다. 군집화 결과에서 특히 흥미로운 점은, 네이버 뉴스의 science 주제가 다른 주제들과의 연관성이 현저히 떨어진다는 것이다. 이는 군집화에서 science 주제가 다른 군집들과 명확하게 분리되는 경향으로 나타났다.

이러한 차이를 수치적으로 확인하기 위해, science 주제와 다른 주제들 간의 코사인 유사도를 측정하였다. 그림 6은 각 주제별로 임의로 추출한 10개의 임베딩 벡터를 활용해 코사인 유사도를 계산한 후, 이를 히트맵으로 시각화한 결과를 보여준다. 그림에서 확인할 수 있듯이, science 주제와 다른 주제들 간의 코사인 유사도가 다른 조합들에 비해 확연히 낮게 나타났다. 이는 science 주제가 다른 뉴스 주제들과 상대적으로 유사성이 적다는 것을 의미하며, 해당 주제의 내용적 특수성이나 표현 방식의 차이가 반영된 결과로 해석될 수 있다.

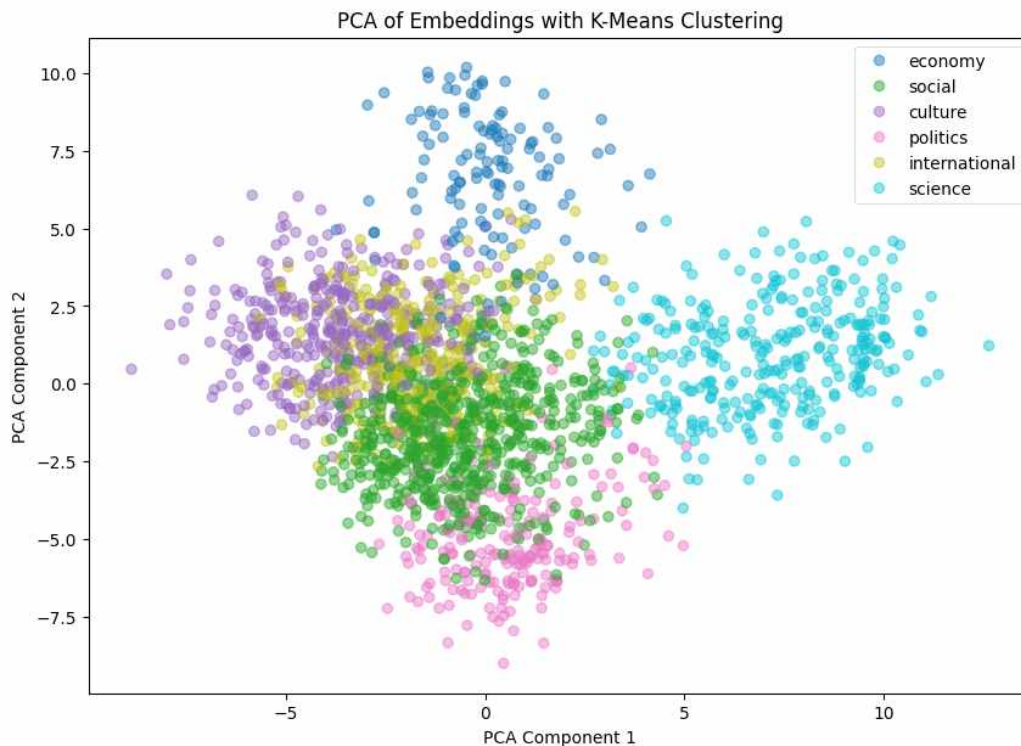


그림 4. 뉴스 요약본 기반 PCA를 이용한 K-Means Clustering

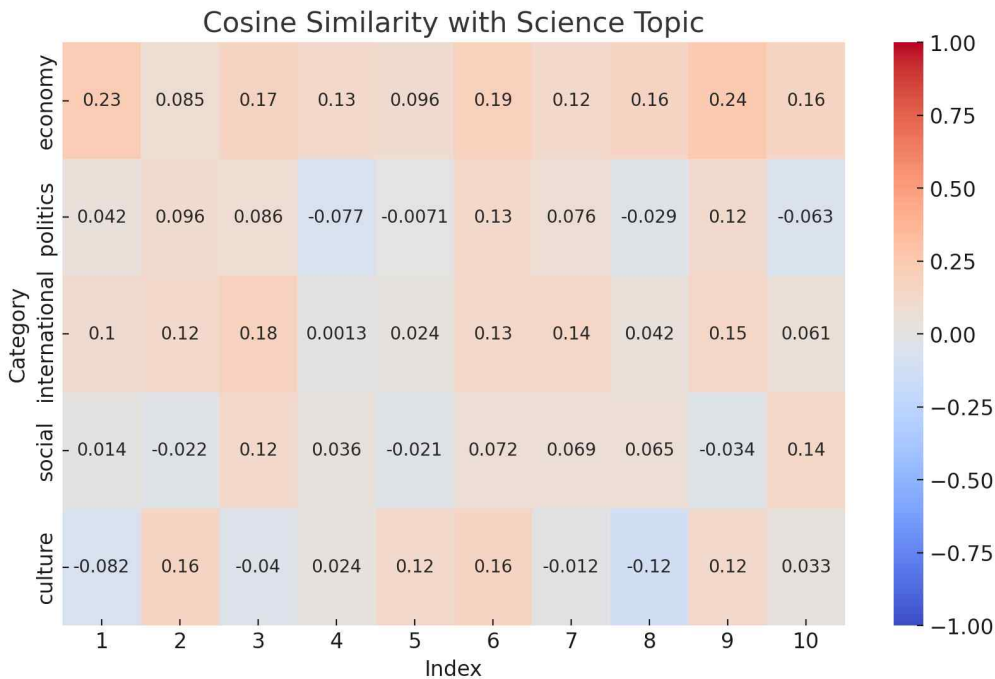


그림 5. Science Topic과 다른 Topic과의 Cosine Similarity

3.8 같은 뉴스 topic 내에서의 코사인 유사도

