

국립국어원 2021-01-32

발 간 등 록 번 호

11-1371028-000868-01

2021년 말뭉치 감성 분석 및 연구

사업 책임자
이 영 희



국립국어원

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 '2021년 말뭉치 감성 분석 및 연구'에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업기간: 2021년 8월 ~ 2021년 12월

2021년 12월 6일

사업 책임자: 이 영 희((주)버즈메트릭스)

사업 수행자 (주)버즈메트릭스

사업 책임자 이영희

사업 참여자 김수진, 정종서, 신현주, 유지현,
박나경, 추승아, 옥승희, 서은미,
김도현, 이진상

<사업 수행자>

(주)버즈메트릭스

사업 책임자	이영희(주)버즈메트릭스
사업 참여자	김수진(주)버즈메트릭스
	정종서(주)버즈메트릭스
	신현주(주)버즈메트릭스
	유지현(주)버즈메트릭스
	박나경(주)버즈메트릭스
	추승아(주)버즈메트릭스
	옥승희(주)버즈메트릭스
	서은미(주)버즈메트릭스
	김도현(주)버즈메트릭스
	이진상(주)버즈메트릭스

2021년 말뭉치 감성 분석 및 연구

본 사업은 4차 산업혁명을 대비하여 인공지능 기술의 개발 및 활용을 위한 대규모 말뭉치 데이터를 구축하고 향후 국립국어원의 감성 및 감성 분석 말뭉치 구축 사업의 방향성을 제시하는 데 목적이 있다. 우선 2020년도 ‘말뭉치 감성 분석 및 연구’의 후속 사업으로서 기존 감성 분석 말뭉치의 미흡한 점을 보완하고 부착된 감성 데이터의 정밀성을 높이고자 하였다. 아울러 국내외 감성 및 감성 분석 말뭉치 관련 현황을 면밀히 살펴 학계 및 산업계 활용에 용이한 데이터 구축 기준을 수립하는 것에 중점을 두었다. 이에 따른 주요 과업과 연구 성과를 요약하면 다음과 같다.

첫째, 한국어의 특성을 반영한 속성 기반 감성 분석(Asspect based sentiment analysis)의 방법론을 제시하고 세부적인 지침을 수립하였다. 분석 대상 말뭉치에 적합한 도메인을 분류하였으며 각 도메인에 해당하는 개체 및 속성 라벨을 구성하여 보다 정밀한 감성 분석이 가능하도록 하였다.

둘째, 지침에 따른 분석을 위해 기존 20년도 감성 분석 저작 도구를 수정·보완하여 감성 분석 말뭉치를 재분석하였다. 또한 주요 도메인에 대한 테스트 문서 데이터를 추가로 수집하여 확보하였으며, 추가 데이터 역시 동일한 분석 과정을 거쳐 AI 성능 검증을 위한 테스트 데이터를 구축하였다.

셋째, 국내 및 국외의 학계·산업계 감성·감정 분석 말뭉치 구축 및 활용 현황을 조사하여 동향을 살폈다. 또한 AI, NLP(자연어 처리), 데이터 마이닝 관련 학계 및 산업계 전문가 인터뷰를 통해 각 분야의 동향을 파악하고 차후 국어원의 감성 분석 말뭉치 구축 사업이 나아갈 방향에 대하여 조언을 구했다.

넷째, 향후 국립국어원의 감성 분석(Emotion Detection) 말뭉치 구축 사업의 방향을 제시하였다. 국내외 동향 조사와 전문가 인터뷰를 통해 학계 및 산업계 내 활용에 실질적으로 유효한 데이터 자료 유형, 규모, 형식 등을 조사하였고 이에 따른 향후 감성·감정 분석 말뭉치 구축 사업의 구체적인 기준을 제안하였다.

주요어: 감성 분석, 감정 분석, 감성 분석 말뭉치 구축, 감정 분석 말뭉치 활용, 속성 기반 감성 분석

차례

제1장 서론

1. 사업 목적	3
2. 사업 수행 범위	3
3. 사업 수행 절차	4
3.1. 말뭉치 감성 분석 지침 수립 및 말뭉치 구축	4
3.2. 국내외 동향 연구 및 제안	5

제2장 말뭉치 감성 분석

1. 문장 분할 작업 및 도메인 선정	9
1.1. 문장 분할 작업	9
1.2. 세부 도메인 선정	11
2. 감성 분석 지침 수립	13
2.1. 감성 분석의 대상	13
2.2. 속성 기반 감성 분석 정보	13
2.3. 분석 과정	17
3. 추가 데이터 수집	19
3.1. 테스트 데이터 구축 과정	19
3.2. 테스트 데이터 구축 결과	20
4. 저작 도구 구축	21
4.1. 저작 도구 구축 방향	21
4.2. 저작 도구 구현 기능	21
5. 말뭉치 분석 데이터	23
5.1. 분석 데이터 정보	23
5.2. 데이터 납품 형태	23

차 례

제3장 말뭉치 감성 분석 활용 및 동향 연구

1. 말뭉치 동향 연구	29
1.1. 국내외 감성 및 감정 분석 말뭉치 구축 현황	29
1.2. 국내외 감성 및 감정 분석 말뭉치 활용 현황	45
2. 국어원 구축 말뭉치 관련 제언	56
2.1. 한국어 감정 분석 말뭉치 구축 시뮬레이션	56
2.2. 국어원 감정 분석 말뭉치 구축 방향성 제언	58

제4장 결론

1. 사업 요약	69
1.1. 말뭉치 감성 분석 지침 수립 및 말뭉치 구축	69
1.2. 국내외 동향 연구 및 제언	70
2. 사업의 의의 및 기대 효과	73

<부록1> 말뭉치 감성 분석 지침

<부록2> 국내외 감성·감정 분석 현황 자료

<부록3> 감정 분석 말뭉치 구축 시뮬레이션

참고 문헌

<Astract>

표 차례

표 1 종결 표현 기준 문장 분할 작업 예시	11
표 2 주제별 문서 및 문장 수 통계	12
표 3 도메인 최종 구성	14
표 4 도메인별 개체 유형 및 속성 라벨	14
표 5 분석 예시 문장	18
표 6 버즈룸 주요 기능 변화	21
표 7 Json 파일 내 정보값	23
표 8 말뭉치 감성 분석 데이터 Json 형식 납품 데이터 예시 (Text level)	24
표 9 말뭉치 감성 분석 데이터 Json 형식 납품 데이터 예시 (Sentence level)	25
표 10 IMDb 데이터 예시	30
표 11 SST-2 데이터 예시	30
표 12 SST-5 데이터 예시	31
표 13 SemEval-2016 Task 4 데이터 예시 (Subtask A)	31
표 14 SemEval-2016 Task 5 데이터 예시	32
표 15 EmoInt 데이터 예시	33
표 16 SemEval 2018 Task 1 E-c 데이터 예시 (an: anger, ant:anticipation, dis: disgust, fe: fear, lo: love, opt: optimism, pes: pessimism, sad: sadness, sur: surprise, tr: trust)	35
표 17 SemEval 2019 Task 3 데이터 예시	35

표 차례

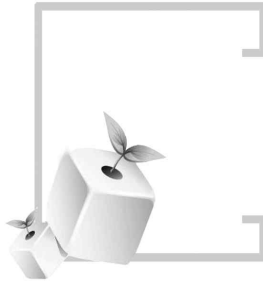
표 18	감성 분석 말뭉치 주제별 분포	36
표 19	감성 분석 말뭉치 예시	37
표 20	NSMC 데이터 예시	39
표 21	한국어 감정 정보가 포함된 단발성 대화 데이터셋 예시	40
표 22	연속성 대화 데이터셋 예시	41
표 23	웰니스 대화 스크립트의 감정 분류	45
표 24	웰니스 대화 스크립트 예시	45
표 25	한국어 감정 분석 데이터 구축 시 고려 사항	59
표 26	감성-감정 연계 모델	62
표 27	감성-감정 연계 모델을 적용한 예시	62
표 28	플루칙의 1, 2, 3차 이중 감정 예시	63
표 29	한국어 감정 분석 데이터 구축 방안 제언	66
표 30	국내외 감성 및 감정 분석 말뭉치 주요 구축 현황	70
표 31	국내외 감성 및 감정 분석 말뭉치 주요 활용 현황	71

그림 차례

그림 1	감성 분석 세부 지침 수립 및 감성 분석 말뭉치 구축 절차	4
그림 2	국내외 동향 연구 및 제안 절차	5
그림 3	블로그 게시글 문장 분할 작업 예시	9
그림 4	누리소통망 게시글 문장 분할 작업 예시	10
그림 5	의견 극성에 따른 문장 분류 예시(영화 도메인)	16
그림 6	문장 감성 분석 과정	17
그림 7	테스트 데이터 확보 과정	19
그림 8	테스트 데이터 감성 분석 과정	20
그림 9	2021 말뭉치 속성 기반 감성 분석 룬	22
그림 10	감정 분류를 위한 음성 데이터셋 예시	42
그림 11	감성 대화 말뭉치의 감정 분류	43
그림 12	감성 대화 말뭉치 예시	44
그림 13	Empath에서 제공하는 감정 분석기(좌), 감정 대화 인공지능 Beluga Box(우)	46
그림 14	아마존의 인공지능 스피커 ‘에코’	47
그림 15	Huang et al. (2019)에서 제안한 모델 아키텍처	48
그림 16	Lai et al. (2020)의 데이터 정보	49
그림 17	GoEmotions에 부착된 감정	49
그림 18	Wang et al. (2021)에서 사용한 Affective Text 데이터 예시	50
그림 19	하이퍼클로바의 기능이 적용된 쇼핑 리뷰 예시	51
그림 20	클로바センチメント API 작동 방식	51

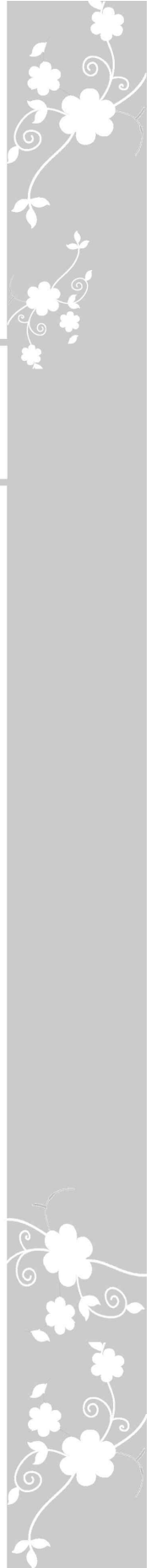
그림 차례

그림 21 야구 댓글에서의 구어체 현상 및 감성 분석	52
그림 22 화장품 A에 대한 속성별 감성 점수	53
그림 23 사용자의 감정을 분석하는 챗봇 ‘티티’	53
그림 24 안세훈 외(2021)의 감성 분류 정확도	55
그림 25 안세훈 외(2021)의 감정 원인 분류 정확도	55
그림 26 플루칙의 감정 바퀴(좌), 플루칙의 1, 2, 3차 이중 감정(우)	63
그림 27 2021 감성 분석 말뭉치 구축 결과	69



제 1 장

서 론



1. 사업 목적

본 사업의 목적은 4차 산업 혁명에 대비하여 인공지능 기술의 개발 및 활용을 위한 대규모 말뭉치를 분석하여 국어 자원의 활용도와 가치를 제고하는 것에 있다. 또한 국내 표준화 및 참고 지반 자료가 될 수 있는 정밀한 언어 정보 부착 말뭉치를 분석 및 배포를 목적으로 한다. 본 사업의 추진 목표는 아래와 같다.

□ 말뭉치 감성 분석 추진 목표

- 언어 데이터 기반 인공지능 기술에 적용할 수 있는 정밀 감성 분석 말뭉치 구축
- 일반 기업이 활용할 수 있는 공공재로서의 텍스트 파일(json 형식) 말뭉치 구축
- 한국어 텍스트 대상 속성 기반 감성 분석 지침 기반 마련

2. 사업 수행 범위

본 사업의 주요 내용은 감성 분석 말뭉치 구축과 말뭉치 활용 및 동향 연구로 구성된다. 구체적인 사업 수행 범위는 다음과 같다.

□ 말뭉치 감성 분석 지침 수립 및 말뭉치 구축

○ 속성 기반 감성 분석 지침 수립

- ABSA-2016을 기반으로 한국어 특성을 반영한 속성 기반 감성 분석 방법론 수립
- 12개 도메인별 개체 및 속성 라벨을 구성하여 분석 방법론 구체화
- OTE(Opinion Target Expression) 설정 및 추출을 위한 관련 지침 마련

○ 테스트 데이터 추가 구축

- 주요 도메인 영화, 화장품·세정제, 전자기기에 해당하는 테스트 데이터 추가 구축
- 참여자 모집·데이터 적합성 평가를 거쳐 210개 유효 문서 확보

○ 속성 기반 감성 분석 말뭉치 구축

- 분석 대상 말뭉치: 문서 2,081건 규모의 웹 말뭉치,
(블로그·게시판 273건, 누리소통망 1,808건)
추가 구축 데이터
(블로그 210건)
- 분석 내용: 속성 기반 감성 분석 지침에 따른 감성 분석 정보 부착

- 속성 기반 감성 분석 결과 말뭉치 구축 결과 Json 형태 출력

□ 국내외 동향 연구 및 제안

○ 문헌 연구, 학계·산업계 전문가 인터뷰를 통한 현황 및 전망 조사

- 국내외 감성 및 감정 분석 말뭉치 구축 및 활용 동향 조사
- 국내 산업 수요 및 전망 조사

○ 국어원 말뭉치 구축 방안 제안

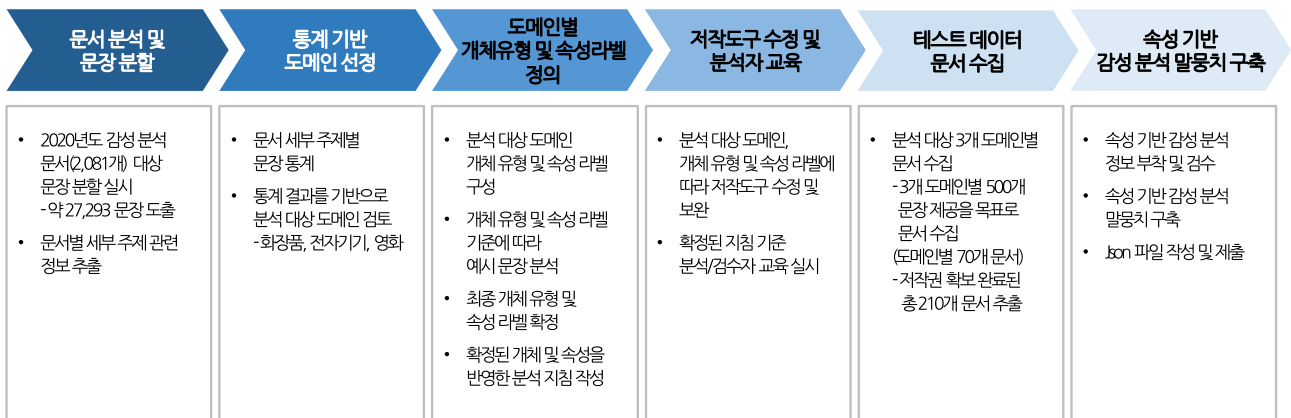
- 국어원에서 수집할 자료의 유형, 구축할 말뭉치의 유형 및 규모, 형식, 품질 평가 방안, 활용 방안 제안

3. 사업 수행 절차

3.1. 말뭉치 감성 분석 지침 수립 및 말뭉치 구축

감성 분석 세부 지침 수립 및 감성 분석 말뭉치 구축은 총 6단계 수행 절차에 따라 진행되었다. 문서 분석 및 문장 분할을 기초로 주요 도메인을 선정하고, 주요 도메인 관련 테스트 데이터를 추가 구축하여 기본 문서와 테스트 데이터를 대상으로 속성 기반 감성 분석 말뭉치를 구축하는 절차로 진행되었다.

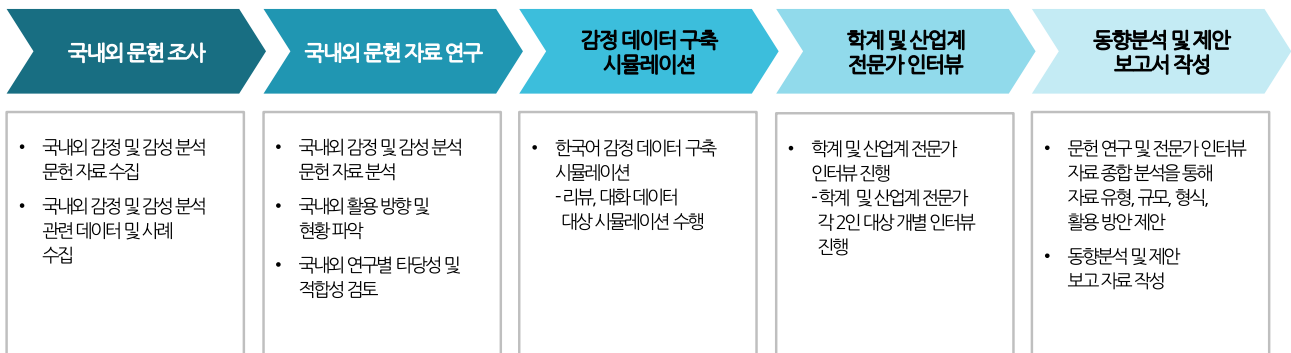
그림 1 감성 분석 세부 지침 수립 및 감성 분석 말뭉치 구축 절차

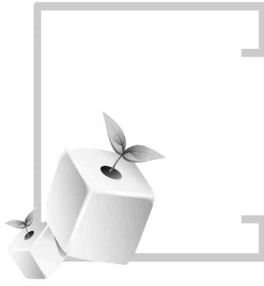


3.2. 국내외 동향 연구 및 제안

국내외 동향조사 및 향후 말뭉치 유형과 규모 제안을 위해 국내외 감정 및 감성 분석에 대한 문헌 자료 연구를 진행하였다. 문헌 자료를 토대로 국내외 감정 및 감성 분석 현황을 파악하고, 실제 감정 데이터 구축 시뮬레이션 진행을 통해 제안을 위한 실효성을 검토하였다. 또한, 학계 및 산업계 전문가 각 2인을 대상으로 개별 인터뷰를 진행하여 현재 학계와 산업계에서 접근하고 있는 감정 및 감성 분석 현황과 활용 방향에 대한 의견을 조사하였다. 문헌 조사와 전문가 인터뷰, 시뮬레이션 결과를 토대로 국어원에서 구축할 자료의 유형과 규모, 형식, 활용 방안을 위한 제언을 수행하였다.

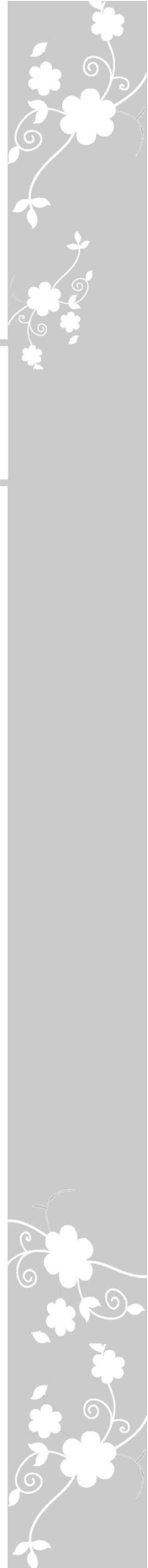
그림 2 국내외 동향 연구 및 제안 절차





제 2 장

말뭉치 감성 분석



1. 문장 분할 작업 및 도메인 선정

1.1. 문장 분할 작업

1.1.1. 문장 분할 작업 개요

본 연구에서는 본격적인 분석 작업을 하기에 앞서 말뭉치 데이터의 정제 작업을 진행하였다. 말뭉치 감성 분석은 문장 단위로 감성 분석 정보를 부착하는 작업이다. 그런데 분석 대상 말뭉치 문서는 웹 게시물의 특성상 문장들이 무분별하게 통합되거나 분절된 경우가 많아 정확한 문장 단위 분석 작업에 어려움이 있었다. 이 때문에 일관된 기준으로 문장 단위 분석을 하기 위해서는 분석 전 문서 전반에 걸친 정제 작업이 필요하였다. 이에 본 연구에서는 문장 분할 작업을 통해 행 단위로 단일 문장이 나타날 수 있도록 하는 전처리 작업을 진행하였다.

블로그 게시글의 경우 주로 줄글이 문단 단위로 무분별하게 결합되어 있는 오류가 많았다. 문장이 끝났음에도 불구하고 줄바꿈 처리를 하지 않아 문장의 적절한 분할이 이루어지지 않은 유형이다. 본 연구에서는 문장 분할 작업을 통해 줄바꿈 처리를 하여 문장을 자연스럽게 분할하였으며 그 예시는 아래와 같다.

그림 3 블로그 게시글 문장 분할 작업 예시

문서 제목	데쓰 프루프 (Death Proof, 2007)
<p>어쨌거나 영화 시작. 타란티노 특유의 장난질이 시작되었다. 화면에 지직거리는 노이즈, 중간중간 끊긴 듯 넘어가는 컷이나 대사가 튀는 것처럼 2번 반복된다던지. 마치 6~70년대 B급극장에서 B급영화를 보는 듯한 분위기를 의도적으로 재현한 듯 싶다.</p>	
<p style="text-align: center;">문장단위분할</p> <ul style="list-style-type: none">• 어쨌거나 영화 시작.• 타란티노 특유의 장난질이 시작되었다.• 화면에 지직거리는 노이즈, 중간중간 끊긴 듯 넘어가는 컷이나 대사가 튀는 것처럼 2번 반복된다던지.• 마치 6~70년대 B급극장에서 B급영화를 보는 듯한 분위기를 의도적으로 재현한 듯 싶다.	