Topic이 같은 뉴스 내에서 요약본의 임베딩 벡터 간의 유사도를 비교할 수 있다. 같은 토픽 내에서 뉴스 요약본의 임베딩 벡터 간의 코사인 유사도를 비교함으로써 해당 토픽 내 문서들이 얼마나 일관성 있게 다루어지는지 확인할 수 있다. 예를 들어, 정치 토픽에서 유사도가 높은 문서 쌍이 많다면, 이는 해당 토픽의 뉴스들이 비슷한 내용이나 관점을 공유하고 있음을 시사한다. 반대로 유사도가 낮다면, 정치 뉴스가 다양한 측면을 다루고 있거나 서로 다른 사건들을 포함하고 있을 수 있다.

이러한 분석은 뉴스 소비자나 언론사에게 유용한 정보를 제공할 수 있다. 소비자의 관점에서, 특정 토픽에 대한 뉴스가 얼마나 다양하고 깊이 있게 다루어지는지 알 수 있어 균형 잡힌 시각을 얻는 데 도움이 된다. 언론사 입장에서는 자사의 보도 경향을 파악하고, 필요에 따라 특정 주제에 대한 보도 전략을 조정하는 데 활용할 수 있다. 예를 들어, 한쪽으로 치우친 보도를 줄이고 다양한 관점을 제공하고자 할 때 유용하다.

그림 7은 같은 뉴스 topic 내에서의 코사인 유사도를 히트맵으로 시각화한 것이며, 그림 8은 각 뉴스 topic의 코사인 유사도의 분포를 나타낸 것이다. 이때, 이상치는 제외하였다.

정치(Politics) 분야에서 코사인 유사도가 높게 나타났다. 이는 같은 주제 내에서 여러 언론사들이 보도하는 내용이 상당히 일관되거나 유사한 경향을 보이고 있음을 의미한다. 즉, 정치적 사건이나 이슈에 대한 분석 및 해석이 유사

한 방식으로 이루어지고 있거나, 동일한 정보원에 기반해 보도가 이루어지고 있을 가능성이 크다.

Cosine Similarity Matrices for Different News Topics

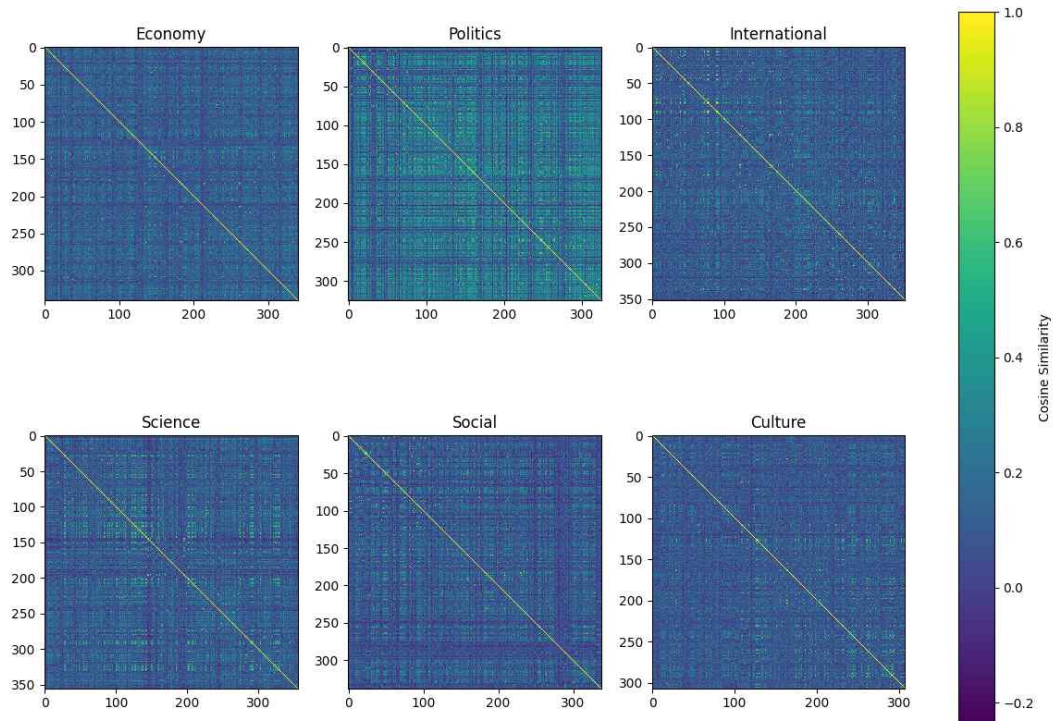


그림 6. 같은 뉴스 Topic 내에서의 코사인 유사도 히트맵

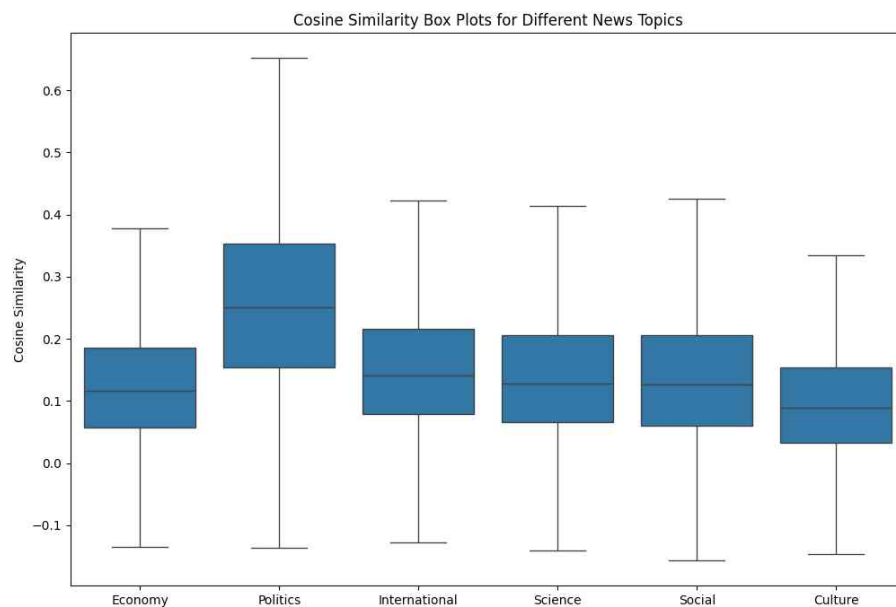


그림 7. 같은 뉴스 Topic 내에서의 코사인 유사도 분포 (이상치는 제외)

3.9 뉴스 Topic 추천 및 워드클라우드 생성

3.7절에서의 각 뉴스 Topic마다 embedded vector의 평균을 구하고, 사용자가 멀티턴 대화가 학습된 Llama3 모델과 대화를 나누며 입력한 문장, 문장 속의 감정을 이용하여 뉴스를 추천하고자 한다. 추천하는 방법은 각 뉴스 Topic 별 평균 embedded vector과 코사인 유사도를 구한 후, 가장 유사도가 높은 Topic을 추천한다.