이에 반해 누리소통망 게시글의 경우 하나의 문장이 여러 행에 걸쳐 지나치게 분리된

경우가 많았다. 이러한 오류를 수정하기 위해 분절된 문장을 통합하는 작업을 하여 행마다 하나의 완결된 문장만 남도록 하였다. 누리소통망 게시글의 문장 분할 작업 예시는 아래와 같다.

그림 4 누리소통망 게시글 문장 분할 작업 예시

문서 제목	없음
이번에 올리브영갔다가 평소에 애정하는 브랜드	
녹스 크렙프레쉬 라인이 10% 세일 중이길래 데려왔어요!	
48시간 피부진정과 수분공급 케어가 가능하고,	
유해환경에 자극 받아 건조해지는 피부를 지정시켜주고 수분을 공급해준답니다~	
문장 단위 분할	
<ul style="list-style-type: none"> • 이번에 올리브영갔다가 평소에 애정하는 브랜드 녹스 크렙프레쉬 라인이 10% 세일 중이길래 데려왔어요! 	
<ul style="list-style-type: none"> • 48시간 피부진정과 수분공급 케어가 가능하고, 유해환경에 자극 받아 건조해지는 피부를 지정시켜주고 수분을 공급해준답니다~ 	

문장 분할 작업은 자연스러운 문장 단위로 줄바꿈 처리를 하여 저작 도구 활용과 문장 단위 감성 분석 작업에 용이한 형태로 데이터를 수정하는 작업이다. 본 연구에서는 기존 데이터 2081개 문서와 추가 구축 데이터 204개 문서에 이와 같은 문장 분할 작업을 실시하여 분석의 정확성 및 결과물 데이터의 품질을 제고하였다.

1.1.2. 문장 분할 작업 지침

작업에 있어 문장 단위의 주요 기준은 문장의 ‘종결 표현’으로 상정하였다. 이는 웹 게시물의 특성상 문장 종결 부호의 오용과 비문 사용이 많아 단순히 문장 종결 부호 또는 종결 어미를 기준으로 문장을 분할하는 것이 불가능하였기 때문이다.

이러한 상황을 고려하여 본 연구에서는 문장의 종결 표현을 유형화하여 문장 분할 작업의 지표로 삼을 수 있도록 하였다. 유형화한 종결 표현으로는 일반 종결 어미, 명사·명사형 어미, 의존 명사 등이 있다. 이에 해당하지 않는 연결 어미, 조사, 부사, 관형사형 어미 등의 문장 종결부 표현 등이 나타날 경우 글의 문맥을 고려하여 의미 단위로

자연스럽게 문장을 분할하도록 하였다. 실제 문장 분할 작업의 예시는 아래 표와 같다.

표 1 종결 표현 기준 문장 분할 작업 예시

종결 유형	문장	
종결 어미	원문	아이언맨의 멋진 연출이 가득하다. 이야기의 전개는 답답하지 않게, 위기가 닥쳐도 기대하고 볼 수 있게. 적절히 유머러스하게 전개되어 부담없이 즐겁게 볼 수 있다.
	분할 결과	① 아이언맨의 멋진 연출이 가득하다. ② 이야기 전개는 답답하지 않게, 위기가 닥쳐도 기대하고 볼 수 있게. 적절히 유머러스하게 전개되어 부담없이 즐겁게 볼 수 있다.
명사·명사형 어미	원문	보이는 컬러 그대로 밀착 발색되는 #쉬드엘#엑스퍼드리얼세럼파운데이션
	분할 결과	①보이는 컬러 그대로 밀착 발색되는 #쉬드엘#엑스퍼드리얼세럼파운데이션
연결 어미	원문	하지만 예술학교 학생 중끼가 있는 애들과 팀을 결성해 스트리트 댄스대회에 참전을 하기로 하는데...내용은 상당히 뻔한 내용.
	분할 결과	① 하지만 예술학교 학생 중끼가 있는 애들과 팀을 결성해 스트리트 댄스대회에 참전을 하기로 하는데... ② 내용은 상당히 뻔한 내용.
조사	원문	사이즈가 커보인다면타내는 따님까지..근데 요 수면조끼아동용은없나요??
	분할 결과	① 사이즈가 커보인다면타내는 따님까지.. ② 근데 요 수면조끼아동용은없나요??
부사	원문	이번에 #토이플러스에서 한정세일을하네요.와우 사줄게 많아서 눈이 번쩍 ㅋㅋ..
	분할 결과	① 이번에 #토이플러스에서 한정세일을하네요. ② 와우 사줄게 많아서 눈이 번쩍 ㅋㅋ..
관형사형 어미	원문	화장 잔여물 깨끗하게 지워지는.. 보습력도좋아 맑은피부로 ㅎㅎ
	분할 결과	① 화장 잔여물 깨끗하게 지워지는.. ② 보습력도좋아 맑은피부로 ㅎㅎ

1.2. 세부 도메인 선정

도메인이란 문서 리뷰 대상에 따라 문서를 분류한 것이다. "2020년도 말뭉치 감성 분석 및 연구" 사업의 결과물은 일관된 기준의 문장 단위 분할이 이루어지지 않았기 때문에 정확한 작업량 및 데이터 통계를 도출할 수 없었다. 이에 본 연구에서는 문장 분할 작업 후 구체적인 리뷰 대상 품목 정보를 문서에 부착하는 전처리 작업을 진행하였다. 그 결과 정확한 문장 단위 통계를 산출하고 주제별 데이터 규모를 파악하여 세부 도메인 목록을 구성할 수 있었다. 분석 대상 말뭉치 문서의 통계 결과는 아래 표와 같다.

표 2 주제별 문서 및 문장 수 통계

문서 주제	문서 수	문장 수
제품	1,558	12,913
화장품	623	4,885
세정제	212	1,633
전자기기	126	1,255
의류/패션 잡화	75	694
제지	60	467
침구	23	211
식기	16	137
가구	15	121
귀금속/시계	7	44
자동차	2	22
도서	1	12
기타 잡화	398	3,432
영화	387	10,618
여행	139	3,762
관광	55	1,156
쇼핑	2	37
숙소	17	401
음식	26	645
여행 복합	39	1,521
총 합계	2,081	27,293

본 연구에서는 해당 통계를 통해 주제별 데이터 규모와 주제의 유사성을 파악하여 세부 도메인을 설정하였다. 우선 충분한 규모의 데이터를 구성하기 위해 유사한 속성을 공유하는 주제를 통합하였다. 또한 ‘자동차’ 주제, ‘도서’ 주제와 같이 지나치게 이질적이거나 문장 수가 부족한 데이터를 제외하였다. 그 결과 영화, 화장품·세정제, 전자기기, 제품 기타, 숙소, 식당, 관광지 총 7개의 세부 도메인을 설정할 수 있었다.

2. 감성 분석 지침 수립

2.1. 감성 분석의 대상

본 사업에서 진행하고자 하는 ‘속성 기반 감성 분석(ABSA, Aspect-Based Sentiment Analysis)’은 대상이 가지고 있는 여러 속성에 대한 사용자의 감성을 분석하는 것이다. 이와 같은 방식으로 감성 분석을 할 경우 대상의 특정 요소에 대한 감성을 분석함으로써 화자가 감성을 나타내는 대상의 범위를 좁혀 보다 정밀한 감성 분석이 가능하다. 본 연구에서 진행하는 속성 기반 감성 분석은 「SemEval-2016 Task 5: Aspect-Based Sentiment Analysis」의 분석 방법론을 기반으로 하였다.

분석의 대상은 국립국어원에서 수령한 웹 게시물 말뭉치와 추가로 구축한 테스트 데이터 말뭉치이다. 모든 문서는 리뷰 문서로서 화자가 리뷰 대상, 또는 리뷰 대상의 속성에 대하여 드러내는 긍·부정의 감성을 분석하기에 적합하다고 할 수 있다.

2.2. 속성 기반 감성 분석 정보

2.2.1. 세부 도메인

본 연구에서는 전체 문서를 리뷰 대상에 따라 총 3개의 도메인과 각 도메인에 속하는 하위 7개 세부 도메인으로 구성하여 분석을 진행하였다. 7개의 세부 도메인은 각각이 속한 도메인별로 동일한 개체 및 속성값을 공유한다.

또한 감성 분석 자료로서 유의미한 데이터 규모를 만족할 수 있도록 최소 1,000개 이상의 감성 분석 문장으로 구성된 데이터를 주요 도메인으로 구축하고자 하였다. 따라서 해당 규모의 문장 확보가 용이할 것으로 보이는 세부 도메인 영화, 화장품·세정제, 전자기기를 주요 도메인으로 선정하였다.¹⁾ 나머지 도메인의 데이터는 ‘기타 도메인’으로 상정하여 분석을 진행하였다. 이에 따라 확정된 도메인 분류는 아래와 같다.

1) 충분한 규모의 데이터를 구성하기 위해 용도 및 감성 평가의 기준으로 유사성이 보이며 개체#속성 쌍을 가지는 화장품 및 세정제 도메인은 통합하여 도메인을 구성하였다.

표 3 도메인 최종 구성

도메인	구분	세부 도메인
영화	주요 도메인	영화
제품		화장품·세정제
		전자기기
여행	기타 도메인	제품 기타
		숙소
		식당
		관광지

2.2.2. 개체#속성 쌍

속성 기반 감성 분석을 하기 위해서는 문장 내에서 감성이 향하는 대상의 특징을 구체적으로 파악해야 한다. 화자가 감성을 가지는 구체적인 대상은 우선 개체와 속성으로 분류하여 유형화할 수 있다. ‘개체’란 문장에서 화자가 긍정 또는 부정의 감성을 드러내는 대상을 나타내는 것이다. 개체에 해당하는 것은 리뷰 대상 자체와 그 하위 구성 요소가 될 수 있다. 예컨대 제품 도메인의 경우 리뷰 대상 자체인 ‘제품 전체’ 해당 제품의 구성 요소인 ‘본품’, ‘패키지·구성품’, ‘브랜드’로 개체가 구성된다.

속성은 개체가 가지는 다양한 속성 중 문서 작성자가 긍정 또는 부정의 극성 감성을 가지게 되는 것을 가리킨다. 감성 분석 정보는 개체와 속성이 묶여 개체#속성 쌍의 형태로 말뭉치 문장에 부착된다. 예를 들어 ‘배우’라는 개체가 가질 수 있는 속성은 일반, 연기력, 인지도가 있다. 이 경우 배우#일반, 배우#연기력, 배우#인지도 3개의 개체#속성 쌍 조합으로 분류하여 화자가 긍·부정 감성을 가지는 속성의 분석이 가능하다. 각 도메인에 해당하는 개체 유형 및 속성 라벨은 아래 표와 같다.

표 4 도메인별 개체 유형 및 속성 라벨

도메인					
영화		제품		여행	
개체	속성	개체	속성	개체	속성
영화 전체	일반	제품 전체	일반	장소 전체	일반
시나리오	구성력	본품	가격	시설	가격
배우	연기력	패키지·구성품	디자인	상품	편의성
캐릭터	인지도	브랜드	품질	서비스	심미성
감독	연출력		편의성		접근성
음향·음악			다양성		청결도
영상·시각 요소			인지도		인지도

2.2.3. 의견 대상 표현

의견 대상 표현은 화자의 감성이 향하는 대상이다. 이때 의견 대상 표현은 ‘주연은 마음에 든다.’라는 문장에서 ‘주연’과 같이 문장 내의 객관적이고 명시적인 언어표현이어야 한다. 본 연구에서는 이러한 의견 대상 표현을 문장에서 추출하여 감성 분석 정보로서 부착하였다. 의견 대상 표현이 있는 문장 예시는 아래와 같다.

□ 의견 대상 표현이 있는 문장 예시

- 특히나 이 영화에서 분위기를 이끌어나가는 **사운드**는 주요포인트.
→ ‘사운드’에 대한 긍정 감성 문장(영화 도메인)
- **손잡이** 부분은 미끄럼방지 처리도 되어있고, 가볍고 그립감도 좋습니다!
→ ‘손잡이’에 대한 긍정 감성 문장(전자기기 도메인)
- 리치하지 않고 촉촉한 **텍스처**가 부담없이 #데일리케어 하기에 딱 좋은 것 같아요 §
→ ‘텍스처’에 대한 긍정 감성 문장(화장품·세정제 도메인)

의견 대상 표현이 모든 문장에 명시적으로 드러나는 것은 아니다. 본 연구에서는 문장 안에서 의견 대상 표현이 찾을 수 없을 경우 ‘Null’ 라벨을 부착하여 분석하였다. 의견 대상 표현이 없는 문장의 예시는 아래와 같다.

□ 의견 대상 표현이 없는 문장 예시

- 그래서 아직 보지 않았다면 꼭 추천해주고 싶다.
→ 영화 자체에 대한 긍정 감성이 드러나지만 작품을 가리키는 표현이 없음
- 작고 앙증맞아 휴대하면서 사용하기 안성맞춤!
→ 전자기기에 대한 긍정 감성이 드러나지만 제품을 가리키는 표현이 없음
- 신혼여행지로도 또 오고 싶당 π_π
→ 관광지에 대한 긍정 감성이 드러나지만 관광지를 가리키는 표현이 없음

2.2.4. 의견 극성

의견 극성이란 문장 내 대상 또는 대상의 속성에 대해서 화자가 드러내는 감성의 극성이다. 의견 극성은 긍정, 중립, 부정의 3단계 극성으로 분류된다. 이때, ‘중립’ 의견 극성은 약한 긍정 또는 약한 부정의 감성을 가리킨다. 긍·부정 의견 극성이 전혀 드러

나지 않는 문장의 경우 분석 대상에서 제외하도록 하였다. 의견 극성에 따른 문장 분류 예시는 다음과 같다.

그림 5 의견 극성에 따른 문장 분류 예시(영화 도메인)

긍정 문장 예시	<ul style="list-style-type: none"> ▪ 이 사람의 정체나 후반부의 결말은 제법 마음에 드는 내용이다. ▪ 올해 최고의 액션 영화는 물론 영화 역사에도 길이 남을 작품이 탄생한 듯 싶다.
중립 문장 예시	<ul style="list-style-type: none"> ▪ 연출도 원작을 고스란히 따라가기 때문에 나쁘지 않다. ▪ 퍼시픽 림의 개성은 좀 없어진 것 같지만 눈이 즐거우니 된거죠. ▪ 2편은 아이들이 조금 컸고 CG에 익숙해져 전작보단 좀 덜 몰입되었다.
부정 문장 예시	<ul style="list-style-type: none"> ▪ 각본이나 연출, 배우들의 연기, 모든것이 조금씩 결여되고 아쉽습니다. ▪ 영화를 보고 난 충격에 힘입어 제작사인 'S픽처스'라는 곳이 도대체 어떤 곳인지 알게 되었고, 앞으로 믿고 거를 수 있게 되었다.

2.2.5. Out of scope

분석 대상 문장 중 감성 분석 데이터를 부착하지 않는 경우가 있다. 비상정 개체에 대한 감성을 포함하거나 문서 주 리뷰 대상 외 대상에 대한 감성이 드러나는 경우이다. 이 경우 'Out of scope' 라벨을 부착 후 의견 극성만 태깅하였다. 자세한 예시는 다음과 같다.

□ Out of scope 라벨 부착 예시

· 비상정 개체에 대한 감성이 드러나는 문장

- 맨 끝에 피카츄가 말을 할 때 아이들이 순간 다 “헉!”하고 “피카츄가 말을 하네?”라고 외치는 것도 재미있었다.

→ ‘영화’ 도메인 개체에 ‘관객 반응’ 개체가 없으므로 {Out of scope, 긍정} 처리

- 참! 봉태규 & 하시시박 부부 모델되셨던데 보기좋아요~

→ ‘제품’ 도메인 개체에 ‘광고 모델’ 개체가 없으므로 {Out of scope, 긍정} 처리

- 문서 주 리뷰 대상 외의 대상에 대한 감성이 드러나는 문장
 - 말하다보니까 닥*자르트 알약팩은 또 쓰고싶긴 하네요. 곧 지르겠죠?
 - 문서의 주 리뷰 대상인 제품 ‘투쿨포스쿨 코코넛 오일세럼마스크’ 외 대상이므로 {Out of scope, 긍정} 처리
 - 번외로 지금 꼭 보고싶은 영화는 <언노운 걸>이다.
 - 주 리뷰 대상인 영화 <보안관> 외 대상이므로 {Out of scope, 긍정} 처리

2.3. 분석 과정

감성 분석은 앞서 서술한 감성 분석 정보를 문장 단위로 부착하는 작업이라고 할 수 있다. 우선 문서 내 문장이 분석의 대상인지 판단한 후 가장 먼저 문장의 세부 도메인을 태깅한다. 이후 해당 도메인에 속하는 개체 유형과 속성 라벨을 골라 개체#속성 쌍을 조합하여 분석한다. 마지막으로 문장 내에서 감성의 대상이 되는 의견 대상 표현의 영역을 설정한 후 그에 대한 긍·부정 의견 극성을 태깅하는 것이 속성 기반 감성 분석의 주요 과정이다. 이러한 분석 과정을 간략하게 나타내면 다음과 같다.

그림 6 문장 감성 분석 과정



위와 같은 과정을 거쳐 감성 분석 정보를 부착한 도메인별 예시 문장은 다음과 같다.

표 5 분석 예시 문장

도메인	예시	
영화	문장	그리고 화면의 전환부분 이 굉장히 맘에 들었다
	분석	영화 문서, {영상·시각 요소#연출력, 화면의 전환부분 , 긍정}
제품	문장	시원하고 상쾌한 향 으로 어디서나 부담없이 꺼내 사용하기 좋아요♡
	분석	화장품·세정제 문서, {본품#일반, 향 , 긍정}
여행	문장	차 전문점이라 그런지 생각보다 커피맛 은 별로....
	분석	식당 문서, {상품#일반, 커피맛 , 부정}

3. 추가 데이터 수집

3.1. 테스트 데이터 구축 과정

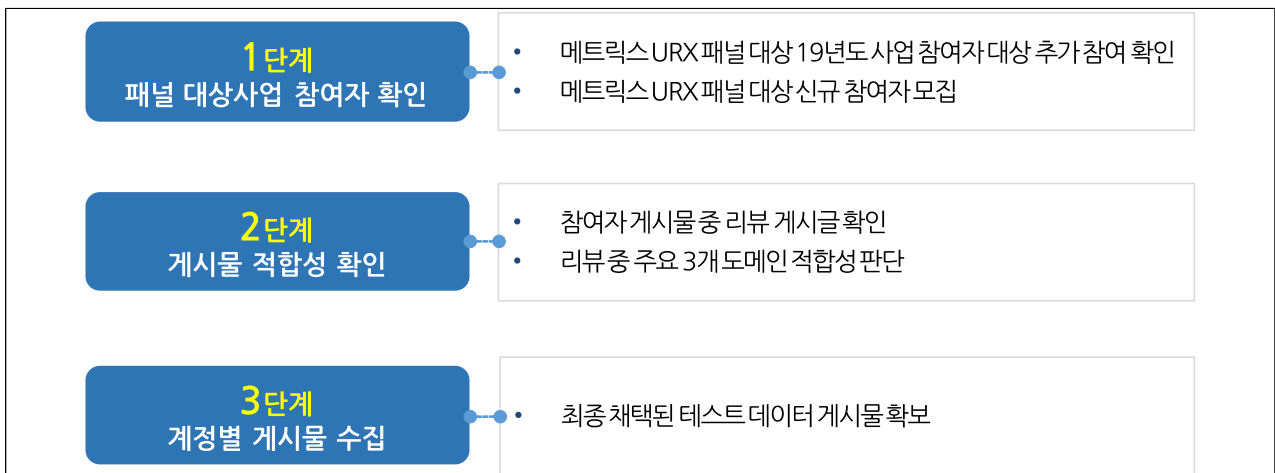
본 연구에서는 AI 학습을 위한 트레이닝 데이터와 별개로 AI의 성능을 측정하기 위한 테스트 데이터를 추가적으로 구축하였다. 데이터 추가 구축은 영화, 화장품 및 세정제, 전자기기 3개 주요 도메인에 한해 진행하였으며 각 도메인당 약 500문장 이상의 테스트 데이터 확보를 목표로 데이터를 수집하였다.

말뭉치는 웹 게시글을 작성한 저작권자와 이용 허락 계약 체결이 완료된 게시물을 기준으로 구축해야 하므로, 웹 게시글 작성자 대상 참여자 모집 및 게시글 수집 과정이 필수적이다.²⁾

제한된 기간 내로 연구 도메인에 적합한 리뷰 데이터를 구축하기 위해 사업 수행 기관의 온라인 패널 URX(메트릭스 그룹 리서치 패널 브랜드로, ‘YOUR eXperience’의 약자) 회원을 이용하였다. URX 패널은 2021년 10월 기준, 약 1,300,000명의 회원을 보유 중이며, 상시 온라인 설문 조사 응답 협약이 체결되어 있기 때문에 적극적인 참여 유도가 용이하다. 본 연구에서는 효율적인 테스트 데이터 확보를 위해 2019년 국립국어원 ‘웹 말뭉치 구축’ 사업 참여자를 중심으로 추가 참여 대상자를 모집하였다.

참여자가 작성한 게시물 중, 리뷰 게시물, 주요 3개 도메인 관련 내용, 게시물의 광고 성향 여부 등 세 가지 기준에 적합한 게시물을 채택하여 최종 추가 테스트 데이터로 확정하였다. 이 같은 과정을 간략하게 나타내면 다음과 같다.