3.9.1 감정 문장 군집화

그림 9는 3.1절의 7가지 감정 레이블이 있는 문장들을 embedded vector로 변환하고 pca로 차원 축소 후 k-means 군집화를 한 것이다.

각 감정은 전반적으로 잘 구분된다. '행복(happiness)', '공포(fear)', '놀람(surprise)' 등 대부분의 감정은 특정한 영역에 집중되어 있어 시각적으로도 명확히 구분된다. 그러나 일부 감정은 분포 상에서 서로 겹치는 경향을 보인다. 특히 '혐오(disgust)'와 '분노(angry)' 감정은 색깔상 분홍색과 초록색으로 구분되지만, 그 분포가 서로 상당 부분 겹쳐 있다. 이는 혐오와 분노라는 두 감정이 의미적으로나 표현적으로도 유사성이 강하다는 것을 반영할 수 있다. 두 감정 모두 강한 부정적 감정을 내포하고 있어, 문장 내에서 그 표현이 서로 혼재될 가능성이 크다.

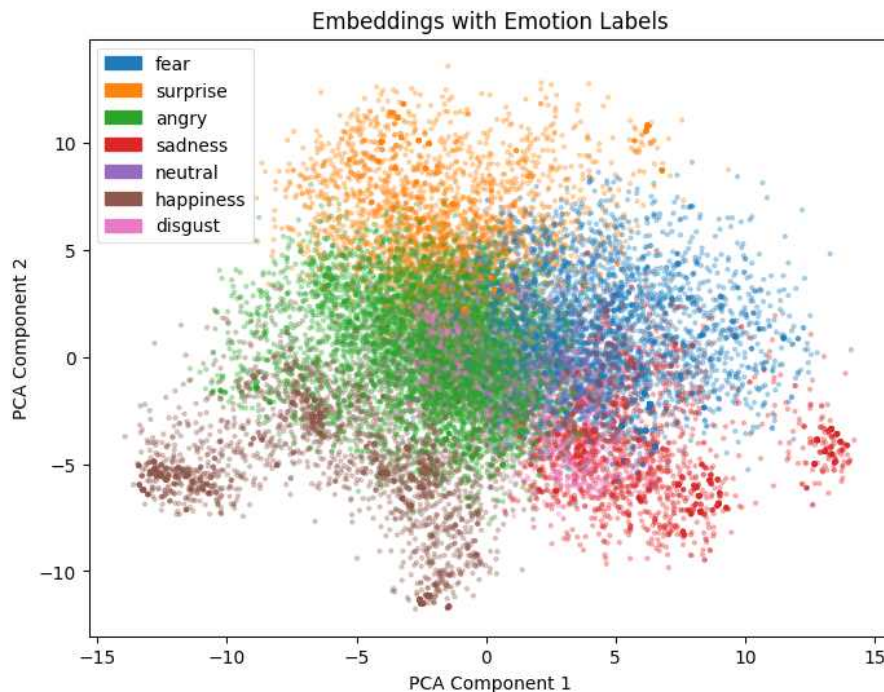


그림 8. 감정 레이블이 있는 문장의 K-Means Clustering

이러한 겹침은 감정 간의 미세한 차이를 구분하는 데 있어 모델의 한계일 수 있지만, 동시에 자연 언어에서 감정 표현이 명확히 구분되지 않을 때도 있

다는 것을 보여준다.

3.9.2 감정과 뉴스 Topic의 코사인 유사도

그림 10은 3.7절에서 뉴스 Topic의 평균 임베딩 벡터와 3.9.1절에서 설명된 7가지 감정의 평균 임베딩 벡터 간의 코사인 유사도를 나타낸 것이다. 각 셀은 특정 주제(예: 경제, 정치, 국제)와 특정 감정(예: 두려움, 놀라움, 슬픔) 사이의 코사인 유사도를 나타낸다.

밝은 노란색일수록 유사도가 높은 것을 의미한다. 예를 들어, "경제"와 "슬픔" 사이의 유사도(0.38), "사회"와 "중립" 사이의 유사도(0.37), "문화"와 "놀라움" 사이의 유사도(0.32)가 특히 높게 나타난다. 이는 특정 감정이 특정 뉴스 주제와 밀접하게 연관되어 있음을 보여주며, 각 뉴스 주제가 주로 어떤 감정적 반응을 유발할 가능성이 높은지를 시사한다.