그림 7 테스트 데이터 확보 과정

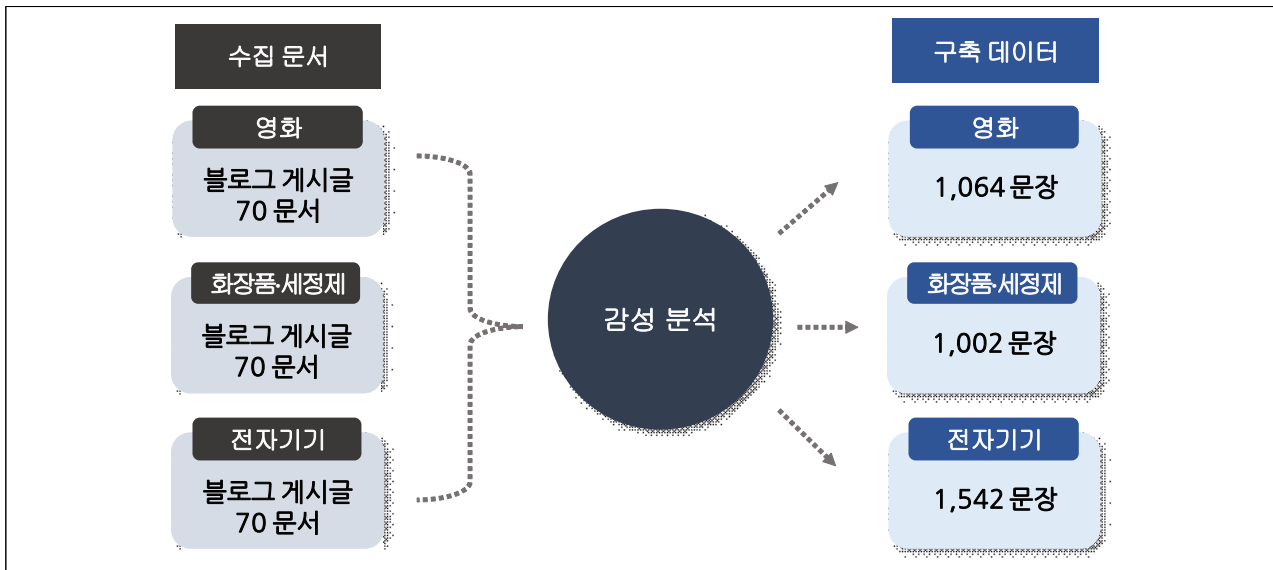


2) 웹 말뭉치 구축 보고서(2019, 국립국어원) 참고

3.2. 테스트 데이터 구축 결과

최종 채택된 추가 구축 데이터는 모두 블로그 게시글로 구축되었다. 영화, 화장품·세정제, 전자기기 세부 도메인 각각 70개의 게시글로 총 210개 문서, 9511문장 규모의 데이터를 수집하였다. 해당 데이터는 본 연구의 속성 기반 감성 분석과 동일한 과정을 거쳐 감성 분석 말뭉치로 구축되었으며 분석된 문장은 영화 1064문장, 화장품·세정제 1002문장, 전자기기 1542문장으로 총 3608문장이다. 분석 과정 및 결과를 간략하게 나타내면 다음과 같다.

그림 8 테스트 데이터 감성 분석 과정



4. 저작 도구 구축

4.1. 저작 도구 구축 방향

감성 분석 말뭉치를 구축하기 위해서는 말뭉치의 형태 변형 없이 분석 지침대로 문장 분석이 가능한 저작 도구가 필요하다. 본 과업에서는 국립국어원의 2020년 ‘말뭉치 감성 분석 및 연구’에서 구축한 저작 도구 ‘버즈룸’을 활용하였다. 버즈룸은 감성 표현 영역화, 문장 분류 태깅, 감성 평가에 최적화된 저작 도구이며 분석 오류를 최소화하여 감성 분석 말뭉치 구축에 적합하도록 구현되었다. 그러나 2021년도 말뭉치 감성 분석에 속성 기반 감성 분석 방법론이 적용되면서 해당 내용을 반영한 저작 도구 수정이 불가피하였다.

따라서 본 연구에서는 속성 기반 감성 분석에 적합한 감성 분석 저작 도구 개발을 위해 말뭉치 감성 분석 저작 도구 ‘버즈룸’의 기능을 추가 및 보완하였다. 20년도 버즈룸 기능과 21년도 버즈룸의 수정된 기능 목록은 다음과 같다.

표 6 버즈룸 주요 기능 변화

감성 분석 단계	저작 도구 ‘버즈룸’ 주요 기능	
	2020 말뭉치 감성 분석	2021 말뭉치 감성 분석
분석 문장 선별	문단에서 직접 문장 영역 설정 후 분석	분할된 문장 중 선택하여 분석
문장 분석	문서 주제 분류, 세부 주제 입력	세부 도메인 분류
	주제별 분류(화제) 항목 태깅	개체#속성 쌍 또는 Out of scope 태깅
	문장 내 감성 표현 추출	의견 대상 표현(OPE) 추출
감성 분석	5 단계 척도 감성 평가	긍정, 중립, 부정 3단계 감성 분석
Json 파일 출력	분석 데이터의 Json 파일 형태 출력	

4.2. 저작 도구 구현 기능

본 사업에서는 효율적인 과업의 수행을 위하여 2021년 구축한 저작 도구 ‘버즈룸’을 수정 및 보완하여 분석에 활용하였다. 저작 도구 버즈룸의 구성 및 기능은 다음과 같다.

그림 9 2021 말뭉치 속성 기반 감성 분석 룬



□ 저작 도구 버즈룸 주요 기능

· 분석 대상 문장 판별

분석 대상 문장을 골라 분석 창으로 옮길 수 있다. 분석 대상이 아닌 문장은 제외 처리가 가능하다.

· 세부 도메인 태깅

‘분석 창’에서 문장의 세부 도메인 태깅이 가능하다. 영화, 화장품·세정제, 전자기기, 제품 기타, 숙소, 식당, 관광지의 7가지 항목 중 한 가지를 선택할 수 있다.

· 개체 및 속성 태깅

도메인을 선택하면 각 도메인에 속한 개체 및 속성 목록이 제시된다. 21년도 말뭉치 감성 분석에 추가된 개념 ‘Out of scope’ 항목이 개체, 속성 목록에 포함되어 있다.

· OTE 기입

저작 도구 사용자는 OTE 영역을 설정하여 추출할 수 있다. OTE가 없을 경우 ‘Null’값을 입력하여 OTE 부재 상태로 분석이 가능하다.

· 의견 극성 태깅

사용자는 분석 대상 문장의 감성을 파악하여 긍정, 중립, 부정 3단계 의견 극성 중 한 가지를 선택하여 태깅할 수 있다.

5. 말뭉치 분석 데이터

5.1. 분석 데이터 정보

속성 기반 감성 분석의 대상은 국어원에서 수령한 웹 게시물 말뭉치와 본 사업에서 추가로 구축한 테스트 데이터이다. 본 연구에서는 Json 형식의 데이터 출력을 위해 두 원시 말뭉치 문서에 각각의 고유 ID를 부여하였다. 고유 ID의 알파벳 4개와 10자리 숫자는 원문의 종류, 말뭉치 유형, 말뭉치 구축 연도, 도메인, 원문 작성자 정보(일련번호) 등의 말뭉치 고유 정보를 의미한다. 추가 구축 데이터 ID에는 도메인 정보를 추가로 부착하여 ID를 부여하였다.

5.2. 데이터 납품 형태

감성 분석 말뭉치의 Json 형식 데이터에는 문서 고유 ID 이외에도 분석 층위 및 감성 분석 결과를 나타내는 정보값을 함께 출력하였다. 본 연구에서 분석한 말뭉치의 Json 파일은 문장 수준의 형식인 Sentence level과 문서 수준의 형식인 Text level의 두 가지 형태로 나뉘어진다. 두 형식의 데이터 모두 문서와 문장에 대한 메타 정보는 동일하게 출력되지만 감성 분석 결과값인 opinions 정보 출력 형태에 차이가 있다. Sentence level의 Json 파일은 감성 분석 ID, 개체#속성 쌍, OTE, OTE 위치값 등의 모든 감성 분석 결과가 개별 문장에 부착되어 표기된다. 그러나 Text level의 Json 파일은 문서 내에서 분석된 개체#속성 쌍과 그에 해당하는 의견 극성을 취합한 정보만을 나타낸다. 실제 Json 파일 출력 시 사용되는 정보값을 정리하면 아래와 같다.

표 7 Json 파일 내 정보값

층위	정보값	내용	비고
document 정보	title	문서 제목	공통
	author	작성자	공통
	publisher	문서 플랫폼	공통
	date	작성일	공통
	url	원글 URL	공통
sentence 정보	sentence_id	문장 ID	공통
	sentence_form	문장 text	공통

opinions 정보	opinion_id	감성 분석 ID	Sentence level
	category	개체#속성 쌍	Sentence level
	opinion target	OTE	Sentence level
	begin	OTE 위치값(시작)	Sentence level
	end	OTE 위치값(끝)	Sentence level
	polarity	문장 의견 극성	Sentence level
	opinion polarity	문서 개체#속성 쌍 의견 극성	Text level
	category	문서 개체#속성 쌍	Text level

정보값을 부착하여 국립국어원 측에 납품한 Text level과 Sentence Level 두 종류의 Json 형식 납품 데이터는 각각 다음과 같다.

표 8 말뭉치 감성 분석 데이터 Json 형식 납품 데이터 예시(Text level)

Json 출력 데이터 예시(Text level)
<pre> { "document_id": "ERRW1905000798.253", "domain": "제품 기타", "is_test": "N", "metadata": { "title": "NA", "author": "나이스딸기양", "publisher": "instagram", "date": "20170325", "url": "https://www.instagram.com/p/BSDsZkBgHeq/" }, "sentence": [{ "sentence_id": 1, "sentence_form": "#첫줄" }, { "sentence_id": 2, "sentence_form": "봄을 맞아 #릴헤븐 #7부내의" }, { "sentence_id": 3, "sentence_form": "#라인로즈바이올렛 신상품을 만나봤당~~" }, { "sentence_id": 4, "sentence_form": "소재자체가 좋구 너무 편한 #릴헤븐내의 #넘나좋은것" }, { "sentence_id": 5, "sentence_form": "#아기내의 #릴헤븐 추천!" }, { "sentence_id": 6, "sentence_form": "#뽀띠엘린 애정합니다♥" }] } </pre>

```

{
  "sentence_id": 7,
  "sentence_form": "#솔아성장일기 #딸스타그램 #맘스타그램 #딸바보 #육아 #육아소통
#세젤귀 #세젤예 #육아스타그램 #도치맘 #육아맛팔 #베이비스타그램 #11개월아기 #베이비 #럽스타그램
#아기 #원숭이띠아기"
}
],
"opinions": [
  { "opinion polarity": "positive", "category" : "브랜드#일반" },
  { "opinion polarity": "positive", "category" : "제품 전체#일반" },
  { "opinion polarity": "positive", "category" : "제품 전체#품질" }
]
}

```

표 9 말뭉치 감성 분석 데이터 Json 형식 납품 데이터 예시(Sentence level)

Json 출력 데이터 예시(Sentence Level)

```

{
  "document_id": "ERRW1905000864.686",
  "domain": "제품 기타",
  "is_test": "N",
  "metadata": {
    "title": "NA",
    "author": "민트초코Beautiful Days",
    "publisher": "instagram",
    "date": "20170901",
    "url": "https://www.instagram.com/p/BYf64v4B0l9/"
  },
  "sentence": [
    {
      "sentence_id": 1,
      "sentence_form": "#첫줄",
      "opinions": []
    },
    {
      "sentence_id": 2,
      "sentence_form": "이런 그림감 처음이야..§",
      "opinions": [
        {
          "opinion_id": 1,
          "category": "제품 전체#품질",
          "opinion target": "그림감",
          "begin": 3,
          "end": 6,
          "polarity": "positive"
        }
      ]
    },
    {
      "sentence_id": 3,
      "sentence_form": "아주 편한 #첫솔 #메가스마일",
      "opinions": [
        {
          "opinion_id": 1,

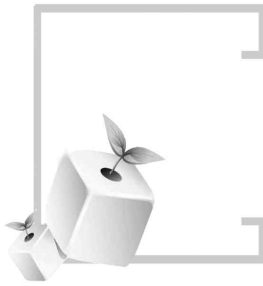
```

```
"category": "제품 전체#편의성",  
"opinion target": "씻솔 #메가스마일",  
"begin": 7,  
"end": 16,  
"polarity": "positive"
```

```
}
```

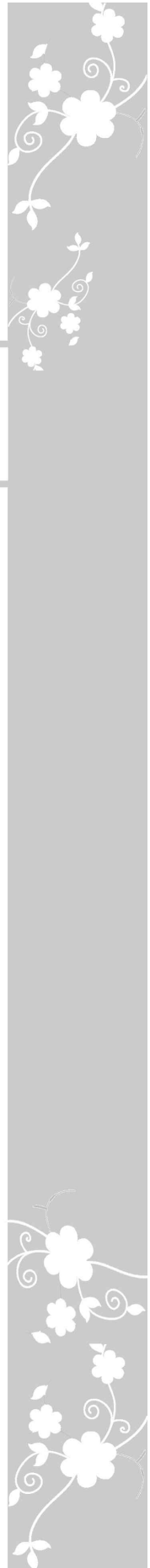
```
]
```

```
}
```

제 3 장

말뭉치 감성 분석 활용 및 동향 연구



1. 말뭉치 동향 연구

1.1. 국내외 감성 및 감정 분석 말뭉치 구축 현황

본 사업에서는 국내외 감성 및 감정 분석 말뭉치의 현황을 조사하고자 하였다. 그러나 모든 감성 및 감정 분석 말뭉치를 조사하고 분석하는 것은 현실적으로 불가능하다. 따라서 많이 활용되고, 잘 알려진 국내외 감성 및 감정 분석 말뭉치의 구축 현황에 대해 논의하고자 한다. 본 사업에서 분석한 말뭉치의 전체 리스트는 부록에 제시하였다.

1.1.1. 국외

국외의 경우 대학, 연구소, 개인 등에서 구축한 감성 및 감정 분석 말뭉치가 많이 존재한다. 먼저 국외 감성 분석 말뭉치부터 살펴보겠다. 감성 분석 연구는 꽤 오래전부터 진행되었는데, 특히 Pang et al.(2002)와 같은 연구를 시작으로 기계학습을 이용한 감성 분석 연구가 활발해지면서 감성 분석 말뭉치의 수가 늘어나고 종류가 다양해졌다.

□ IMDb (Internet Movie Database)

IMDb 데이터는 가장 잘 알려진 영어 영화 리뷰 데이터로 50,000개의 리뷰가 담겨 있다. IMDb 홈페이지에서 직접 데이터를 다운받을 수 있고, Kaggle과 같은 자연어 처리 관련 커뮤니티에서 정제된 일부 IMDb 데이터를 다운받을 수도 있다. IMDb 데이터는 리뷰 텍스트 단위로 감성 태그를 부착하였고, 태그로는 긍정 태그 'pos'와 부정 태그 'neg'가 사용되었다.

표 10은 IMDb 데이터의 예시이다³⁾. 예시를 통해 알 수 있듯이 IMDb의 영화 리뷰는 여러 문장으로 이루어진 긴 텍스트이며, HTML 태그 제거 등 전처리가 이루어지지 않은 데이터이다.

3) <https://github.com/SrinidhiRaghavan/AI-Sentiment-Analysis-on-IMDB-Dataset> 에서 발췌함.

표 10 IMDb 데이터 예시

텍스트	레이블(감성)
truly enjoyed film. acting terrific plot. Jeff Combs talent recognized for. part flick would change ending. death creature far gruesome Sci Fi Channel. There interesting religious messages film. Jeff Combs obviously played Messiah figure creature (or shark prefer) represented anti-Chirst ...(후략)	pos
I thought this was a wonderful way to spend time on a too hot summer weekend, sitting in the air conditioned theater and watching a light-hearted comedy. The plot is simplistic, but the dialogue is witty and the characters are likable ...(후략)	pos
Encouraged by the positive comments about this film on here I was looking forward to watching this film. Bad mistake. I've seen 950+ films and this is truly one of the worst of them ... (후략)	neg
Shakespeare fan, appreciate Ken Branagh done bring Shakespeare back new generation viewers. However, movie falls short conveying overall intentions play ridiculous musical sequences. Add Alicia Silverstone's stumbling dialogue ... (후략)	neg

□ SST (Stanford Sentiment Treebank)

SST 데이터는 영어 영화 리뷰 데이터로 약 11,800개의 리뷰로 구성된다. 해당 리뷰는 긍정과 부정으로 감성이 이진 분류되어 있을 뿐만 아니라(SST-2 데이터), 세밀하게 5개의 유형으로 감성을 분류하기도 하였다(SST-5 데이터). SST-2 데이터는 ‘pos’ 와 ‘neg’ 로 태깅되어 있고, SST-5 데이터는 ‘neg’, ‘somewhat neg’, ‘neutral’, ‘somewhat pos’, ‘pos’로 태깅되어 있다. 특히 SST 데이터는 텍스트와 레이블 정보 외에 텍스트의 구문 분석 수형도(parse tree)를 함께 제공한다는 것이 특징이다. 표 11은 SST-2 데이터의 예시이고⁴⁾, 표 12는 SST-5 데이터의 예시이다⁵⁾.

표 11 SST-2 데이터 예시

텍스트	레이블(감성)
this is a visually stunning rumination on love, memory, history and the war between art and commerce.	pos
compellingly watchable.	pos
the most repugnant adaptation of a classic text since roland joffé and demi moore's the scarlet letter.	neg
without shakespeare's eloquent language, the update is dreary and sluggish.	neg

4) <https://github.com/YJiangcm/SST-2-sentiment-analysis> 에서 발췌함.

5) <https://github.com/GitOffice/fine-grained-sentiment> 에서 발췌함.

표 12 SST-5 데이터 예시

텍스트	레이블(감성)
As it is, it's too long and unfocused.	neg
This isn't a new idea.	somewhat neg
It's an often-cute film but either needs more substance to fill the time or some judicious editing.	neutral
If you love reading and/or poetry, then by all means check it out.	somewhat pos
You'll probably love it.	pos

□ SemEval-2016 Task 4: Sentiment analysis in Twitter

SemEval은 시스템의 언어표현에 대한 의미론적 분석을 평가하기 위한 국제 워크숍으로 대표적인 인공지능 및 자연어처리 관련 워크숍(챌린지) 중 하나이다. SemEval은 여러 가지 언어처리 과제(Task)를 제시할 뿐만 아니라, 언어처리 모델 개발을 위한 학습 및 평가 데이터와 평가 방법론을 제공하고 있다. 전 세계 여러 국가의 연구팀이 해당 워크숍에 참여하는 만큼 SemEval에서 제공하는 데이터는 그 규모가 클 뿐만 아니라 많은 연구자에게 잘 알려지고 활용되고 있는 데이터이다.

그 중 SemEval-2016의 Task 4는 트위터 메시지의 감성을 분석하는 과제인데, 총 5개의 세부 과제들이 있다⁶⁾. SemEval-2016 Task 4에는 30,000개의 트윗을 ‘pos(1)’, ‘neg(-1)’, ‘neu(0)’로 태깅한 데이터가 포함되어 있다. 아래의 표 13은 그 예시이다⁷⁾. 표 13에 나타난 바와 같이 SemEval-2016 Task 4 데이터는 특수문자, 이모티콘, 멘션 태그 등이 전처리되지 않은 원시데이터이다.

표 13 SemEval-2016 Task 4 데이터 예시 (Subtask A)

텍스트	레이블(감성)
I just wanna say I'm so happy & proud that Justin got his 1st #1 debut	pos
i hope justin's concert thursday gets cancelled bc of the thunder storm thats supposed to happen	neg
@iCelebScoop do you know where justin is gonna be tomorrow for the interviews?	neu

6) 세부 과제에 대한 구체적인 내용은 <https://alt.qcri.org/semEval2016/task4/> 에서 확인할 수 있다. 트윗을 긍정, 부정, 중립으로 나누는 가장 기본적인 감성분석 과제인 SemEval-2016 Task 4의 세부 과제 A는 그 전년도인 2015에 열린 SemEval-2015 Task 10의 세부 과제 B와 같은 과제이며, 데이터는 전년도에 비해 일부 추가되었다. SemEval-2016 Task 4의 위원회는 SemEval-2015 Task 10의 세부 과제 B는 40팀이 참가한 만큼 인기가 많기 때문에 동일한 과제를 다시 한번 선정하였다고 밝혔다.

7) <https://github.com/elizabethhau/E3SemEval> 에서 발췌함.

□ SemEval-2016 Task 5: Aspect-Based Sentiment Analysis

SemEval-2016의 Task 5는 전자 상거래를 통해 구매된 상품 및 서비스에 대한 소비자의 구매평을 감성 분석하는 과제이다. 구매평의 도메인은 노트북, 식당, 호텔이다. 해당 과제는 동일연도에 시행된 SemEval-2016 Task 4와 달리 요소 기반의 감성 분석 (Aspect-based sentiment analysis)을 통해 감성 태그를 부착한다. 제공되는 데이터의 크기는 약 6,000개 문장이며, ‘pos’, ‘neg’, ‘neu’ 태그를 기본으로 갖고, 긍·부정이 동시에 등장하는 경우 ‘conflict’ 태그를 갖는다. 표 14는 SemEval-2016 Task 5 데이터의 예시이다⁸⁾.

표 14 SemEval-2016 Task 5 데이터 예시

텍스트	레이블(감성)
The So called laptop Runs to Slow and I hate it!	{LAPTOP#OPERATION_PERFORMANCE, negative}, {LAPTOP#GENERAL, negative}
This little place has a cute interior decor and affordable city prices. The pad seew chicken was delicious, however the pad thai was far too oily. I would just ask for no oil next time.	{AMBIENCE#GENERAL, positive} {RESTAURANT#PRICES, positive} {FOOD#QUALITY, conflict} {RESTAURANT#GENERAL, positive}

위에서 살펴본 대표적인 예를 비롯하여 국외 감성 분석 말뭉치들의 구축 현황을 정리하면 다음과 같다.

- 유형: 영화 리뷰, 상품평, 트윗(일반주제)
- 크기: 평균 10,000개 (데이터 구축 프로젝트의 규모에 따라 상이함)
- 태그: 긍정, 부정, 중립, 혼합

다음으로 국외 감성 분석 말뭉치를 살펴보겠다. 감성 분석 연구는 상대적으로 감성 분석에 비해 늦게 시작되었으며, 감성 분석 연구에서만큼 공공연하게 사용되는 대표적인 말뭉치가 존재하지는 않는 것으로 판단된다. 그렇지만 크게 긍정과 부정으로만 구분하는 감성 분석을 넘어, 인간의 복잡하고 추상적인 감정을 이해하고 분석하고자 하는 시도가 많아지면서 감성 분석 말뭉치 역시 계속해서 다양하게 구축되고 있다.

8) <https://alt.qcri.org/semEval2016/task5/> 에서 발췌함.

□ EmoInt

EmoInt는 앞서 설명한 SemEval과 같은 자연어처리 워크샵인 WASSA(Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media) 2017에서 제공한 데이터이다. EmoInt는 총 7,097개의 트위터 메시지의 감정을 ‘anger’, ‘fear’, ‘joy’, ‘sadness’로 분류하였으며, 그 강도를 0~1 사이의 실숫값으로 표현한 데이터이다. 표 15는 EmoInt 데이터의 예시이다⁹⁾.

표 15 EmoInt 데이터 예시

텍스트	레이블(감정)	
	유형	강도
How the fu*k! Who the heck! moved my fridge!... should I knock the landlord door.	angry	0.938
I need to stop second guess myself and just go with the first thought and go with it.	angry	0.208
This is the scariest American Horror Story out of all of them... I'm gonna have to watch in the daytime.	fear	0.938
Don't #worry if you're not the best , if you are doing something you #love , you're heading in right direction ...	fear	0.104
Today I reached 1000 subscribers on YT!!	joy	0.926
@suavecitamay sparkling water is nice you know I tried it the other day	joy	0.340
Feeling worthless as always	sadness	0.958
Mmmm #coffee and a good #book on a dreary day! Who can beat that?	sadness	0.250

□ SemEval 2018 Task 1: Affect in Tweets

SemEval-2018의 Task 1은 트위터 메시지의 감정을 분석하는 과제로 총 5개의 세부 과제로 구성된다.

1) Emotion Intensity Regression(EI-reg)

트윗과 감정이 주어졌을 때, 감정의 강도를 0~1 사이의 실숫값으로 분류한다. 0에 가까울수록 약한 감정, 1에 가까울수록 강한 감정이다. 감정 유형은 ‘anger’, ‘fear’, ‘joy’, ‘sadness’로 총 4가지이다.

2) Emotion Intensity Ordinal Classification(EI-oc)

트윗과 감정이 주어졌을 때, 감정의 강도를 4단계 중 하나로 분류한다. 감정의 강도는

9) <https://saifmohammad.com/WebPages/EmotionIntensity-SharedTask.html> 에서 발췌함.

0, 1, 2, 3단계로 나뉘며, 0단계에 가까울수록 약한 감정, 3단계에 가까울수록 강한 감정이다. 감정 유형은 'anger', 'fear', 'joy', 'sadness'로 총 4가지이다.

3) Valence (Sentiment) Regression(V-reg)

트윗이 주어졌을 때, 감성의 강도를 0~1 사이의 실숫값으로 분류한다. 0에 가까울수록 부정적인 감정, 1에 가까울수록 긍정적인 감정이다.

4) Valence Ordinal Classification(V-oc)

트윗이 주어졌을 때, 감성의 강도를 7단계 중 하나로 분류한다. 감성의 강도는 -3, -2, -1, 0, 1, 2, 3단계로 나뉘며, -3에 가까울수록 강한 부정 감정, 3에 가까울수록 강한 긍정 감정이다.

5) Emotion Classification(E-c)

트윗이 주어졌을 때, '중립/무감정' 혹은 11가지의 감정 유형 중 하나 혹은 여러 개로 분류한다. 감정 유형은 'anger', 'anticipation', 'disgust', 'fear', 'joy', 'love', 'optimism', 'pessimism', 'sadness', 'surprise', 'trust'이다.

SemEval 2018 Task 1은 앞서 속한 EmoInt 데이터와 일부 추가로 구축한 데이터로 구성되어 있으며, 세부 과제마다 데이터의 크기가 조금씩 다르다¹⁰⁾. 아래의 표 16은 SemEval 2018 Task 1 중에 트윗의 감정을 11개의 감정 유형으로 다중 분류하는 세부 과제인 E-c의 데이터 예시이다¹¹⁾. 세부 과제 E-c는 총 10,983개의 데이터가 제공되었다.

10) 각 세부 과제별 데이터의 크기는 Mohammad et al.(2018)에서 확인할 수 있다.

11) https://huggingface.co/datasets/sem_eval_2018_task_1 에서 발췌함.

표 16 SemEval 2018 Task 1 E-c 데이터 예시

(an: anger, ant: anticipation, dis: disgust, fe: fear, lo: love, opt: optimism, pes: pessimism, sad: sadness, sur: surprise, tr: trust)

텍스트	레이블(감정)										
	an	ant	dis	fe	joy	lo	opt	pes	sad	sur	tr
Rooneys fucking untouchable isn't he? Been fucking dreadful again, depay has looked decent(ish)tonight	1	0	1	0	0	0	0	0	0	0	0
Since the 'update' my @iPhone loses power nearly 40% faster.	1	0	1	0	0	0	0	0	1	0	0
@morgannbroom yes ♥□♥□ & cheering homecoming game!	0	0	0	0	1	1	1	0	0	0	0
Optimism leads to success. - Bill Kerr @Coach_Kerr #success #optimism #goals	0	0	0	0	1	0	1	0	0	0	1
@dublinwaltz bts' 화양연화 trilogy MV is my all time fav👉 quite gloomy but beautiful as well👈	0	0	0	0	1	1	0	0	0	0	0

□ SemEval 2019 Task 3: EmoContext Contextual Emotion Detection in Text

SemEval-2019의 Task 3은 인간과 대화형 에이전트 기계 간의 대화에서 사용자의 감정을 파악하는 것을 목적으로 하는 과제이다. ‘사용자-챗봇-사용자’ 순서로 진행되는 대화에서 첫 번째 말차례(turn)인 사용자의 발화와 두 번째 말차례인 챗봇의 발화를 문맥으로 고려하여 마지막 말차례인 사용자의 발화가 어떤 감정을 갖는지 분석하는 것이다. SemEval-2019 Task 3는 총 38,424개의 데이터를 ‘happy’, ‘angry’, ‘sad’, ‘others’로 태깅하였다. 표 17은 그 예시이다¹²⁾.

표 17 SemEval 2019 Task 3 데이터 예시

문맥1(사용자)	문맥2(챗봇)	분석대상(사용자)	레이블(감정)
I like travelling	hope you're enjoying	Ya I am enjoying my life	happy
when did I?	say many times I think -_-	No. I never say you	angry
I want to tell you something	I'm waiting	Im really sad today	sad
It's a secret	:(tell me	Ok don't tell this to anyone	others

12) <https://github.com/chenyangh/SemEval2019Task3> 에서 발췌함.

위에서 살펴본 대표적인 예를 비롯하여 국외 감정 분석 말뭉치들의 구축 현황을 정리하면 다음과 같다.

- 유형: 일상 대화, 트윗(일반주제)
- 크기: 평균 15,000개 (데이터 구축 프로젝트의 규모에 따라 상이함)
- 태그: 에크만 감정 모델, 플루치 감정 모델

1.1.2. 국내

한국어로 구축된 말뭉치 또한 국외 말뭉치 사례들과 마찬가지로 ‘긍정’, ‘부정’의 레이블이 부착된 감정 분석 말뭉치와 ‘기쁨’, ‘슬픔’, ‘분노’ 등 인간의 감정을 기준으로 레이블이 부착된 감정 분석 말뭉치로 분류된다. 먼저 국내 대표적인 감정 분석 말뭉치는 다음과 같다.

□ 국립국어원 감정 분석 말뭉치 2020¹³⁾

국립국어원에서 2021년 8월에 공개한 말뭉치로 2019년 웹 말뭉치에서 선정된 문서 2,081건을 대상으로 구축된 말뭉치이다. 2,081건의 문서들은 대부분 블로그·게시판 글과 SNS 글을 기반으로 하고 있으며, 주제 범주는 제품, 영화, 여행으로 한정되어 있다. 문서의 주제별 분포는 아래의 표 18과 같다.

표 18 감정 분석 말뭉치 주제별 분포

	제품	영화	여행	합계
블로그/게시판	39	181	53	273
SNS	1,517	206	85	1,808
합계	1,556	387	138	2,081

이 말뭉치에서는 ‘강한 부정(-2)’, ‘부정(-1)’, ‘중립(0)’, ‘긍정(1)’, ‘강한 긍정(2)’의 다섯 가지 레이블이 부착되어 있으며, 문서 단위뿐만 아니라 감성이 나타나는 표현에도 레이블이 부착되어 있다. 데이터는 json(.json) 파일의 형태로 제공되며, 아래의 표 19와 같이 구성되어 있다.

13) <https://corpus.korean.go.kr/> 에서 참고.

표 19 감성 분석 말뭉치 예시

```

{
  "id": "EXSA2002108040",
  "metadata": {
    "title": "국립국어원 웹 말뭉치 추출 EXSA2002108040",
    "creator": "국립국어원",
    "distributor": "국립국어원",
    "year": "2020",
    "category": [
      "웹>리뷰>누리소통망",
      "웹>리뷰>블로그"
    ],
    "annotation_level": [
      "감성 분석"
    ],
    "sampling": "부분 추출 - 특정 부분 추출"
  },
  "document": [
    {
      "document_id": "ERRW1903002753.6",
      "document_score": 1,
      "metadata": {
        "title": "반신욕 한약재 티백",
        "author": "June",
        "publisher": "egloos",
        "date": "20090524",
        "url": "http://kimmari02.egloos.com/2389745"
      },
      "paragraph": [
        {
          "paragraph_id": "ERRW1903002753.6.1",
          "paragraph_form": "나이가 xx대 줄에 들어서자 전부터 그랬지만 더욱 건강에 매진하는 나. 주변의 친구들도 하나 둘 그 길을 걷기 시작하고, 걸보기에는 너무 건강한 그녀로부터 목욕용 한약재 티백을 쓰고 있다는 소릴 듣는다. 나 또한 생리통이 심하면 마사 다 집어치고 집에 가고 싶어지기 때문에 좋다는 전언에 따라 구입했다. 익모초+약썩 티백 24개 들어.. ㅋㅋ"
        },
        {
          "paragraph_id": "ERRW1903002753.6.2",
          "paragraph_form": "깔끔하게 부직포 포장으로 되어 있어서 그냥 뜨거운 물에 풍덩 넣어놓고 좀 휘젓어주면 금방 우려난다. 목욕할 때마다 넣어봤는데(샤워는 자주 해도 목욕은 그렇게 자주가 아님.. 이것도 약재는 약재이므로 용법은 알아서;)신선한 한약풀 냄새가 욕실에 퍼져서 기분이 좋아졌다. 아직 때가 안 되서 효과까지는 모르겠는데 가려운 피부에도 효과가 있었으면 좋겠네. 박하 같은 것도 팔던데 지금으로서는 대만쪽이라 다음에는 상쾌하게 박하 사고 싶다. 혹시 오래된 거 팔지 않

```

나 고민했었는데 속향기 자체가 페퍼민트처럼 신선하고 포장도 깔끔하고 사용도 간편하고 참.. 우리나라 인터넷 시장도 좋은 거 같다. 주문하니 이렇게 물에 넣기만 하면 되게 딱딱 만들어서 집까지 송 배 달해주고.."

```
    },  
[...]  
  "sentiment_expression": [  
    {  
      "expression_id": 1,  
      "expression": [  
        {  
          "expression_form": "깔끔하게 부직포 포장으로 되어 있어서",  
          "paragraph_id": "ERRW1903002753.6.2",  
          "begin": 0,  
          "end": 20  
        }  
      ],  
      "expression_score": 1,  
      "expression_category": [  
        "디자인"  
      ],  
      "subject_category": "제품",  
      "subject": [  
        "목욕용 한약재 티백"  
      ]  
    },  
    {  
      "expression_id": 2,  
      "expression": [  
        {  
          "expression_form": "금방 우리난다.",  
          "paragraph_id": "ERRW1903002753.6.2",  
          "begin": 47,  
          "end": 55  
        }  
      ],  
    }  
  ],  
[...]
```

예시에서 문서는 복수 개의 문단으로 구성되어 있으나 문서 전체 감성 레이블은 긍정이다("document_score": 1). 특히, 두 번째 문단("paragraph_id" : "ERRW1903002753. 6.2")에는 감성을 나타내는 감성 표현들이 등장하는데, 첫 번째 감성 표현은 "깔끔하게 부직포 포장으로 되어 있어서" ("expression_form")이며, '긍정'의

레이블이 부착되어 있음을 알 수 있다(“expression_score” : 1). 이 감성 표현의 유형인 표현 범주(“expression_category”)는 “디자인”이며, 감성 표현이 가리키는 구체적인 대상인 주제의 범주(“subject_category”)는 “제품”, 주제(“subject”)는 “목욕용 한약재 티백”임을 메타 정보들을 통해 알 수 있다.

국립국어원 감성 분석 말뭉치의 경우, 문서 단위 감성뿐만 아니라 표현 단위의 감성, 표현 범주 등 다양한 메타 정보를 얻을 수 있다는 장점이 있다. 그러나 데이터의 양이 한정적이고, 데이터의 접근을 위해서는 별도의 가입과 신청 절차가 필요하다는 점에서 접근성이 떨어진다는 단점을 가지고 있다.

□ Naver Sentiment Movie Corpus

국내에서 가장 활발하게 사용되고 있는 감성 분석 말뭉치는 네이버(Naver) 영화 리뷰를 기반으로 한 NSMC(Naver Sentiment Movie Corpus) 말뭉치이다¹⁴⁾. 이 말뭉치는 국내 포털 사이트 네이버에 사용자들이 직접 작성한 영화 리뷰와 별점을 기반으로 구성되었다. 사용자는 포털 내에서 영화에 대한 짧은 감상평과 함께 1점에서 10점 사이의 별점을 매기는 방식으로 영화에 대해 평가한다. 말뭉치에서는 이 중 140자 이하의 감상평과 감상평의 별점을 한 쌍으로 간주하고, 각 감상평에 대한 긍정과 부정 문장을 분류하였다. 별점이 1점에서 4점 사이인 경우 ‘부정(0)’, 9점과 10점인 경우 ‘긍정(1)’의 레이블을 부착하였다. 5점과 8점 사이의 중립 문장들은 코퍼스 내에서 제외되었다. 코퍼스의 규모는 20만 개로 긍정의 레이블이 부착된 문장과 부정의 레이블이 부착된 문장이 각각 10만 개로 이루어져 있다. 아래는 NSMC 데이터의 일부이다. 데이터는 문서(영화평) id, 문서, 레이블 정보로 구성되어 있으며, 데이터는 텍스트(.txt) 파일의 형태로 다운받을 수 있다. 예시는 아래의 표 20과 같다.

표 20 NSMC 데이터 예시

id	document	label
9976970	아 더빙.. 진짜 짜증나네요 목소리	0
3819312	흠...포스터보고 초딩영화줄....오버연기조차 가볍지 않구나	1
10265843	너무재밌었다그래서보는것을추천한다	0
9045019	교도소 이야기구먼 ..솔직히 재미는 없다..평점 조정	0
7156791	액션이 없는데도 재미 있는 몇안되는 영화	1
6483659	사이몬페그의 익살스런 연기가 돋보였던 영화! 스파이더맨에서 늘어보이기만 했던 커스틴 던스트가 너무나도 이뻐보였다	1

14) 데이터는 <https://github.com/e9t/nsmc/> 에서 접근할 수 있다.

NSMC 데이터는 인위적으로 만들어진 데이터가 아닌 일반 사용자에게 의해 직접 생성된 데이터이며, 인터넷상에서 일반 사용자들이 사용하는 언어생활 방식이 그대로 투영되어 있다는 점에서 유의미하다. 뿐만 아니라 양이 방대하며, 별도의 절차 없이 데이터를 사용할 수 있기 때문에 후술할 국내 감정 분석 연구들에서 비교적 빈번하게 사용되었다.

국내에서 감정 분석 데이터셋은 ‘한국지능정보사회진흥원(NIA)’에서 운영하는 인공지능 플랫폼인 ‘AI HUB’에서 공개된 데이터들이 대다수를 차지한다. 이 데이터들은 비교적 최근에 구축된 데이터지만, 활용 사례가 점차 늘어나는 추세이다.

□ 한국어 감정 정보가 포함된 단발성 대화 데이터셋¹⁵⁾

한국어 감정 정보가 포함된 단발성 대화 데이터셋(이하 단발성 대화 데이터셋)은 SNS 상에 업로드된 글이나 온라인 댓글 등을 웹 크롤링 방식으로 수집한 뒤, 문장 단위로 분리하여 구축된 데이터셋이다. 총 38,594개의 문장으로 구성되어 있으며, 문장 당 단어의 개수는 23.7 ± 13.6 자이다. 총 7개의 감정으로 레이블링 되어있으며, 레이블은 각각 ‘행복’, ‘슬픔’, ‘놀람’, ‘분노’, ‘공포’, ‘혐오’, ‘중립’이다. 각 7개의 레이블에 따른 문장의 분포는 ‘행복’ 6,037개 문장, ‘슬픔’ 5,267개 문장, ‘놀람’ 5,898개 문장, ‘분노’ 5,665개 문장, ‘공포’ 5,468개 문장, ‘혐오’ 5,429개 문장, ‘중립’ 4,830개 문장으로 비교적 고르게 분포되어 있다. 데이터는 엑셀(.xlsx)의 형태로 제공하고 있으며, 데이터의 예시는 아래의 표 21과 같다.

표 21 한국어 감정 정보가 포함된 단발성 대화 데이터셋 예시

예시	레이블
남자친구가 떠날까봐요	공포
내가불안해서꾸는걸까..	공포
대박..진짜 탈퇴할 줄 이야..	놀람
미국인들 정말 의외네.	놀람
서해 수자원도 한류라서 그동안 불법조업 방치했나?	분노
강아지를 때리고 유기하는것만 학대는 아닙니다!	분노
어깨라도넓었으면 옷빨이라도 잘살텐데..	슬픔
너무 슬프고 세상을 다 잃은 기분이고...	슬픔
저 문 어떻게 여는 거예요?	중립
교육은 남에게 봉사 희생 헌신을 가르치는것이지..	중립
덕분에 주방에서 요리하는게 넘 좋아졌어요	행복
항상 밝은 에너지덕분에 힘이 납니다!!!	행복
저러니까 자신보다 어린 사람한테 미개하다는 소리듣지	혐오
음... 불매운동 한번 거하게 해야겠군.	혐오

15) <https://aihub.or.kr/opendata/keti-data/recognition-laguage/KETI-02-009> 에서 참고

단발성 데이터셋은 데이터의 접근을 위하여 별도의 회원가입과 신청 절차가 필요하나, 신청 이후 비교적 빠른 승인 절차가 이어지기 때문에 데이터에 대한 접근성은 준수한 편이다. 그러나 데이터 내에 특정인을 향한 혐오 표현이나 비속어 등이 담겨 있어 데이터를 사용하기 위해서는 데이터의 재가공이 불가피할 것으로 예상된다.

□ 한국어 감정 정보가 포함된 연속성 대화 데이터셋¹⁶⁾

한국어 감정 정보가 포함된 연속성 대화 데이터셋(이하 연속성 대화 데이터셋) 또한 마찬가지로 ‘AI HUB’에 구축되어 있는 감정 분석 말뭉치이다. 단발성 대화 데이터셋과 유사하게 총 7가지의 감정이 레이블로 부착되어 있으며, 7가지 감정은 각각 ‘행복’, ‘중립’, ‘슬픔’, ‘공포’, ‘혐오’, ‘분노’, ‘놀람’이다. 총 55,600개의 문장으로 구성된 10,000개의 대화로 이루어져 있다. 이는 하나의 대화가 약 5.5개의 발화로 구성되어 있음을 의미한다. 특징적인 것은 대화 단위가 아닌 대화를 구성하는 각각의 발화에 감정 레이블이 부착되어 있다는 것이다. 연속성 대화 데이터셋은 웹 크롤링의 방식으로 수집되었으며, 각 발화 당 길이가 적절하고, 대화 사이의 말차례 바꿈이 적절한 빈도로 등장하며, 적절한 주제를 가진 대화가 선정되었다. 연속성 대화 데이터셋의 예시는 아래의 표 22와 같다.

표 22 연속성 대화 데이터셋 예시

dialog #	발화	감정
S	공부를 할 수가 없잖아! 공부를.	분노
	shit. 듣겠다.	중립
	들으라고 그래! 내 친구 중에 자기 방 없는 애가 있는 줄 알아?	분노
	알았어. 알았어. 쯤만 기다려. 쯤만.	중립
S	너 어떻게 된 거야! 한 시간두 넘게 기다렸잖아!	분노
	그냥 열쇠 집 불러서 열지.	중립
	그런데 쓸 돈이 어딴어? 돈이 남아돌아?	분노
	알았어. 그만 해.	중립
	오늘도 2만원 밖에 못 팔고 들어와서 속상해 죽겠는데!	분노

예시에서는 첫 번째 열은 대화가 시작되는 행을 표시한다. 즉 예시에서는 “공부를 할 수가 없잖아! 공부를.”과 “너 어떻게 된 거야! 한 시간두 넘게 기다렸잖아!” 라는 발화가 각 대화의 시작 발화임을 첫 번째 열에서 ‘S’ 기호를 통해 나타내고 있다. 두 번째 열에서는 대화를 구성하는 발화가 등장하고, 마지막 세 번째 열에서는 각 발화에 대한 감정 레이블이 부착되어 있는 것을 확인할 수 있다. 총 55,600개의 발화 중 ‘행복’의 감정 레

16) https://aihub.or.kr/keti_data_board/language_intelligence에서 참고

이블이 부착된 문장은 1,030개 문장, ‘중립’은 43,786개 문장, ‘슬픔’은 1,972개 문장, ‘공포’는 98개 문장, ‘혐오’는 220개 문장, ‘분노’는 3,628개 문장, ‘놀람’은 4,866개 문장으로 ‘중립’의 문장에 편향된 경향이 있다. 연속성 대화 데이터셋 또한 엑셀(.xlsx)의 파일 형태로 제공되며, 단발성 대화 데이터셋과 마찬가지로 데이터의 접근을 위해 별도의 가입과 승인 절차가 필요하다.