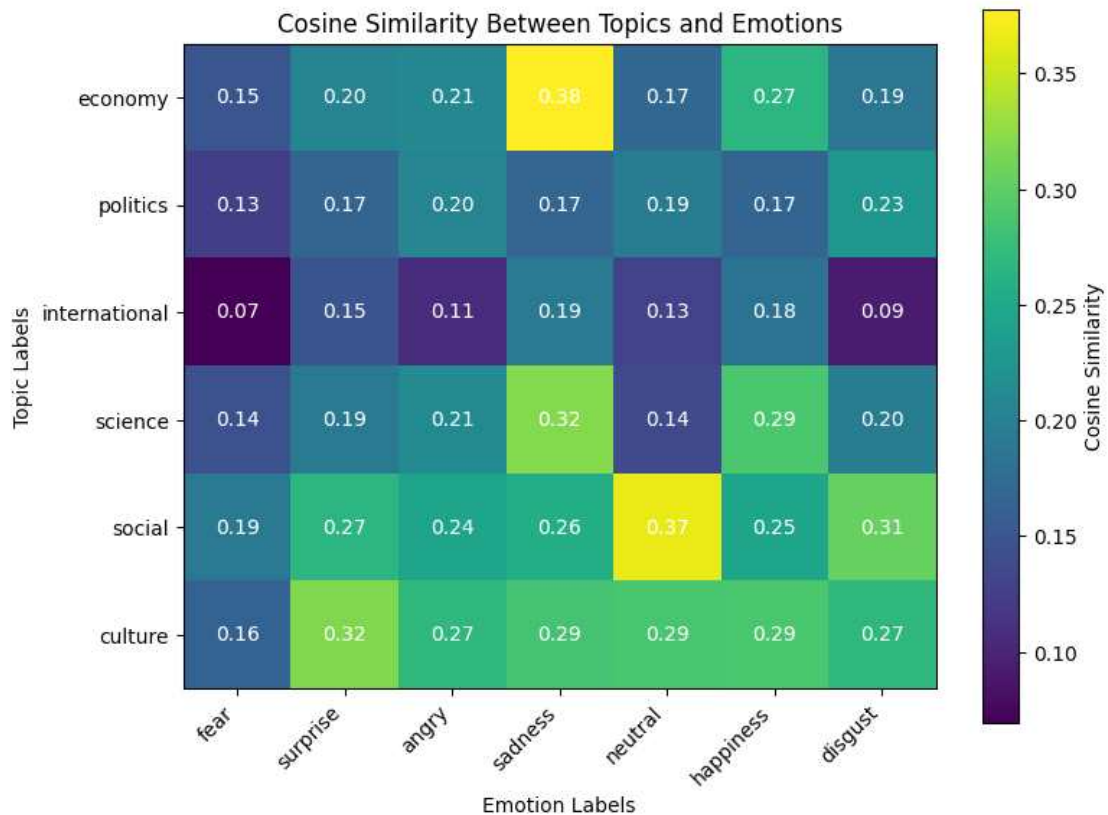


그림 9. 뉴스 Topic과 감정의 코사인 유사도

예를 들어, 경제 관련 뉴스가 슬픔과 높은 유사도를 보이는 것은 경제적 불안이나 불평등과 같은 주제들이 많다는 점을 시사할 수 있다. 또한 사회 관련 뉴스가 중립적 감정과 연결되는 것은 사회 관련 뉴스 특성상 다양한 관점을 다루기 때문이라고 해석될 수 있다. 반면에, 문화적 주제와 놀라움 간의 높은

유사도는 뉴스가 예상치 못한 정보나 새로운 경험에 주목할 가능성이 크다는 것을 암시한다.

3.9.3 뉴스 문장에서의 키워드 추출 및 텍스트 클라우드 생성

기존의 텍스트 클라우드가 단어의 빈도수에만 치중하는 반면, 우리의 연구에서는 사용자가 미세 조정된 Llama3 모델과 대화를 하며 입력한 문장의 평균 벡터들과, 그에 알맞는 감정의 평균 벡터를 이용하여 뉴스 문장에서의 단어들과 코사인 유사도가 높은 단어들을 중심으로 시각화하고자 한다. 아래 수식 (1)은 사용자 입력 문장 평균 벡터와, 사용자의 감정 평균 벡터를 이용한 수식이다. \vec{V} 는 뉴스 내에서의 문장의 단어들과 코사인 유사도를 구하기 위한 최종적인 벡터이다.

$$(1) \vec{V} = k \times \overrightarrow{sentence} + (1 - k) \times \overrightarrow{sentiment}$$

실험에서 k 값을 0.3, 0.5, 0.7로 조정하며 실험해 보았다. 이때, 감정은 중립(neutral)이며, 추천한 뉴스 Topic은 사회(Social)이다. 그림 11은 미세 조정된 Llama3와 사용자의 대화를 보여준다.

HUMAN
안녕 넌 누구니?

AI
저는 AI입니다. 당신의 질문에 답하고자 합니다

HUMAN
오늘의 날씨는 어때?

AI
오늘은 맑고 화창한 날씨입니다. 기온은 20도이고, 습도는 60%입니다.

HUMAN
부산대학교에 대해서 알고 있어?

AI
네, 부산대학교는 대한민국의 국립 종합대학으로, 1946년에 설립되었습니다. 현재 21개 단과대학과 1개 대학원으로 구성되어 있습니다. 부산광역시 금정구 장전동에 위치하고 있으며, 학생 수는 약 3만 명입니다.

그림 10. 미세 조정된 Llama3와의 대화



그림 11. k 값이 0.3일 때 텍스트 클라우드



그림 12. k 값이 0.5일 때 텍스트 클라우드



그림 13. k 값이 0.7일 때 텍스트 클라우드

k	추천된 뉴스의 제목
0.3	답답하다 정말. 선배들 때는 수사하기 쉬웠죠"[조준영의 검찰聽]
0.5	답답하다 정말. 선배들 때는 수사하기 쉬웠죠"[조준영의 검찰聽]
0.7	아내 나가면 전화줘요"..유부남과 불륜 과거 있는 여친과 헤어져야 하나?

표 12. k 값에 따른 추천 뉴스의 제목

그림 11, 12, 13은 각각 k값이 0.3, 0.5, 0.7일 때, 나타나는 텍스트 클라우드이며, 표 12는 k 값에 따른 추천된 뉴스이다.

4. 결론 및 추후 연구 과제

우리는 Llama3 모델 미세 조정과, 감정, 뉴스 Topic에 대해 Llama3 모델과 사용자간의 대화를 통해 사용자의 발화와 사용자의 감정을 추출하고, 이에 따라 뉴스 Topic과 해당하는 Topic 내에서의 뉴스를 추천하는 연구를 하였다. 주요 결과들은 다음과 같다.

- Embedded Vector은 Llama3 모델보다는, Embedded Vector을 추출하는데

특화된 모델을 쓰는 것이 정확도가 더 높다.

- 사용자 감정 분류 또한 Llama3 모델보다는, 감정 분류에 특화된 모델을 쓰는 것이 정확도가 높다.
- MBTI의 경우, 단순한 대화 데이터만으로 이 복잡한 특성을 완전히 학습하기는 어렵다. 특히, 짧거나 맥락이 부족한 대화의 경우, 모델이 해당 대화를 특정 MBTI 유형에 맞게 분류하기가 매우 어려울 수 있다.
- 한국어 데이터셋이 미리 학습되어 있는 모델보다는, 한국어 데이터셋이 학습되어 있지 않은 모델에서 더 좋은 수치적인 결과가 나온다.
- 뉴스 Topic 중, 사회(social) 분야는 국제(international), 정치(politics)와 분포가 많이 겹친다. 하지만 과학(science) 분야는 다른 분야들과 분포가 제일 적게 겹친다.
- 같은 뉴스 Topic 중, 정치(politics) 분야가 기사 간 유사도가 가장 높다. 이는 뉴스들이 일관적인 보도를 한다는 것을 시사한다.
- 화남(angry) 감정은 혐오(disgust) 감정과 유사도가 매우 높다. 두 감정 모두 강한 부정적 감정을 내포하고 있어, 문장 내에서 그 표현이 서로 혼재될 가능성이 크다.

본 결과와 관련하여 추후 연구 과제는 다음과 같다.