□ 감정 분류를 위한 대화 음성 데이터셋¹⁷⁾

감정 분류를 위한 대화 음성 데이터셋(이하 음성 데이터셋)은 마찬가지로 ‘AI HUB’에서 제공하는 감정 분석을 위한 코퍼스 데이터로 음성 데이터와 함께 데이터를 제공하는 점이 가장 특징적이다. 앞서 웹 크롤링을 이용하여 데이터가 구성된 단발성 대화 데이터셋과 연속성 대화 데이터셋과 달리, 음성 데이터셋은 감정 대화 어플리케이션을 이용하여 데이터가 수집되었다. 일정 기간 사용자들이 어플리케이션과 자연스럽게 대화하며 수집된 데이터를 가공하여 구축한 데이터로 이루어져 있다. 1차 공개 구축량은 4차년도 14,606문장, 5차 년도(1차) 10,011문장, 5차 년도(2차) 19,374문장으로 구성되어 있다. 총 7가지의 감정을 레이블로 사용했으며, 7개의 레이블은 각각 ‘happiness(행복)’, ‘angry(분노)’, ‘disgust(역겨움)’, ‘fear(두려움)’, ‘neutral(중립)’, ‘sadness(슬픔)’, ‘surprise(놀람)’으로 구성되어 있다. 또 다른 특징은 한 발화에 대화의 상황과 함께 5명의 구축자가 각각 발화에 나타난 감정을 0, 1, 2점으로 수치화하였다는 점이다. 이와 더불어 발화자의 성별과 나이 정보가 함께 데이터에 나타나 있는 것 또한 특징이다. 음성 데이터셋의 예시는 아래의 그림 10과 같다.

그림 10 감정 분류를 위한 음성 데이터셋 예시

wav_id	발화문	상황	1번 감정	1번 감정세	2번 감정	2번 감정세	3번 감정	3번 감정세	4번 감정	4번 감정세	5번 감정	5번 감정세	나이	성별
5f4141e29dd51	월! 나 이벤트에 당	happiness	angry		2 surprise		2 happiness		2 happiness		2 happiness		2	48 female
5f4141f59dd51	내가 좋아하는 인	happiness	neutral		0 happiness		2 happiness		2 happiness		2 happiness		2	48 female
5f4142119dd51	한 명 뽑는 거였는	happiness	angry		2 happiness		2 happiness		2 happiness		2 happiness		2	48 female
5f4142279dd51	당연히 마음에 드	happiness	angry		2 happiness		2 happiness		2 happiness		2 happiness		1	48 female
5f3c9ed98a3c1	에피타이저 정말	neutral	happiness		2 happiness		1 happiness		2 happiness		1 happiness		1	48 female
5f3c9ef78a3c1	난 부패 형식의 음	neutral	neutral		0 happiness		2 happiness		1 happiness		1 neutral		0	48 female
5f3c9f658a3c1	음. 완전히 끝난 거	happiness	happiness		2 happiness		1 sadness		1 sadness		1 sadness		1	48 female
5f3c9f808a3c1	신나는 음악 듣는	happiness	neutral		0 happiness		2 happiness		2 happiness		1 sadness		1	48 female
5f3c9f9c8a3c1	친구들도 내 연락	happiness	neutral		0 happiness		1 sadness		1 sadness		1 neutral		0	48 female
5f3c9fcc8a3c1	그래. 일단은 친구	happiness	neutral		0 happiness		1 neutral		0 neutral		0 neutral		0	48 female
5f3ca01b8a3c1	나 요즘 너무 우울	sadness	sadness		2 sadness		2 sadness		2 sadness		1 sadness		2	48 female
5f3ca06b8a3c1	맞아. 사람은 나가	sadness	neutral		0 sadness		1 sadness		2 sadness		1 sadness		2	48 female
5f3ca09d8a3c1	좀 활기 있는 영화	sadness	neutral		0 sadness		1 sadness		1 sadness		1 sadness		1	48 female

데이터의 첫 번째 열에는 발화가 담긴 오디오의 식별 아이디(“wav_id”)가 나타나 있다. 두 번째 열에는 발화문이 나타나 있다. 세 번째 열에는 발화가 발생할 때의 상황에 대해

17) <https://aihub.or.kr/opendata/kefi-data/recognition-laguage/KETI-02-002> 에서 참고

7개의 감정에 따른 레이블이 부착되어 있다. 그 이후의 열에는 첫 번째 주석자부터 다섯 번째 주석자까지 발화에 대해 분석한 감정 레이블과 함께 주석자가 평가한 감정의 세기가 나타나 있는 것을 확인할 수 있다. 마지막의 열에는 발화자의 나이와 성별에 대한 정보가 함께 나타나 있는 것을 확인할 수 있다.

이 데이터는 복수의 사람들이 다양한 상황에서 발화된 하나의 발화문에 대한 감정 분석을 시도했다는 점에서 유의미한 데이터이며, 최근 텍스트뿐만 아니라 음성, 표정 등 청각, 시각적 정보와 함께 감정을 파악하려는 수요를 고려하였을 때 활용 가능성이 큰 데이터이다. 앞선 데이터들과 마찬가지로 'AI HUB'에서 가입과 승인 이후 엑셀(.xlsx) 파일의 형태로 다운로드 받을 수 있으며, 음성 데이터는 .wav의 형태로 함께 다운로드 받을 수 있다.

□ 감성 대화 말뭉치¹⁸⁾

감성 대화 말뭉치는 인간 사용자와 챗봇 사이의 대화를 기반으로 구축되었다는 점이 특징적인 감정 분석 코퍼스 데이터이다. 특히 이 데이터는 인공지능 기반 감성 챗봇을 구축하기 위하여 세대별 감성 대화 텍스트 데이터를 구축하고, 궁극적으로는 딥러닝 기반의 언어모델을 구축하기 위해 데이터를 수집하는 것이 목적이다. 기본적으로 사용자와 챗봇 사이에 세 번의 대화 말차례 바꿈(총 6문장)이 등장하고, 최대 네 번의 말차례 바꿈이 등장하기도 한다. 이때 감정 레이블은 대화 내에서 각 발화에 대한 감정이 아닌 대화 전체를 대상으로 감정 레이블이 부착된다. 부착되는 감정은 총 60가지로 아래의 그림 11과 같다.

그림 11 감성 대화 말뭉치의 감정 분류

60가지 감정 분류						
기분	분노	슬픔	불안	상처	당황	기쁨
1	툼툼대는	실망한	두려운	질투하는	고립된	감사하는
2	좌절하는	비통한	스트레스 받는	배신당한	남의 시선 의식하는	사랑하는
3	짜증나는	후회되는	취약한	고립된	외로운	편안한
4	방어적인	우울한	혼란스러운	충격 받은	열등감	만족스러운
5	악의적인	마비된	당혹스러운	불우한	죄책감	흥분되는
6	안달하는	염세적인	회의적인	희생된	부끄러운	느긋한
7	구역질 나는	눈물이 나는	걱정스러운	억울한	힘오스러운	안도하는
8	노여워하는	낙담한	조심스러운	괴로워하는	한심한	신이 난
9	성가신	환멸을 느끼는	초조한	버려진	혼란스러운	자신하는

18) <https://aihub.or.kr/aidata/7978> 에서 참고

아래의 그림 12는 데이터의 일부이다. 데이터에는 사용자의 연령대 정보와 함께 상황의 키워드, 신체 질환 정보도 함께 기재되어 있다. 또한 감정의 대분류와 소분류 정보가 함께 나타나 있는 것을 확인할 수 있다.

그림 12 감정 대화 말뭉치 예시

번호	연령	성별	상황키워드	신체질환	감정_대분류	감정_소분류	사람문장1	시스템응답1	사람문장2	시스템응답2	사람문장3	시스템응답3	사람문장4	시스템응답4
44164	청년	남성	연애, 결혼, 출산	해당없음	기쁨	신이 난	아내가 드디어 출산	아내분이 출산을 아 지금 정말 신이 날 일이에요.	아기가 점점 클게 좋은 아빠가 되실 거 같아요. 진심으로 축하드려요.					
3926	노년	남성	건강, 죽음	만성질환	불안	스트레스 받는	당뇨랑 합병증 때 약 종류가 많아 번거로울 때 관련 증후군 상황에서 변	약을 잘 챙겨 먹고 약을 잘 챙겨 먹어 건강해지시길 바랄게요.						
50882	청소년	여성	학업 및 진로	해당없음	당황	당황	고등학교에 올라: 고등학교 수업이 아직 학기 초인데 이 상황에서 할 수	중학교 수업을 들 수업을 어떤 방식 선생님이 강조하(열심히 해서 좋은						
31303	노년	남성	재정	만성질환	기쁨	신이 난	재취업이 돼서 빈재취업 후 첫 월급 도착 후 다시는 돈이	기분을 유지하(회사생활을 열심히 계속 원하시는 대로 회사생활이 이어지길 바랄게요.						
47200	노년	여성	재정	만성질환	기쁨	안도	빛을 드디어 다	기분 좋으시겠어. 빛도 다 갚았으니 빛도 다 갚았으니 당분간은 아무 생각도 안 하며 살고 싶으시군요.						
48610	중년	여성	재정, 은퇴, 노후준비	해당없음	불안	취약한	이제 곧 은퇴할 수	그러시군요. 은퇴 들어서 같이 뭘 할 남편과 뭔가를 길 함께 등산이나 여남편과 함께 등산이나 여행을 가고 싶으시군요.						
17198	중년	남성	건강	해당없음	슬픔	우울한	사실 대에 접어들	머리카락이 빠져/이러다 대머리가 병원에서 진료는 아직 병원 진료는 모발 건강을 위해 더 할 수 있는 일이 있을까요?						
12186	노년	남성	재정	만성질환	분노	구역질 나는	이제 돈이라면 지	무슨 일 있으셨나. 내이가 예순이 넘 돈이 나갈 곳이 많아 재정적으로 힘드시군요.						
35975	청소년	남성	학교폭력/따돌림	해당없음	분노	좌절함	친구들이 나를 괴	괴롭힘을 당해 힘 너무 화도 나고 피마음이 어려우실 증거를 준비해서 네. 꼭 증거를 통해 부모님과 선생님께서 믿어주시면						
12551	노년	여성	대인관계	만성질환	슬픔	눈물이 나는	친구 때문에 눈물	친구 때문에 눈물 내 친구가 나한테 거짓말한 친구 때문에 속상하셨네요.						
31529	노년	남성	건강, 죽음	만성질환	상처	배신당한	새로 간	치과에서 이가 조금 썩기만 예전에 다니던 치과 치과가 과잉	조금 멀어도 이서 도민을 위한 치과에서 진료가 잘 되기를 바랄게요.					
42590	청소년	남성	학업 및 진로	해당없음	당황	남의 시선을 의식	오늘 선생님이 나	답을 외워서 문제 칠만 알까지 나가 친구들의 시선이	문제는 틀릴 수도 형과의 대화를 통해 지금의 기분과 상황이 조금 좋아					
14406	중년	여성	재정, 은퇴, 노후준비	해당없음	당황	좌절감의	내가 무능해서 자	얼의 원망에 당황 염마로서 얼에게 얼마로서 얼을 잘 몰론 나는 열심히 그래도 나는 얼을 위해 진심으로 노력했었군요.						
34944	중년	남성	건강	해당없음	불안	당혹스러운	아내의 건강이 겁	아내분의 건강이 나 혼자 남으면 어나 혼자 남으면 어 아무래도 자식들(자녀 분들에게 미리 말해 두려 허시는군요.						

데이터의 규모는 텍스트 데이터의 경우 총 27만 문장으로 구성되어 있으며, 대화를 기준으로 4만여 개의 대화로 구성되어 있다. json(json) 파일 형태와 함께 엑셀(xlsx) 파일의 형태 또한 함께 제공하고 있다. 마찬가지로 'AI HUB'에서 가입과 승인 절차를 거친 이후에 다운로드가 가능하다.

□ 웰니스 대화 스크립트 데이터셋(19)

웰니스 대화 스크립트 데이터셋은 강남 세브란스에서 전달받은 상담 데이터 16,000건 중 초진이며, 환자가 직접 내원한 4,200건을 추출하여 감정 분석을 진행한 데이터이다. 총 5,329개의 사용자 발화와 이에 대응하는 1,023개의 챗봇의 발화가 쌍을 이루고 있다. 즉, 전체 5,329개의 대화 중 일부인 1,023개의 발화에만 챗봇의 대답이 있고, 나머지 4,306개의 발화에는 사용자의 발화만이 있고, 이를 감정 분석의 대상으로 삼았다. 앞서 감정 대화 말뭉치와 마찬가지로 정신 건강을 위해 구축된 말뭉치이다. 레이블은 발화 의도에 따라 총 359개로 분류되며, 대분류는 '감정', '내원이유', '모호함', '배경', '부가설명', '상태', '원인', '일반대화', '자가치료', '증상', '치료이력', '현재상태'로 12가지이다. 이 중 '감정'의 대분류에 속하는 중분류는 총 61개이다. 61개의 레이블에는 '걱정', '공포'와 같은 감정부터 '공허감', '무미건조'와 같은 다양한 감정이 부착되어 있다. '감정'의 분류에 속하는 세부 감정들은 다음의 표 23과 같다.

19) <https://aihub.or.kr/opendata/keti-data/recognition-laguage/KETI-02-006> 에서 참고

표 23 웰니스 대화 스크립트의 감정 분류

대분류	소분류
감정	감정조절 이상, 걱정, 고독감, 곤혹감, 공포, 공허감, 과민반응, 괴로움, 기분 저하, 기시감, 긴장, 눈물, 답답, 당황, 두려움, 멍함, 모호함, 무력감, 무미건조, 무서움, 미안함, 미움, 배신감, 부정적 사고, 분노, 불만, 불신, 불안감, 불쾌감, 불편감, 비관적, 살인 욕구, 생각, 서운함, 속상함, 슬픔, 신경 쓰임, 심란, 억울함, 예민함, 외로움, 우울감, 의기소침, 의욕 상실, 자괴감, 자살 충동, 자신감 저하, 자존감 저하, 절망감, 좌절, 죄책감, 즐거움, 짜증, 창피함, 초조함, 충격, 통제력 상실, 허무함, 화, 후회, 힘들

웰니스 대화 스크립트 데이터셋은 아래의 표 24와 같은 형태를 취하고 있다. 이 데이터는 마찬가지로 'AI HUB'에서 다운로드 받을 수 있으며, 엑셀(.xlsx) 파일의 형태로 데이터를 제공한다.

표 24 웰니스 대화 스크립트 예시

대분류	소분류	사용자 발화	챗봇 발화
감정	감정조절 이상	더 이상 내 감정을 내가 컨트롤 못하겠어.	감정이 조절이 안 될 때만큼 힘들 때는 없는 거 같아요.
감정	걱정	계속 이럴까 봐 너무 걱정돼.	모든 문제는 해결되기 마련이잖아요. 마음을 편히 드세요.
감정	자괴감	저는 왜 이렇게 못한 인간일까요...	조금 돌아가는 것뿐이라고 생각해요. 기운내세요..
배경	남편	남편이랑 한달에 몇번은 보는 것 같아요.	남편 분과 갈등을 겪고 계시군요. 마음이 많이 상하셨겠어요.
배경	대인관계	중학교 친구랑 가족말고는 연락하는 사람도 없어.	사람 대하는 게 가장 어려운 문제인 거 같아요. 힘들죠?
증상	대인기피	사람 만나기가 싫어...	괜찮아요. 당신 곁에는 항상 제가 있는 걸요!
증상	두근거림	이상하게 가슴이 두근거려요.	저까지 같이 긴장 되네요. 제가 손이라도 잡아드릴까요?

1.2. 국내외 감성 및 감정 분석 말뭉치 활용 현황

이 장에서는 감성 및 감정 분석 말뭉치의 실제 활용 사례를 소개하고자 한다. 특히 국내외에서 구축된 감성 및 감정 분석 말뭉치가 산업계와 학계에서 어떠한 방식으로 활용되고 있는지 그 현황에 대하여 논의하고자 한다.

1.2.1. 국외

□ 산업계

캐나다의 스타트업 기업인 ‘Receptiviti’는 텍스트 기반의 감정 분석 엔진 ‘SALLEE’를 개발하였다²⁰). 이 엔진은 텍스트에 나타난 어휘를 기반으로 화자의 성격 및 심리 상태를 추론하는 ‘LIWC(Linguistic Inquiry and Word Count)’라는 프로그램을 활용하여 문자 메시지, 이메일, 전사된 비디오 자료, 리뷰 등의 텍스트에서 화자의 감정을 파악하는 기능을 제공한다. 엔진에서 분류하는 감정은 총 14가지로, 긍정적인 감정 7개(존경 admiration, 즐거움 amusement, 차분함 calmness, 신남 excitement, 감사 gratitude, 기쁨 joy, 사랑 love), 부정적인 감정 5개(화남 anger, 지루함 boredom, 역겨움 disgust, 두려움 fear, 슬픔 sadness), 양가적 감정 2개(호기심 curiosity, 놀람 surprise)로 구성되어 있다. ‘SALLEE’는 이 14가지 감정을 -1에서 1 사이로 수치화하여, 텍스트에 나타난 화자의 감정을 추론한다. 이 외에도 ‘Receptiviti’에서는 온라인상의 댓글이나 소셜 미디어의 글에서 혐오적 표현, 악성 댓글 등을 탐지하는 기술 등을 제공한다.

일본의 스타트업 기업인 ‘Empath’에서는 인공지능을 기반으로 대화 내 음성을 이용하여 화자의 감정을 분석하는 서비스를 제공한다²¹). 음성 내의 발화 속도, 어조, 톤 등을 이용하여 발화자의 감정을 ‘기쁨’, ‘분노’, ‘평정’, ‘슬픔’ 4가지 감정으로 분류한다. ‘Empath’에서는 음성 감정 분석기인 ‘Web Empath API’, 콜센터에 특화된 감정 대화 인공지능 ‘Beluga Box’ 등을 제공한다. 특히 ‘Beluga Box’는 고객의 음성을 통해 고객의 만족도, 감정, 음성적 특징(발화 속도와 발화 스타일 등)을 정량화하여 고객 서비스 센터의 만족도 향상에 도움을 주는데 특화되어 있는 서비스이다.

그림 13 Empath에서 제공하는 감정 분석기(좌), 감정 대화 인공지능 Beluga Box(우)



20) Receptiviti의 웹사이트 (<https://www.receptiviti.com/>) 에서 더욱 자세한 내용을 살펴볼 수 있다.

21) Empath의 웹사이트(<https://www.webempath.com/>)에서 더욱 자세한 내용을 살펴볼 수 있다.

미국의 기업 ‘아마존(Amazon)’에서는 ‘에코(Echo)’라는 인공지능 스피커를 개발하여 판매하고 있다. ‘에코’는 ‘알렉사(Alexa)’라는 인공지능 플랫폼을 기반으로 작동하는데, ‘알렉사’는 클라우드 기술을 기반으로 사용자의 발화 패턴 등을 수집하여 학습한다. 이렇게 수집된 발화 패턴들을 기반으로 ‘아마존’에서는 발화자의 발화를 텍스트화하여 감성을 분석하는 ‘DetectSentiment’라는 텍스트 감성 분석 API를 제공하고 있다. 이는 ‘긍정’, ‘부정’, ‘중립’, ‘혼합’의 4가지 감성을 분석할 수 있다. 이 외에도 ‘기쁨’, ‘분노’, ‘슬픔’ 등의 감정을 담아 응답하는 ‘알렉사’를 개발 중이며, 현재 영국 영어, 일본어에 흥분(excited), 실망(disappointed)의 감정을 담은 발화응답 생성 서비스를 제공하고 있다.