- 감정과 대화 내역을 기반으로 사용자의 뉴스 선호도를 더욱 정교하게 반영하는 알고리즘을 연구할 필요가 있다. 개인화 추천 시스템의 성능 향상을 목표로, 다양한 추천 알고리즘을 실험해볼 수 있다.
- 다양한 모델들(Llama3, 감정 분석 모델, 임베딩 특화 모델 등)을 기반으로 한 미세 조정 실험을 통해, 어떤 모델이 사용자의 대화 데이터와 감정 데이터에 가장 적합한지 비교하는 연구 또한 필요하다.
- MBTI와 대화 패턴 간의 상관관계를 더 깊이 탐구하기 위해, 더 많은 데이터와 복잡한 대화 시나리오를 활용한 연구가 필요하다. 특히, 장기 대화 데이터를 통해 MBTI 유형과 대화 방식 간의 상관관계를 도출하는 것이 유의미할 것이다.
- 뉴스 기사 간 유사도 분석을 확장하여, 시간에 따른 유사도의 변화나 다른 뉴스 카테고리 간의 관계를 분석할 수 있다. 또한, 새로운 주제가 등장할 때 유사도가 어떻게 변하는지를 추적하는 것도 흥미로운 연구 주제가 될 수 있다.

5. 참고 문헌

[1] DUBEY, Abhimanyu, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

[2] SASSE, Laura K., et al. Selective control of attention supports the positivity effect in aging. PloS one, 2014, 9.8: e104180.

[3] WILLIAMS, Paige; KERN, Margaret L.; WATERS, Lea. Exploring selective exposure and confirmation bias as processes underlying employee work happiness: An intervention study. Frontiers in Psychology, 2016, 7: 878.

[4] VENETIS, Petros; KOUTRIKA, Georgia; GARCIA-MOLINA, Hector. On the selection of tags for tag clouds. In: Proceedings of the fourth ACM international conference on Web search and data mining. 2011. p. 835-844.

[5] KASER, Owen; LEMIRE, Daniel. Tag-cloud drawing: Algorithms for cloud visualization. arXiv preprint cs/0703109, 2007.

[6] LEE, S. Kr-bert: A small-scale korean-specific language model. arXiv preprint arXiv:2008.03979, 2020.

[7] PARK, Sungjoon. KLUE: Korean Language Understanding Evaluation. arXiv preprint arXiv:2105.09680, 2021.

[8] PAPINENI, Kishore, et al. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002. p. 311-318.

[9] LIN, Chin-Yew. Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. 2004. p. 74-81.

[10] CHOI, ChangSu, et al. Optimizing Language Augmentation for Multilingual Large Language Models: A Case Study on Korean. arXiv preprint arXiv:2403.10882, 2024.

6. 부록

“system” : 정확한 챗봇으로서 상대방의 입력에 대해 감정을 맞추자. 모든 대답은 '행복', '분노', '슬픔', '중립', '혐오', '놀람', '공포' 중 하나로 대답해줘.

“user” : { }

“assistant” : { }

표 13. Llama3 감정 분류 프롬프트

<p>“system” : 공감하는 챗봇으로서 상대방의 입력에 대해 공감을 하자. 모든 대답은 한국어 (Korean)으로 대답해줘.</p> <p>“user” : { }</p> <p>“assistant” : { }</p>

표 14. Llama3 공감 대화 프롬프트 (감정 인식 제외)

<p>“system” : 정확한 챗봇으로서 상대방의 입력에 대해 MBTI를 맞추자. 모든 대답은 MBTI 16개중 하나로 대답해줘.</p> <p>“user” : { }</p> <p>“assistant” : { }</p>
--

표 15. Llama3 MBTI 분류 프롬프트

<p>“system” : 정확한 챗봇으로서 상대방의 입력에 대해 요약을 하자. 모든 대답은 한국어 (Korean)으로 대답해줘.</p> <p>“user” : { }</p> <p>“assistant” : { }</p>
--

표 16. Llama3 요약 생성 프롬프트

<p>“system” : You are a helpful AI assistant. Please answer the user's questions kindly. 당신은 유능한 AI 어시스턴트 입니다. 사용자의 질문에 대해 친절하게 답변해주세요.</p> <p>“user” : { }</p> <p>“assistant” : { }</p>
--

표 17. Llama3 멀티턴 학습 프롬프트