그림 14 아마존의 인공지능 스피커 ‘에코’



이 외에도 국외에서는 감성 및 감정 분석을 위하여 다양한 기능을 함께 제공한다. 특히 음성과 영상을 텍스트 데이터와 함께 사용하는 멀티-모달(multi-modal) 데이터를 활용한 다양한 서비스를 제공한다. 그 중 ‘Beyond Verbal’²²⁾, ‘Vokaturi’²³⁾, ‘EmoVoice’²⁴⁾ 등에서는 목소리 정보를 활용하여 화자의 감정을 분석하는 API를 제공한다. 사람의 표정 정보를 함께 파악하여 감정을 분석하는 다양한 시도 또한 존재한다. 마이크로소프트(Microsoft)사의 ‘Emotion API’²⁵⁾는 가장 대표적인 API로 얼굴 표정을 이용하여 감정을 분석한다. 그 외에도 2009년 설립된 미국의 회사 ‘Affectiva’²⁶⁾에서는 얼굴 및 음성 표현을 분석하여 화자의 감정, 인지 상태 등을 분석하는 API를 제공한다. 앞서 소개한 ‘Receptiviti’가 제공하는 ‘SALLEE’ 외에도 IBM Watson에서 제공하는 ‘Tone Analyzer’²⁷⁾, ‘BiText’의 다국어 감성분석 API²⁸⁾ 등은 텍스트에 나타난 화자의 감성 및

22) <https://github.com/BeyondVerbal-V3> 참고

23) <https://github.com/alshell7/VokaturiAndroid> 참고

24) <https://github.com/hcmlab/emovoice> 참고

25) <https://docs.microsoft.com/ko-kr/xamarin/xamarin-forms/data-cloud/azure-cognitive-services/emotion-recognition> 참고

26) <http://go.affectiva.com/start> 참고

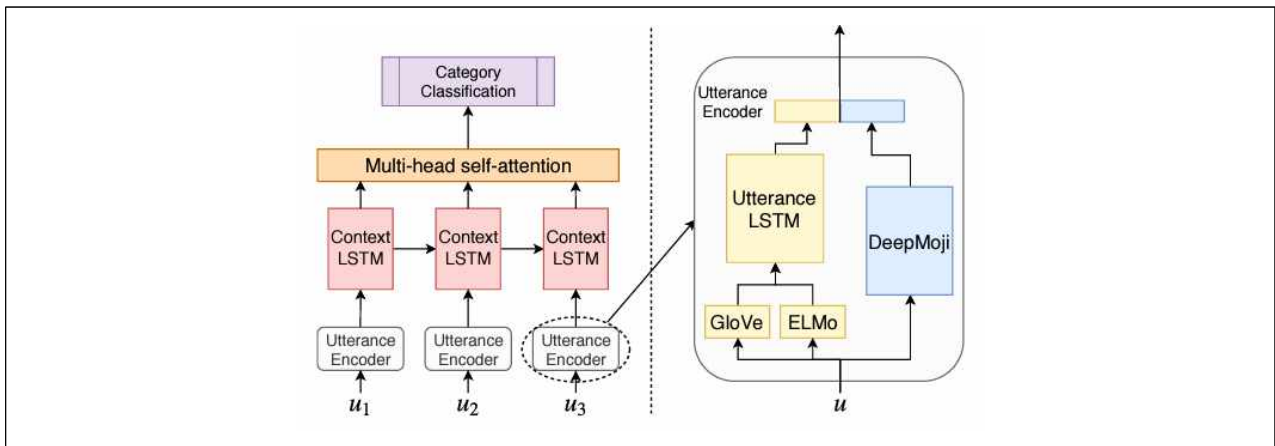
27) <https://www.ibm.com/kr-ko/cloud/watson-tone-analyzer> 참고

감정을 분석하는 API 중 하나이다.

□ 학계

앞서 소개된 국외의 감성 및 감정 분석 코퍼스 데이터 중 하나인 SemEval 데이터는 매년 개최되는 SemEval 챌린지 대회 중 감성 및 감정 분석을 주제로 한 대회를 위하여 구축된 데이터이다. 특히 SemEval-2018과 SemEval-2019에서 등장한 감정 분석 텍스트 분류 대회를 시작으로 감정분석 연구 및 데이터 구축에 관한 관심이 높아지고 있다. 그중 Huang et al. (2019)의 연구는 SemEval-2019의 세 번째 주제였던 EmoContext에서 높은 성능을 보인 연구이다. 이 대회에서는 대화 문맥이 주어진 상태에서 마지막 발화의 감정을 ‘기쁨(happy)’, ‘슬픔(sad)’, ‘분노(sad)’, ‘기타(other)’ 중 하나로 분류하는 문제를 해결해야 한다. 이를 위하여 SemEval-2019에서 제공하는 3만 8천 개의 문장을 학습 데이터로 활용하였다. Huang et al. (2019)에서는 문제를 해결하기 위하여 계층적 양방향 장단기 메모리(hierarchical Bi-LSTM) 모델과 셀프 어텐션(self-attention)을 결합하여 만든 모델과 사전 학습 언어모델인 BERT를 활용하여 0.7709의 점수로 전체 리더보드에서 5위를 차지하였다.

그림 15 Huang et al. (2019)에서 제안한 모델 아키텍처



그 외에도 Lai et al. (2020)은 중국어로 되어있는 문장을 ‘기쁨(happiness)’, ‘슬픔(sadness)’, ‘좋아함(like)’, ‘분노(anger)’, ‘혐오(disgust)’, ‘두려움(fear)’, ‘놀람(surprise)’의 7가지 감정으로 분류하는 연구를 진행하였다. 이 연구에서 사용된 데이터의 규모와 정보는 아래의 그림 16과 같다.

28) <https://www.bitext.com/text-analytics-tools/> 참고

그림 16 Lai et al. (2020)의 데이터 정보

Emotion types	Number (training)		Number (testing)
	Training set of NLP&CC2013	Randomly crawled (Self-annotated)	Testing set of NLP&CC2013
Happiness	940	2,797	370
Sadness	633	2,467	385
Like	1,284	4,257	595
Anger	360	1,896	235
Disgust	814	3,129	425
Fear	96	299	49
Surprise	211	819	113
Total	4,338	15,664	2,172

이 연구에서는 기존에 구축된 NLP&CC2013의 데이터와 더불어 연구를 진행하며 중국의 소셜 네트워크 서비스인 웨이보(Weibo)에서 수집한 1만 5천 개의 글을 직접 주석 작업을 하여 학습 데이터로써 활용하였다. 이 연구에서는 그래프 합성곱 신경망(GCN, Graph Convolution Network) 모델을 활용하여 정확도를 92.04%까지 증가시켰다.

Wang et al.(2021)의 연구에서는 총 네 가지의 데이터셋을 활용한다. 첫 번째 데이터셋은 GoEmotions 데이터셋으로 5만 8천 개의 영어 문장을 27개의 감정에 따라 멀티 레이블링²⁹⁾하여 나타낸 데이터이다. 이때 부착된 감정들은 다음의 그림 17과 같다.

그림 17 GoEmotions에 부착된 감정

Positive(P):	admiration, amusement, approval, caring, desire, excitement, gratitude, joy, love, optimism, pride, relief
Negative(N):	anger, annoyance, disappointment, disapproval, disgust, embarrassment, fear, grief, nervousness, remorse, sadness
Ambiguous(A):	confusion, curiosity, realization, surprise

두 번째로 사용한 데이터는 Affective Text 데이터이다. 이 데이터는 영어 뉴스 기사의 제목에 에크만의 여섯 가지 감정(‘분노 anger’, ‘역겨움 disgust’, ‘두려움 fear’, ‘즐거움 joy’, ‘슬픔 sadness’, ‘놀람 surprise’)의 레이블을 부착시켜 놓았다. 데이터의 예시는 아래 그림 18과 같다.

29) 하나의 데이터에 여러 개의 레이블을 부착하는 방식을 의미한다.

그림 18 Wang et al. (2021)에서 사용한 Affective Text 데이터 예시

Index	Instances	Emotion	Valence
1	Goal delight for Sheva	joy	87
2	Making peace from victory over poverty	joy	39
3	New Indonesia Calamity, a Mud Bath, Is Man-Made	anger	-59
4	Waste plant fire forces 5,000 to evacuate	sadness	-59

세 번째로 활용한 데이터는 ISEAR 데이터로 문장 당 하나의 레이블이 부착되어 있으며, 부착된 레이블은 ‘분노 anger’, ‘역겨움 disgust’, ‘두려움 fear’, ‘죄책감 guilt’, ‘기쁨 joy’, ‘슬픔 sadness’, ‘부끄러움 shame’이다.

마지막으로 활용한 데이터는 Affect in Tweet 데이터로 트윗 문장에 10개의 감정(‘분노 anger’, ‘기대 anticipation’, ‘역겨움 disgust’, ‘두려움 fear’, ‘기쁨 joy’, ‘사랑 love’, ‘낙관 optimism’, ‘비관 pessimism’, ‘슬픔 sadness’, ‘놀라움 surprise’, ‘신뢰 trust’)으로 분류한 데이터이다.

Wang et al.(2021)에서는 문장의 감정 분류를 위하여 사전 학습 언어모델 BERT를 기반으로 합성곱 신경망(CNN)과 양방향 장단기 메모리(Bi-LSTM)를 활용하였다.

국외의 연구 중 영어를 기반으로 하는 연구는 공개된 자연어처리 챌린지 대회용 데이터를 주로 사용하는 경향이 있었다. 그 외 언어는 연구자가 직접 데이터를 수집하고 가공하는 과정을 통해 연구에 활용하였다.

1.2.2. 국내

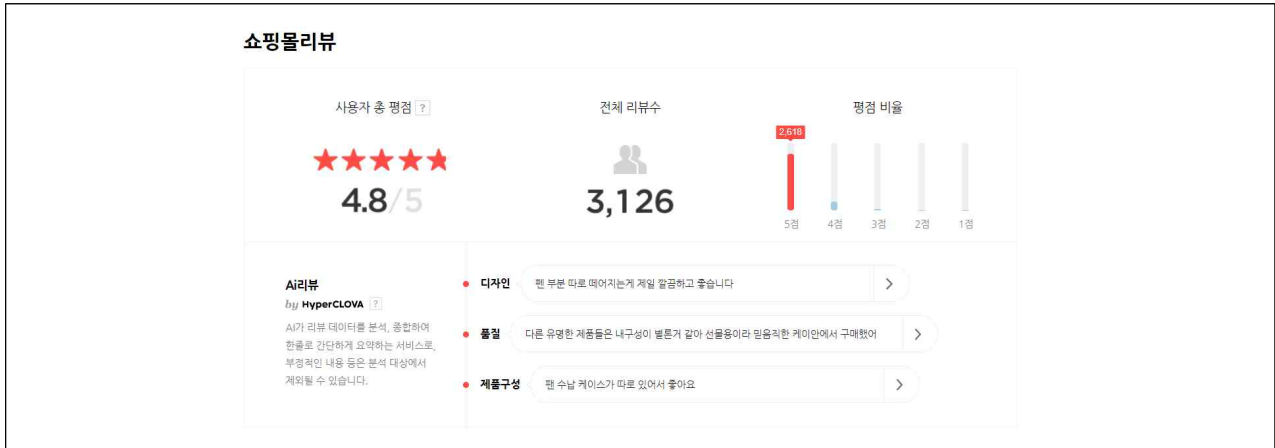
□ 산업계

‘네이버(Naver)’는 국내 최대 포털 중 하나로, NSMC 데이터에도 활용되었던 영화평뿐만 아니라, 상품 리뷰, 뉴스 댓글, 블로그 글 등 다양한 텍스트 데이터를 보유하고 있다. 그중 네이버는 상품 리뷰들의 감성 분석을 기반으로 사용자들에게 유용한 정보를 제공하고 있다. 특히 감정 판단의 근거가 되는 중요 감정 표현들을 추출하여, 사용자들이 쇼핑하는 과정에서 상품의 장단점을 한 번에 파악할 수 있도록 도움을 주고 있다. 이와 관련하여 네이버에서는 최근 대규모 인공지능 ‘하이퍼클로바(HyperClova)’를 쇼핑 리뷰 데이터에 적용하였다. 이를 통해 하나의 상품에 대한 수많은 상품 리뷰에 대한 분석을 기반으로 제품의 특성을 요약하여 정리해주는 기능이 가능해졌다.

텍스트 데이터를 이용한 감성 분석은 상품의 리뷰뿐만 아니라 영화, 노래 등에 대한 이용자들의 감성 분석과 식당, 호텔, 회사 등 특정 업체에 대한 감성 분석에도 활용되고 있

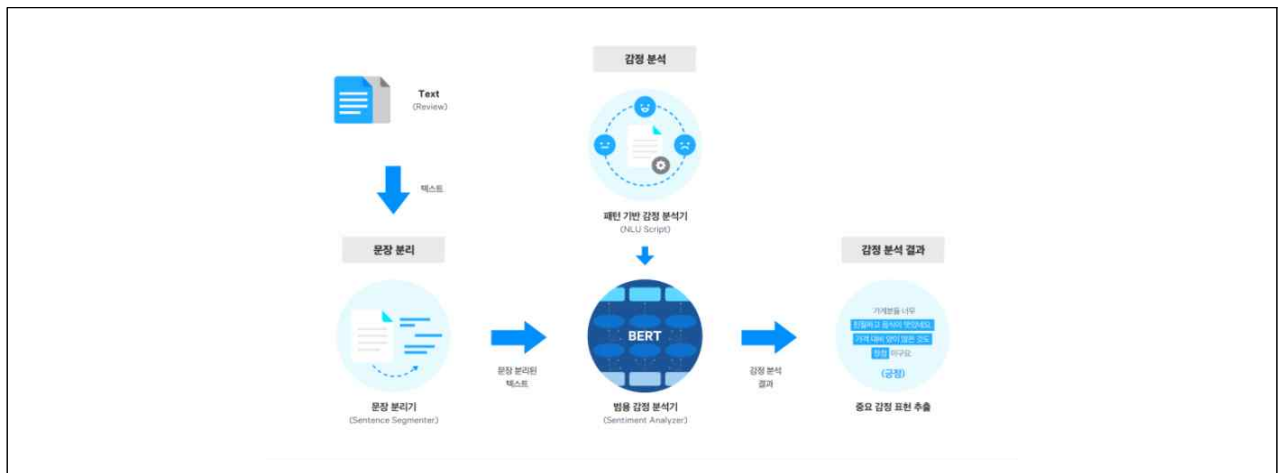
다. 더 나아가 뉴스 댓글을 활용하여 연예인, 정치인 등 특정 인물에 대한 호감도, 주요 사건 및 법안에 대한 여론 분석 등에서도 활용의 범위를 넓혀가고 있다³⁰⁾.

그림 19 하이퍼클로바의 기능이 적용된 쇼핑 리뷰 예시



이 외에도 네이버에서는 해외 사례와 마찬가지로 ‘클로바 센티먼트 (CLOVA Sentiment)’라는 한국어 텍스트의 감성 분석 서비스를 API로 제공한다. 클로바 센티먼트는 입력되는 단어/문장/문단의 내용의 긍정/부정/중립을 분석하는 서비스이다³¹⁾.

그림 20 클로바 센티먼트 API 작동 방식



국내의 또 다른 기업인 ‘엔씨소프트(NCSOFT)’는 야구 정보 제공 서비스인 ‘페이지 (Paige)’를 운영 중이다. 엔씨소프트는 페이지를 통해 수집된 댓글 데이터를 활용하여 사용자의 감성을 분석하였다. 특히 야구 관련 댓글들을 기반으로 사용자의 구단에 대한 선

30) 네이버의 공식 블로그 https://blog.naver.com/naver_search/222080132277 참고

31) <https://medium.com/naver-cloud-platform/이렇게-사용하세요-텍스트-감정-분석-서비스-구현하기-clova-sentiment-활용기> 참고

호도를 파악하는 것을 시도하였다. 야구 댓글은 은어, 문법 오류 등을 모두 포함한 구어체 문장으로 이루어져 있는데, 이러한 상황에서도 ‘긍정’, ‘부정’, ‘중립’의 감성을 분류할 수 있도록 수집된 댓글 데이터뿐만 아니라 앞서 소개한 NSMC 데이터를 함께 활용하였다. 그리고 이러한 기능을 ‘페이지’의 챗봇 기능에서 활용하고 있다³²⁾.

그림 21 야구 댓글에서의 구어체 현상 및 감성 분석

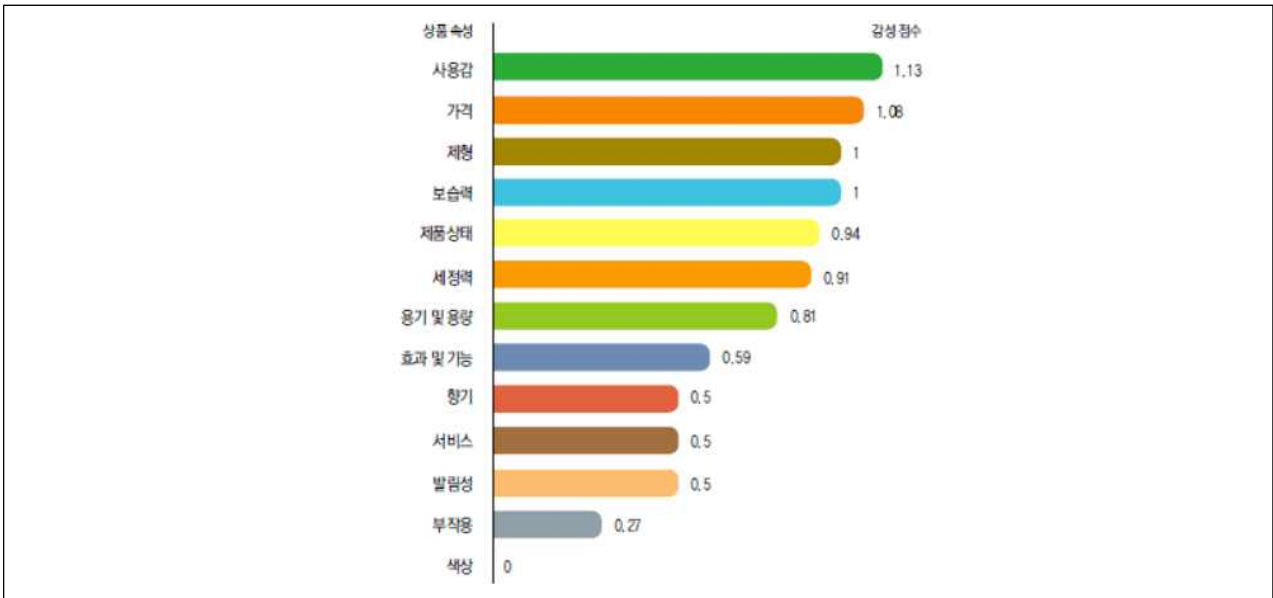


국내 화장품 기업인 ‘에뛰드하우스’에서는 상품 리뷰를 기반으로 마케팅 전략을 수립하였다³³⁾. 먼저 뷰티 관련 웹사이트에서 구매 후기를 품목별로 수집하여 감성을 분석하였다. 수집된 구매 후기를 기반으로 효과적인 감성 분석과 화장품이라는 도메인에 특화된 단어들을 잘 분석할 수 있도록 기존의 형태소 사전에 관련 단어를 첨가하는 방식으로 형태소 사전을 구축하였다. 그 뒤, 속성어 사전을 구축하는데, 속성어 사전에서는 화장품이 가질 수 있는 속성들을 등록하여 각 속성을 기준으로 후기를 분석할 수 있는 토대를 마련하였다. 그 뒤 소비자의 후기에서 상품에 대한 감성을 알 수 있는 감성어 사전을 구축하였다. 이때 감성어 사전에는 단순한 단어의 나열이 아닌, 감성의 극성 및 강도를 함께 설정하여 추후 감성을 도출하는 데에 활용하였다. 마지막으로 구축한 모든 사전을 활용하여 소비자의 후기를 이용하여 속성별 감성 점수를 도출해낼 수 있는 감성 분석 모형을 구축하였다. 그 결과 각 화장품에 대하여 아래의 그림 22와 같은 속성별 감성 점수를 얻을 수 있게 되었다. 에뛰드하우스는 이러한 자사 제품에 대한 속성별 감성 점수와 경쟁사 화장품의 속성별 감성 점수를 비교하여 전략적인 마케팅 방식을 수립하였다.

32) 공식 블로그 <https://blog.ncsoft.com/커뮤니케이션과-ai-2-사람의-말을-이해하는-페이지paige/>와 <https://blog.ncsoft.com/커뮤니케이션과-ai-6-자연어처리-extreme-setting/> 참고

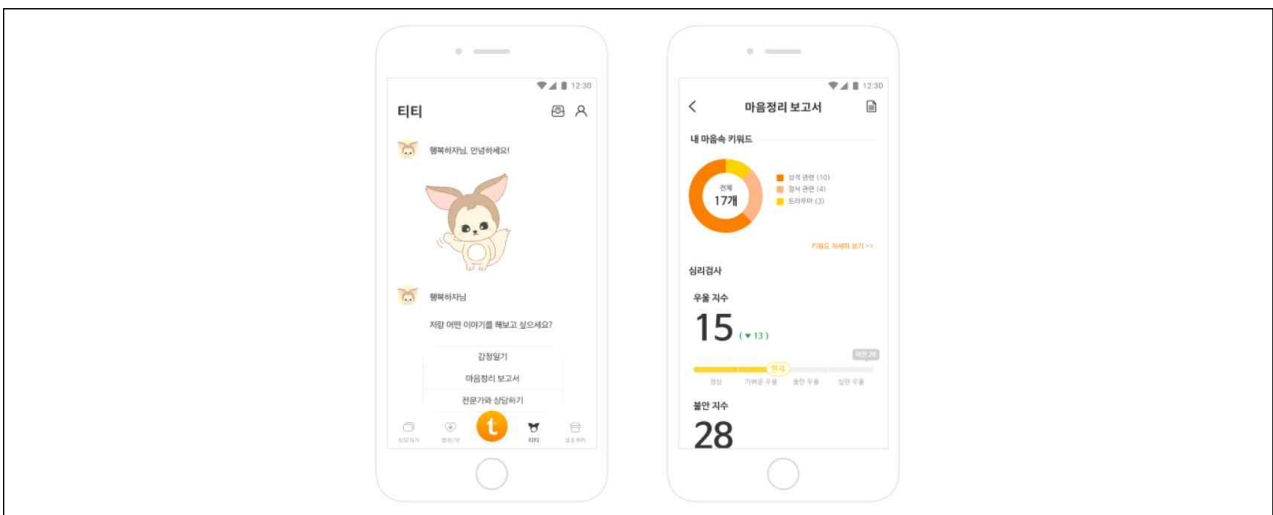
33) https://dbr.donga.com/article/view/1202/article_no/8891/ac/magazine 참고

그림 22 화장품 A에 대한 속성별 감성 점수



이러한 감성 분석 외에도 감정 분석을 기반으로 한 서비스를 제공하는 기업들이 최근 설립되기 시작하였다. 대표적인 예로 ‘트로스트’³⁴⁾는 심리 치료를 위한 감정 분석 챗봇을 제공한다. 자연어처리 기반의 챗봇 ‘티티’를 통해 사용자에게 대한 데이터를 확보하고, 챗봇과의 대화를 기반으로 사용자의 감정을 좌절, 슬픔, 사랑 등 8가지로 분석한다. 또한 사용자의 감정을 챗봇과의 대화를 통해 상시 케어하는 서비스를 제공한다.

그림 23 사용자의 감정을 분석하는 챗봇 ‘티티’



이 외에도 ‘위드마인드’³⁵⁾는 텍스트, 음성, 영상 분석을 기반으로 하여 면접을 상황을

34) <https://trost.co.kr/> 참고
 35) <https://withmind.net/> 참고

연습해볼 수 있는 어플리케이션 ‘IM’을 내놓았다. 이 어플리케이션을 통해 목소리의 톤, 크기, 발화 속도 등을 이용하여 발화의 내용뿐만 아니라 음성을 분석하고, 표정과 시선, 움직임 등을 이용하여 영상을 분석한다. 이러한 정보들을 활용하여 9가지 감정맵 기반의 감정 분석 결과를 제공한다. 또한 이러한 정보를 수집하는 인공지능 면접 영상의 멀티모달 분석을 통해 이용자의 성향 뿐만 아니라 이를 기반으로 한 직군 적합성을 도출해 낸다. ‘위드마인드’는 이러한 서비스를 제공하기 위하여 15,000여 건의 면접 영상 데이터 및 30만건의 음성 데이터를 활용하였다고 밝히고 있다.

마지막으로 ‘스캐터랩’에서는 오픈 도메인 인공지능 챗봇인 ‘이루다’ 서비스를 제공한다. 이 서비스는 100억건 이상의 한국어 카카오톡 대화 데이터를 이용하여 챗봇을 학습시켰으며, 자연스러운 발화를 위해 사용자의 대답으로부터 사용자가 느끼는 감정을 분석할 수 있도록 모델을 학습시켰다. ‘이루다’의 서비스를 제공하기 전에는 채팅 메시지 분석을 통해 사용자의 감정, 호감도 등을 알려주는 ‘텍스트앳’, ‘진저’ 등의 서비스를 제공하기도 하였다.

□ 학계

국내 연구의 경우 감정이 아닌 긍정/부정의 이진 분류를 하는 감성 분석에 관한 연구가 대다수를 차지하였다. 이를 위해 많은 연구에서 NSMC 데이터를 활용하였다. 그러나 AI HUB에서 최근 감정과 관련된 다양한 데이터를 구축하기 시작하며, 이에 관한 연구들이 점점 증가하는 추세를 보인다.

안세훈 외(2021)의 연구에서는 AI HUB의 데이터 중 ‘웰니스 대화 스크립트 데이터셋’과 ‘한국어 감정 정보가 포함된 단발성 대화 데이터셋’을 활용하여 감정분석 기반 심리상담 AI 챗봇 시스템에 관하여 연구하였다. 그러나 감정 정보가 담긴 데이터셋을 활용했음에도 불구하고, 데이터의 감정 레이블을 ‘긍정’과 ‘부정’의 레이블로 재가공하여 각각 5만개의 문장을 활용하였다. 또한 ‘웰니스 대화 스크립트 데이터셋’의 레이블 정보를 이용하여 불안의 원인을 ‘가족’, ‘성’, ‘대인관계’, ‘자기개념’의 네 가지로 분류한 뒤, 대화 속에서 감정 원인을 분류하는 연구 또한 진행하였다.

문장의 분류 연구에서는 BERT, ELECTRA, LSTM 모델을 활용하였으며, 각 모델에 따른 정확도는 그림 24과 같다. 감정의 원인 분류 문제에서는 이전 실험에서 가장 정확도가 높았던 ELECTRA 모델을 이용하여 그림 25과 같은 정확도를 얻었다.

그림 24 안세훈 외(2021)의 감성 분류 정확도

	Precision	Recall	F1-score
BERT	0.81	0.82	0.82
ELECTRA	0.82	0.84	0.83
LSTM	0.38	0.78	0.51

그림 25 안세훈 외(2021)의 감성 원인 분류 정확도

	Precision	Recall	F1-score
가족	0.85	0.81	0.83
성	0.89	0.77	0.83
대인관계	0.78	0.84	0.80
자기개념	0.87	0.79	0.82

이 외에도 이상아 외(2021)과 이현영 외(2021)은 문장의 감성이 ‘긍정’인지 ‘부정’인지를 판단하는 언어모델의 성능을 판별하기 위하여 NSMC 데이터를 활용하였다.

국내의 많은 연구가 주로 NSMC 데이터를 활용하여 감성 분석 연구를 진행하였는데, 그 이유는 첫째, 데이터의 양이 ‘긍정’과 ‘부정’ 각각 10만 개로 총 20만 개의 데이터를 활용할 수 있기 때문이다. 현재 생성되는 다양한 모델들의 성능을 평가하기 위해서는 데이터가 많을수록 성능 평가에서 더욱 정확한 결과값을 얻을 수 있다. 이 때문에 데이터의 양은 매우 중요하다. 둘째, 인위적으로 구축된 데이터가 아닌 실제 사용자들에 의해 수집된 데이터이기 때문이다. 이를 통해 문어체 문장이 아닌, 구어체 문장에 강건한 언어 모델을 만들 수 있기 때문에 실제 언어생활을 반영하는 데이터의 구축은 중요하다. 셋째, 접근이 쉽다. NSMC 데이터는 별도의 절차 없이 데이터를 활용할 수 있으며, 구축된 시점이 오래되었기 때문에 대다수 연구자가 접근하기 쉬운 데이터이다. 따라서 데이터가 제대로 활용되기 위해서는 데이터의 개방성 또한 중요한 과제이다.

2. 국어원 구축 말뭉치 관련 제언

2.1. 한국어 감정 분석 말뭉치 구축 시뮬레이션

국립국어원의 한국어 감정 분석 말뭉치 구축 방향에 대해 논의하기에 앞서, 3장에서 살펴본 국내외 감정 분석 말뭉치 구축 현황을 바탕으로 한국어 감정 분석 말뭉치 구축 시뮬레이션을 진행하고 문제점을 검토하였다. 한국어 데이터 2종에 시뮬레이션을 진행하였다. 시뮬레이션에 사용한 실제 데이터는 부록에서 확인할 수 있다.

□ 시뮬레이션 1

- 데이터: 한국어 감정 정보가 포함된 단발성 대화 데이터셋의 일부³⁶⁾
- 태그: 에크만 감정 모델(6종), 플루치 감정 모델(8종)
- 목적: 에크만, 플루치 등의 기존 감정 유형 검토
- 주석자: 성균관대 연구진 2인 (다른 데이터에 각자 태깅)

□ 시뮬레이션 2

- 데이터: 네이버 영화 리뷰(NSMC)의 일부³⁷⁾
- 태그: 에크만 감정 모델(6종), 플루치 감정 모델(8종)
- 목적: 감정 태그 부착의 일관성 유지 가능성 검토
- 주석자: 성균관대 연구진 2인 (같은 데이터에 동시 태깅)

시뮬레이션 1, 2에 대한 본 연구진의 종합적인 검토 결과를 밝히고자 한다. 먼저 감정 유형에 대한 검토이다. 시뮬레이션에서는 가장 잘 알려지고 많이 사용되고 있는 에크만과 플루치 감정 모델을 이용해서 한국어 데이터의 감정을 태깅하였다. 그러나 기존 감정 모델들로 태깅하는 데에는 몇 가지 문제점이 존재하였다. 실제 한국어 데이터 예시를 보며 어떤 문제점이 있는지 하나씩 살펴보자.

첫째, 에크만과 플루치 모델에는 긍정적인 감정 유형이 충분하지 않았다. 에크만 모델

36) '한국어 감정 정보가 포함된 단발성 대화 데이터셋'은 한국지능정보사회진흥원의 AI허브(AIHub)에 공개된 데이터로, 연구 목적인 경우 AI허브 이용자 누구나 데이터 내려받기를 신청하고 사용할 수 있다. 2021년 10월 20일에 내려받은 데이터로 시뮬레이션을 진행하였다.

(<https://aihub.or.kr/opendata/kefi-data/recognition-laguage/KETI-02-009>)

37) '네이버 영화 리뷰 데이터(v1.0)'는 박은정 박사가 개인 깃헙에 공개한 데이터로, 누구나 접근 가능하다. 2021년 10월 20일 내려받은 데이터로 시뮬레이션을 진행하였다.

(<https://github.com/e9t/nsmc>)

의 경우 ‘기쁨 happiness’과 ‘놀라움 surprise’, 플루치 모델의 경우 ‘기쁨 joy’, ‘놀라움 surprise’, ‘기대 anticipation’, ‘수용 trust’가 긍정류의 감정 유형이다. 그러나 실제 데이터상에는 ‘기쁨’ 외에는 거의 등장하지 않아, 기존 모델로는 한국어 데이터에 등장하는 긍정적인 감정을 표현하기 어려웠다. 아래의 예 1과 2는 긍정적인 감정 유형의 부족으로 인해 감정 태깅이 어려운 경우이다.

예 1) 인상적인 영화였다

예 2) 우지오빠 어떻게 노래 그렇게 잘만들고 노래도 잘해요?ㅠㅠ

둘째, 감정이 뚜렷하지 않거나 없는 경우를 표현하기 위한 감정 유형이 없었다. 감성 분석에서 긍정과 부정 외에 중립 유형이 존재하듯, 감정 분석 역시 ‘감정없음’ 혹은 ‘중립’ 유형이 필요함을 확인할 수 있었다. 아래의 예 3과 4는 감정이 뚜렷하지 않아 기존 모델로는 감정 태깅이 어려운 경우이다.

예 3) 애기들이 좋아하는 맛은 아니였어요ㅋㅋㅋ

예 4) 카톡차단하지않는다는건무슨 의미일까요??

셋째, 많은 감정 모델들이 그러하듯 에크만과 플루치 모델 역시 감정 유형간의 경계가 모호하여 태깅에 어려움을 발생시키는 경우가 빈번하였다. 예를 들어 에크만과 플루치 두 모델에 모두 포함된 ‘혐오 disgust’와 ‘분노 anger’ 유형은 경계가 모호하여 주석자간의 주석 부착 일치도를 떨어트리는 요인 중 하나였다. 아래의 예 5와 6은 이에 대한 실제 예시 데이터이다.

예 5) 평점에속지마시길시간낭비 돈낭비임

예 6) 울면서 손들고 횡단보도 건널때 뛰쳐나올뻔 이범수 연기 드럽게못해

넷째, 복합적인 감정이 드러나는 표현이 존재하였다. 둘 이상의 감정이 동시에 등장하여 하나의 감정 유형으로 태깅하는 것이 어려운 경우가 다수 존재하였다. 이때 예 7처럼 ‘속상’, ‘짜증’과 같이 관련 있는 두 감정이 동시에 나타나기도 하고, 예 8처럼 ‘만족’이라는 긍정적인 감정과 ‘속상’이라는 부정적인 감정이 동시에 나타나기도 하였다. 두 예시 유형 모두 기존 감정 모델로 태깅하기에는 어려웠다. 실제로 예 8에 대해 주석자 1은 ‘sadness’/‘sadness’로 태깅했고, 주석자 2는 ‘happiness’/‘joy’로 태깅했다.

예 7) 애를 왜 더 힘들게 만들어...시상식 참석가 그렇게 중요한가? 회복하는게 우선이지...

[속상+짜증]

예 8) 정말 맘에 들어요. 그래서 또 보고싶는데 또 보는 방법이 없네? >.. ── [만족+속상]

이처럼 단순히 기존 감정 모델을 그대로 차용하는 것이 아니라, 데이터의 도메인을 고려한 감정 유형의 설정과 태깅 방식이 필요하다고 판단된다. 예를 들어, 영화 리뷰의 경우 ‘만족 satisfied’, ‘실망 disappointed’, ‘감동 impressed’, ‘짜증 annoyed’ 등의 유형이 빈번하게 나타날 것이다. 특히 복합적인 감정을 표현해내기 위해 둘 이상의 감정에 동시에 태깅하는 다중 레이블링 방식에 대해서도 고려해야 할 것이다.

본 시뮬레이션을 통해 한국어 감정 분석 말뭉치 구축 시 위에서 논의한 감정 유형뿐만 아니라 다양한 조건들을 고려해야 함을 확인할 수 있었다. 먼저 언어표현 중 어떠한 부분의 감정을 분석할지에 따라 감정 유형이 다르게 태깅될 수 있었다. 예를 들어 예 9는 텍스트의 명제는 ‘불만족’과 같은 부정적인 감정 유형이 나타나지만, 텍스트에서 기술하는 대상인 영화에 대해서는 ‘만족’이라는 긍정적인 감정을 갖는다.

예 9) 왜케 평점이 낮은건데? 꽤 불만한테.. 헐리우드식 화려함에만 너무 길들여져 있나?

뿐만 아니라, 문맥 부족 등의 이유로 문장을 완벽하게 이해할 수 없어서 감정을 태깅하는 것이 어려운 경우도 존재하였다. 아래의 예 10과 11은 단순히 궁금증을 드러내는 표현인지, 비꼬는 표현인지, 불만족 혹은 짜증의 표현인지 불분명하다.

예 10) 키이라 나이틀리가 연기하고자 했던건 대체 정신장애일까 틱장애일까

예 11) 4대강 22조는 어디다 쓴거냐?

위에서 살펴본 것과 같이 감정 정보 부착을 위해서는 데이터 문장에 대한 명확한 이해와 수집 대상의 선별 과정이 필요하다. 본 시뮬레이션을 통해 한국어 감정 분석 말뭉치 구축 시 다양한 조건들을 고려해야 함을 확인할 수 있었다.

2.2. 국어원 감정 분석 말뭉치 구축 방향성 제안

본 사업에서는 국내외 감정 분석 말뭉치 구축 현황을 바탕으로 국립국어원의 감정 분석 말뭉치 구축 방향성을 제안한다. 아래의 표 25와 같이 크게 데이터 수집과 레이블 부

착의 측면으로 나눠 검토하였고, 이를 차례로 논의하고자 한다.

표 25 한국어 감정 분석 데이터 구축 시 고려 사항

데이터 수집	1. 도메인·주제
	2. 활용 분야
	3. 구축량
레이블 부착	4. 감정 유형
	5. 부착 단위
	6. 부착 방식

2.2.1. 데이터 도메인·주제

데이터 수집을 위해서는 개인 정보 비식별화, 차별 및 혐오적인 표현 제거 등의 과정을 반드시 수행해야 하기 때문에 현실적으로 수집 가능한 언어 데이터의 유형이 한정적인 것이 사실이다. 특히 감정 분석 말뭉치의 경우 감정이 겉으로 드러나는 도메인과 주제를 선정하는 것이 중요하다. 또한 데이터의 클래스 균형을 위해서는 어느 특정 감정만 등장하는 것이 아니라 다양한 감정이 등장할 수 있는 데이터를 선정하는 것이 매우 중요하다.

3.1장을 통해 국내외에서 구축된 감정 분석 데이터는 어떤 도메인과 주제의 언어 데이터에 기반하는지 살펴보았다. 그 결과 트위터, 페이스북 등의 SNS 텍스트, 블로그 글, 뉴스 기사 제목과 본문, 대화 등의 데이터가 사용되었다. 그 중에서도 대화와 SNS 텍스트가 가장 많이 사용되었다.

본 사업에서는 인간과 인간 간의 일상 대화나 특정 상품 또는 서비스에 대한 소비자의 리뷰가 가장 적합한 데이터 도메인과 주제일 것이라고 제안한다. 일상 대화의 경우 일반 상식에 관한 대화 혹은 우울증, 트라우마 등의 상담 대화 등이 감정 데이터 구축에 적합하다고 판단된다. SNS 텍스트의 경우 많은 사람들로 부터 생산된 짧은 텍스트를 많이 수집할 수 있기 때문에 특정 상품 및 서비스에 대한 리뷰 텍스트를 수집하는 것이 알맞을 것이다.

2.2.2. 데이터 활용 분야

한국어 감정 분석 말뭉치는 다양한 분야에 활용될 수 있을 것으로 기대된다. 이를 위해 범용적으로 사용 가능한 감정 분석 말뭉치를 구축하는 것이 이상적일 수 있으나, 이는

현실적으로 어렵다. 왜냐하면 감정 레이블의 종류, 데이터의 크기 등의 데이터 구축 방식에 따라 실제 연구에 적용하기 까다로운 경우가 존재하기 때문이다. 예를 들어 우울증을 예방 목적의 챗봇을 개발하기 위해 감정 데이터를 학습 데이터로 활용하는 경우, 해당 데이터는 우울증을 겪고 있거나 비슷한 상태에 있는 사람이 느낄 수 있는 감정이 무엇인지 세밀하게 표현하고 있어야 한다. 반면에 일반적인 감성 대화 챗봇을 개발하기 위해서는 사용자가 기계와 나누는 대화에서 주로 어떤 감정을 표출하는지에 대한 분석이 선행되어야 한다. 이처럼 데이터를 활용하는 분야에 따라 이에 사용하는 데이터의 양상이 다르기 때문에 데이터의 활용 분야를 명확히 하는 것은 매우 중요하다.

3.2장에서 살펴본 국내외의 다양한 감정 데이터 활용 사례와 연구 현황을 고려할 때 데이터의 활용 분야로 제안하는 첫 번째는 대화 데이터를 활용한 감성 대화 챗봇 개발이다. 최근 코로나바이러스감염증-19의 사태로 인해 비대면 의사소통의 비중이 높아지고 이와 함께 챗봇에 대한 관심이 더 늘어나고 있다. 특히 SKT에서 외부활동을 하지 못하고 집에만 있는 자가격리자 혹은 타인과의 교류가 적은 독거 노년층을 위한 챗봇이 개발되면서 단순히 기계와 질의응답을 주고받는 것이 아니라, 마치 인간과 감정을 교류하듯 감성적인 대화가 가능한 챗봇을 개발하려는 시도가 많아지고 있다. 상대방의 긍정적인 감정에 칭찬이나 동조 등의 반응할 수 있는 혹은 부정적인 감정에 격려나 위로를 할 수 있는 챗봇을 개발하기 위해서는 사용자의 감정을 정확하게 파악해야 한다. 이를 위해서는 감정이 태깅된 대화 혹은 구어체 데이터가 매우 필요하다.

두 번째 활용 분야는 특정 상품 또는 서비스에 대한 SNS 리뷰 텍스트를 활용하는 연구이다. 당사의 제품에 대해 소비자가 만족했는지 불만족했는지 파악하고, 만족 또는 불만족했다면 어떤 점에 대해 그러한 감정을 갖게 되었는지 파악할 수 있다면, 해당 회사는 그 제품에 대한 마케팅을 효과적으로 개선할 수 있을 것이다. 또한 당사 브랜드 혹은 상품에 대한 언어표현으로부터 부정적인 감정을 빠르게 분류해낸다면 당사의 브랜드 안정성 도모에 큰 도움이 될 것으로 보인다. 더 나아가 당사 제품 또는 서비스에 대한 사용자의 리뷰와 감정 정보는 무인 콜센터 혹은 제품·서비스 상담 챗봇 개발에도 유용하게 활용될 것이다.

위와 같은 자연어처리 기반의 상품 및 서비스 개발뿐만 아니라, 자연어처리 언어모델의 성능 평가에도 활용될 수 있을 것이다. 최근 BERT, GPT-3처럼 대용량의 일반 텍스트 데이터로 사전 학습(pre-training)하고, 주어진 태스크와 관련된 데이터로 미세 조정(fine-tuning)하여 언어모델을 개발하는 추세가 증가하고 있다. 인간이 사용할 프로그램에 적용될 언어모델을 개발하고 평가하는 데에 인간의 감정 정보가 반영된 데이터를 활용한다면 프로그램의 성능을 향상시키고 관리하는 데에 유용할 것으로 판단된다.

2.2.3. 데이터 구축량

국립국어원의 한국어 감정 데이터 구축을 위해 수집해야 할 데이터의 양을 제안하기 위해 본 사업에서는 여러 국내외 감정 데이터의 크기를 분석하였다. 그러나 부록을 통해 확인할 수 있듯이 국내외에서 구축한 감정 분석 데이터의 크기는 서로 매우 상이하다. 기구축된 한국어 감정 데이터를 보면 가장 적은 경우 약 15,000개의 문장을 포함하고 있지만, 많은 경우 약 27만 개의 문장을 갖기도 한다. 기구축된 감정 데이터로부터 한국어 감정 데이터의 적절한 양을 가늠하는 것은 어려웠다. 특히 수집한 언어표현이 모두 감정을 갖는 것은 아니며, 각 감정 유형들이 언어 데이터에 골고루 등장하는 것은 아니기 때문에 전체 데이터의 크기를 기준으로 데이터를 수집한다면 추후 데이터 불균형의 문제가 예상된다.

이 때문에 본 사업에서는 전체 데이터의 크기가 아닌, 각 클래스별 데이터의 개수를 제안하고자 한다. 각 감정 유형마다 약 5,000개의 문장을 수집하는 것이 적절하다고 판단된다. 하지만 감정 유형에 따라 자주 등장하는 것도 있고, 가끔 등장하는 것도 있기 때문에 유형별로 정확하게 동일한 개수의 데이터를 수집하는 것은 어렵다. 대신 심각한 데이터 불균형을 방지하기 위해 가장 많은 인스턴스를 갖는 감정 유형과 가장 적은 인스턴스를 갖는 감정 유형 간의 비율을 사전에 정의하여 제한하는 것도 하나의 방법일 것이다. 물론 이를 위해서는 다음 장에서 설명 감정 유형 설정에 대한 깊은 분석이 이루어져야 할 것이다.

2.2.4. 감정 유형

어떤 감정 유형을 수집한 언어 데이터에 부착할지와 관련하여 크게 3가지의 방안을 제안해볼 수 있다. 각각에 대해 자세히 논의해보고자 한다.

□ 기존 감정 모델의 적용

에크만 모델은 감정을 별개의 범주들로 간주해야 한다고 가정한다. 에크만 모델에서는 인간의 기본적인 감정을 ‘기쁨(happiness)’, ‘공포(fear)’, ‘슬픔(sadness)’, ‘놀람(surprise)’, ‘혐오(disgust)’, ‘분노(anger)’의 6가지 감정으로 분류한다. 이 모델에서는 감정이 서로 다른 강도를 가질 수는 있지만 섞이는 것은 허용하지 않는다. 이러한 이유로 에크만 모델은 많은 비난을 받았으나 기본 감정이 다른 감정과 어떻게 구별될 수 있는지를 명확하게 정의하였기 때문에 감정 분석 모델을 구축하는 데에 자주 사용된다.

이에 반해 플루치 모델은 기본 감정을 기반으로 모든 감정들이 서로 혼합되고 다양한 조합을 파생시킨다. 플루치 모델은 ‘놀람(surprise)’, ‘기대(anticipation)’, ‘기쁨(joy)’, ‘슬픔(sadness)’, ‘분노(anger)’, ‘공포(fear)’, ‘신뢰(trust)’, ‘혐오(disgust)’의 8가지 유형으로 감정을 분류한다. 그리고 이러한 감정의 유형을 수레바퀴에 비유하여, ‘감정 수레바퀴(wheel of emotions)’ 모델을 제안하였다. 이 모델 내에서 유사한 감정들은 서로 가깝게 배치되어 있고, 반대의 감정은 반대편에 배치되어 있다. 또한 가운데로 갈수록 감정의 강도가 강해진다. 플루치 모델은 바퀴 구조에 대한 경험적인 근거가 없다는 점에서 비판을 받기도 하였으나, 감정 분류의 연구에 많은 영향을 끼쳤다.

에크만과 플루치 모델은 공신력 있는 감정 모델이므로 연구자들의 합의를 이끌어낼 수 있다는 장점이 있다. 하지만 단점 역시 존재하다. 앞서 한국어 감정 데이터 구축을 위한 시뮬레이션에서 밝힌 바와 같이 에크만과 플루치 모델에는 긍정적인 감정 유형이 부족하다. 각 모델에서 가지고 있는 감정 유형들은 그 경계가 모호하며, 복합적인 감정을 갖는 언어표현에는 태깅하기 어렵다는 문제가 있다.

□ 기존 감정 모델의 확장

기존 감정 모델의 한계를 보완하기 위해 이를 확장하여 사용하는 방법이 있다. 표 26은 플루치 감정 모델을 긍정, 부정, 중립으로 이루어진 감성 유형과 연계한 모델이다. 이처럼 감성과 감정을 연계한 확장 모델을 만들어, 예를 들어 같은 ‘놀라움’이라는 감정 유형으로 태깅되더라도 반가운 놀라움인지, 불쾌한 놀라움인지 구분할 수 있도록 감정 모델을 확장하는 것이다. 그러한 경우 표 27의 예문들을 구분하고, 언어학적으로 깊게 이해할 수 있을 것으로 기대된다.

표 26 감성-감정 연계 모델

대분류(감성)	소분류(감정)
긍정	기쁨, 기대, 놀라움, 수용, 슬픔
부정	슬픔, 분노, 공포, 혐오, 놀라움
중립	중립

표 27 감성-감정 연계 모델을 적용한 예시

문장	감성	감정
바빠서 모임에 못 올 줄 알았는데 어떻게 왔어!	긍정	놀라움
갑자기 그렇게 나타나서 너 땀에 간 떨어질 뻔 했잖아!	부정	놀라움

또 다른 확장 방법은 기존 감정의 세부 감정들을 정의하는 것이다. 플루치크는 8개의 기본 감정과 함께 이것들이 2개씩 서로 합쳐져 새로운 감정을 만든다고 주장하였다. 1차 이중 감정은 서로 인접한 두 감정들 사이에서 발생하며, 자주 느낄 수 있는 감정들이다. 2차 이중 감정은 바퀴 그림에서 한 개의 꽃잎을 사이에 두고 떨어져 있는 감정들의 결합으로 발생하며, 가끔씩 느낄 수 있는 감정들이다. 3차 이중 감정은 바퀴 그림에서 두 개의 꽃잎을 사이에 두고 떨어져 있는 감정들의 결합으로 발생하며, 아주 가끔 느낄 수 있는 감정들이다. 그림 26의 왼쪽 그림은 플루치크가 밝힌 8개의 기본 감정과 강도에 따른 구체적 유형들을 표현한 감정 바퀴이다. 오른쪽 그림은 플루치크의 1, 2, 3차 이중 감정을 표현한 것이며, 표 28은 1, 2, 3차 이중 감정의 예시와 결합 과정을 나타낸 것이다.

그림 26 플루치크의 감정 바퀴(좌), 플루치크의 1, 2, 3차 이중 감정(우)

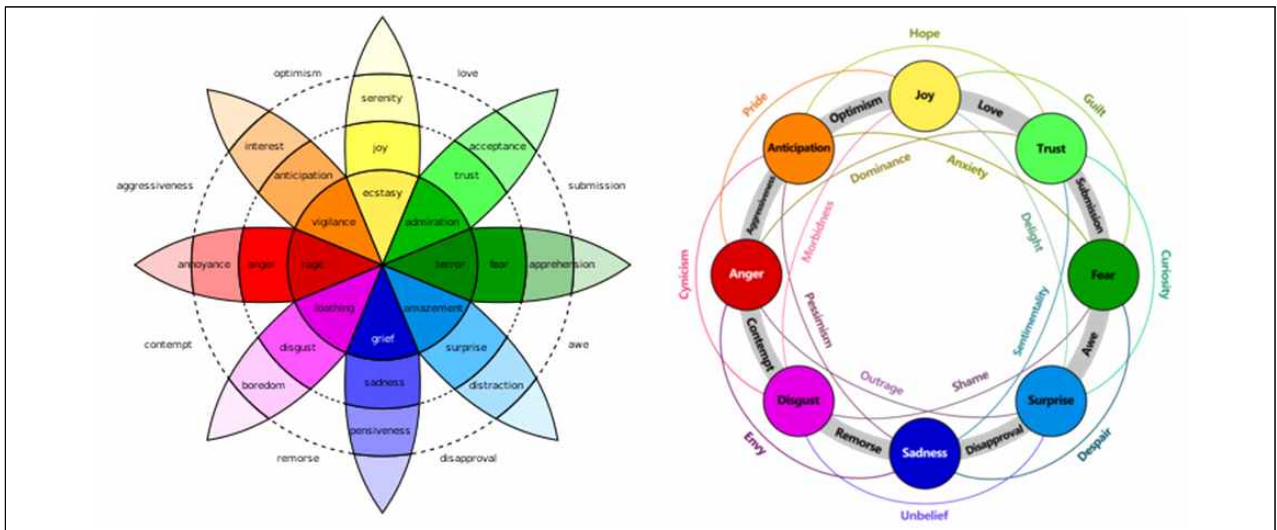


표 28 플루치크의 1, 2, 3차 이중 감정 예시

1차 이중 감정			2차 이중 감정			3차 이중 감정		
기본감정 1	기본감정 2	이중감정	기본감정 1	기본감정 2	이중감정	기본감정 1	기본감정 2	이중감정
Anger	Anticipation	Aggressive-ness	Anger	Joy	Pride	Anger	Trust	Dominance
Anticipation	Joy	Optimism	Anticipation	Trust	Hope	Anticipation	Fear	Anxiety
Joy	Trust	Love	Joy	Fear	Guilt	Joy	Surprise	Delight
Trust	Fear	Submission	Trust	Surprise	Curiosity	Trust	Sadness	Sentiment-ality
Fear	Surprise	Awe	Fear	Sadness	Despair	Fear	Disgust	Shame
Surprise	Sadness	Disapproval	Surprise	Disgust	Unbelief	Surprise	Anger	Outrage
Sadness	Disgust	Remorse	Sadness	Anger	Envy	Sadness	Anticipation	Pessimism
Disgust	Anger	Contempt	Disgust	Anticipation	Cynicism	Disgust	Joy	Morbidness

플루치 모델의 기본 감정 8가지 외에 1차 이중 감정 등을 세부 감정으로 추가하여 기존 감정 모델을 확장한다면, 공신력 있는 기존 감정 모델을 차용하면서 기존 모델만으로는 설명하기 어려운 언어표현에도 감정 유형을 태깅할 수 있을 것이다. 그러나 레이블의 개수가 늘어남에 따라 주석자 간의 일치도가 떨어질 가능성이 존재한다.

□ 신규 감정 유형의 제안

에크만, 플루치 등의 감정 모델은 모두 외국인 화자의 언어적·비언어적 사용에 대한 분석을 토대로 만들어진 것이다. 한국어 감정 데이터 구축을 위해 감정 유형을 설정하는 세 번째 방안은 한국어 표현 고유의 감정 유형을 제안하는 것이다. 감정 유형에 대한 기존 외국 연구 결과를 번역하여 사용하지 않고, 한국어 사용에서 나타나는 고유의 감정을 찾아 한국어 감정 모델을 정립하는 것이다. 이러한 경우 한국어 표현 고유의 감정을 분석하기 때문에 매우 경쟁력 있는 데이터를 구축할 수 있다는 장점은 있지만, 많은 노력이 필요하고 모든 연구자의 동의를 이끌어내기 어려우며 주석자 간의 높은 일치도를 기대하기 어렵다는 문제가 존재한다.

여러 가능성을 비교 분석한 결과 본 사업에서는 기존 감정 모델을 확장하여 감정 유형들을 정의하고, 이를 이용하여 한국어 데이터의 감정을 태깅하는 것을 제안한다.

2.2.5. 레이블 부착 단위

정의한 감정 유형을 수집한 언어 데이터의 어느 부분에 부착할지와 관련하여 크게 3가지의 방안을 고려해볼 수 있다. 첫째, 문서 단위로 감정 레이블을 부착하는 방법이다. 여러 문장으로 구성된 문서가 언어 데이터로 주어진 경우 전체 텍스트에 전반적으로 나타나는 감정이 무엇인지 판단하여 레이블링하는 방법이다. 이 경우 수집한 언어 데이터에 대한 전처리 작업이 비교적 적고 대화 혹은 텍스트의 진행에 따른 전체 감정 맥락을 파악할 수 있다는 장점이 있지만, 말차례 혹은 문장의 개수가 많아지면 하나의 감정으로 정의하는 것이 어렵다는 단점이 있다.

둘째, 문장 단위로 감정 레이블을 부착하는 방법이다. 수집한 언어 데이터가 여러 문장으로 구성된 경우 문장 단위로 분절하여 문장마다 감정을 분석하여 태깅하는 것이다. 이 경우 문서 단위보다는 여러 감정이 동시 등장할 확률이 낮기는 하지만, 여전히 여러 감정이 동시에 등장할 수 있으며, ‘중립’ 감정 유형의 데이터가 월등하게 많아질 가능성이 존재한다. 특히 대화의 경우 문장의 경계가 불분명하여 감정을 부착할 문장 단위로 분절

하는 것이 어렵다는 문제점도 존재한다.

셋째, 요소(aspect) 단위로 감정 레이블을 부착하는 방법이다. 이는 텍스트 혹은 문장에서 감정과 관련된 특정 표현에 감정 레이블을 부착하는 방법으로 구체적이고 세밀한 감정 분석이 가능하다는 특징이 있다. 이로 인해 다양한 태스크로의 확장 가능성이 크다. 하지만 도메인별 요소 정의와 정확한 요소 추출이 선행되어야 한다는 단점이 있다.

본 사업에서는 대화 데이터의 경우 문장 단위의 감정 레이블링을, 리뷰 데이터의 경우 요소 단위의 감정 레이블링을 제안한다. 대화에서의 감정 분석은 일반적으로 화자의 기분이나 상태를 파악하기 위함이기 때문에, 문장 단위의 감정 유형 레이블링이 적합하다. 반면 상품 및 서비스에 대한 리뷰의 감정 분석에서는 고객이 무엇을 만족하고 불만족하는지 파악하는 것이 중요하기 때문에, 요소 단위의 감정 유형 레이블링이 적합하다고 판단된다.

2.2.6. 레이블 부착 방식

정의한 감정 유형을 언어 데이터에 어떻게 부착할지와 관련하여 크게 2가지의 방안을 제안해볼 수 있다. 첫째, 각 단위에 하나의 감정 레이블을 부착하는 단일 레이블링(single labeling)이다. 감정 유형의 정의가 분명하고 유형 태깅에 대한 교육이 잘 이루어진다면 단일 레이블링은 주석자 간의 비교적 높은 일치도를 기대해볼 수 있을 것이다. 그렇지만 언어표현이 갖는 복합적인 감정을 표현하는 것이 어렵다는 문제가 존재한다.

둘째, 각 단위에 둘 이상의 감정 레이블을 동시에 부착하는 다중 레이블링(multi labeling) 방식이 있다. 이는 단일 레이블링의 문제를 보완할 수 있는 방안이다. 다중 레이블링은 다음의 2가지 방법으로 수행해 볼 수 있을 것이다. 먼저 유형별로 해당 감정의 등장 여부를 유무로 표현하는 방법이다. 앞서 표 16에서 본 SemEval 2018 Task1 E-c 데이터가 감정을 레이블링한 방법이 그 예이다. 또 다른 방법은 복합 감정이 등장할 때 유형별로 비중을 수치로 표현하는 것이다. 예를 들어, 표 27에서 본 ‘바빠서 모임에 못 올 줄 알았는데 어떻게 왔어!’ 문장은 ‘기쁨(60%)’과 ‘놀라움(40%)’으로 표현할 수 있을 것이다. 이처럼 감정의 비중으로 다중 레이블링하는 경우 언어표현이 갖는 주된 감정과 부수 감정을 구분지어 이해할 수 있다는 장점이 있다. 하지만 주석자 간의 주석 일치도를 보장하기 매우 어렵다는 한계가 존재한다.

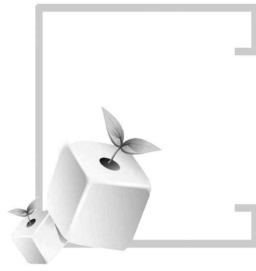
본 사업에서는 등장 여부를 유무로 표현하는 다중 레이블링 방식으로 언어표현의 감정을 태깅하는 것을 제안한다. 이는 복합 감정을 이해할 수 있는 풍부한 언어자료의 토대가 될 것이며, 다양한 활용 가능성을 가질 것으로 기대된다.

2.2.7. 종합

본 사업에서는 아래의 표 29와 같이 국립국어원의 한국어 감정 분석 데이터 구축 방안을 제안하고자 한다.

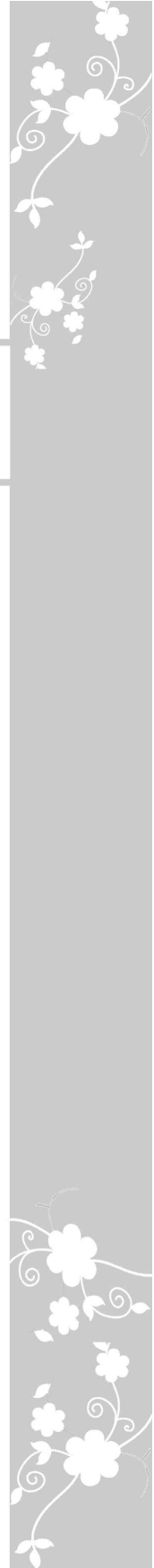
표 29 한국어 감정 분석 데이터 구축 방안 제언

데이터 수집	1. 도메인·주제	<ul style="list-style-type: none"> • 일상 대화 • 상품/서비스 리뷰
	2. 활용 분야	<ul style="list-style-type: none"> • 자연어처리 기반의 상품/서비스 개발 ✓ 감성대화 챗봇 개발 ✓ 상품/서비스 마케팅 개선 • 자연어처리 언어모델의 성능 평가
	3. 구축량	<ul style="list-style-type: none"> • 레이블 당 약 5,000개
레이블 부착	4. 감정 유형	<ul style="list-style-type: none"> • 기존 감정 모델 확장(플루치 모델)
	5. 부착 단위	<ul style="list-style-type: none"> • 문장 단위 • 요소 단위
	6. 부착 방식	<ul style="list-style-type: none"> • 다중 레이블(유/무)



제 4 장

결 론



1. 사업 요약

본 사업의 목적은 4차 산업 혁명에 대비하여 인공지능 기술의 개발 및 활용을 위한 대규모 말뭉치를 분석하여 국어 자원의 활용도와 가치를 제고하는 것에 있다. 또한 국내 표준화 및 참고 지반 자료가 될 수 있는 정밀한 언어 정보 부착 말뭉치를 분석 및 배포를 목적으로 진행되었다.

1.1. 말뭉치 감성 분석 지침 수립 및 말뭉치 구축

감성 분석 세부 지침 수립 및 감성 분석 말뭉치 구축은 문서 분석 및 문장 분할을 기초로 주요 도메인을 선정하고, 주요 도메인 관련 테스트 데이터를 추가 구축하여 기본 문서와 테스트 데이터를 대상으로 속성 기반 감성 분석 말뭉치를 구축하는 절차로 진행되었다. 그 결과 총 2,081개 문서와 추가 수집한 210개 문서를 대상으로 분석한 감성 분석 말뭉치를 구축하였으며 그 결과는 다음과 같다.

그림 27 2021 감성 분석 말뭉치 구축 결과

분석 데이터	제외	영화	제품			여행			합 계
		영화	화장품세정제	전자기기	제품기타	숙소	식당	관광지	
20년도 말뭉치데이터	- (27)	3,243 (379)	2,924 (776)	693 (123)	2,551 (640)	262 (27)	26 (35)	713 (74)	10,652 (2,054)
추가 구축 데이터	- (6)	1,064 (69)	1,002 (67)	1,542 (68)	-	-	-	-	3,608 (204)
합 계		4,307(448)	3,926(843)	2,235(191)	2,551(640)	262(27)	266(35)	713(74)	14,260(2,258)

※ 문장 수(문서 수)

본 사업에서 구축한 감성 분석 말뭉치의 규모는 총 2,258문서, 14,260문장이다. 20년도 감성 분석 말뭉치와 추가 수집 데이터 내 문서 중 소수 도메인 문서와 지나치게 많은 리뷰 대상을 다루는 문서는 분석 대상에서 제외하였다. 그 결과 20년도 말뭉치 데이터에서 27건, 추가 구축 데이터에서 6건의 문서가 제외되어 각각 2,054건, 204건의 문서가 분석되었다.

본 사업에서는 비교적 큰 규모를 이루고 있는 주요 3개 세부 도메인을 대상으로 블로그 문서를 추가로 수집하여 테스트 데이터를 구축하였다. 테스트 데이터 추가 구축의 목표 데이터량은 각 세부 도메인당 500문장이다. 본 사업에서는 최종적으로 영화 1,064문

장, 화장품·세정제 1,002문장, 전자기기 1,542문장을 분석함으로써 목표 규모를 충족하는 테스트 데이터를 구축하였다.

1.2. 국내외 동향 연구 및 제안

본 연구에서는 국내외 감정 및 감성 분석에 대한 문헌 자료 연구를 통해, 국외 13건의 감성 분석 말뭉치 연구 자료와 16건의 감정 분석 말뭉치 자료를 검토하였으며, 국내에서 수행된 각 5건의 감성 및 감정 분석 말뭉치 문헌 자료를 토대로 국내외 감정 및 감성 분석 현황을 파악하였다. 검토 내용을 요약하면 다음과 같다.

표 30 국내외 감정 및 감성 분석 말뭉치 주요 구축 현황

구분	구축 말뭉치 데이터	구축 데이터	레이블 개수	
국외	감성	IMDb (Internet Movie Database)	영화 리뷰 50,000건	2
		SST-2 (Stanford Sentiment Treebank)	영화 리뷰 11,855건	2
		SST-5 (Stanford Sentiment Treebank)	영화 리뷰 11,855건	5
		SemEval-2016 Task 4: Sentiment analysis in Twitter	트위터 메시지 30,000건	3
		SemEval-2016 Task 5: Aspect-Based Sentiment Analysis	구매평 약 6,000개 문장	4
	감정	EmoInt	트위터 메시지 7,097건	4
		SemEval 2018 Task 1: Affect in Tweets	트위터 메시지 12,600건	4
		SemEval 2019 Task 3: EmoContext Contextual Emotion Detection in Text	인간-기계 대화 38,424건	4
국내	감성	국립국어원 감성 분석 말뭉치 2020	SNS, 블로그 게시물 2,081건	5
		네이버 영화 리뷰(NSMC)	댓글 200,000개 문장	2
	감정	한국어 감정 정보가 포함된 단발성 대화 데이터셋	댓글 38,594개 문장	7
		한국어 감정 정보가 포함된 연속성 대화 데이터셋	대화 세트 10,000건 (55,627개 문장)	7
		감정 분류를 위한 대화 음성 데이터셋	음성 대화 4차(14,606개 문장), 5차(29,385개 문장)	7
		감성 대화 말뭉치	사람-챗봇 대화 270,000개 문장	60
		웰니스 대화 스크립트 데이터셋	사람-챗봇 대화 5,232개 문장	61

표 31 국내외 감성 및 감정 분석 말뭉치 주요 활용 현황

구분	활용 사례		기관 및 연구자
국외	산업계	텍스트 기반 감정 분석 엔진 'SALLEE'	캐나다 스타트업 기업 'Receptiviti'
		음성 감정 분석기 'Web Empath API, 감정 대화 인공지능 'Beluga Box'	일본 스타트업 기업 'Empath'
		인공지능 스피커 '에코(Echo)', 인공지능 플랫폼 '알렉사(Alexa)'	미국 기업 '아마존(Amazon)'
	학계	Ana at semeval-2019 task 3: Contextual emotion detection in conversations through hierarchical lstms and bert.(2019)	Huang, Chenyang, Amine Trabelsi, and Osmar R. Zaiane.
		Fine-grained emotion classification of Chinese microblogs based on graph convolution networks.(2020)	Lai, Yuni, et al.
		Distributed representations of emotion categories in emotion space.(2021)	Wang, Xiangyu, and Chengqing Zong.
국내	산업계	대규모 인공지능 '하이퍼클로바(HyperClova)', 텍스트의 감성 분석 서비스를 API '클로바 센티먼트 (CLOVA Senttiment)'	네이버(NAVER)
		야구 정보 제공 서비스인 '페이지(Paige)'	엔씨소프트(NCSOFT)
		상품 리뷰 기반 마케팅 전략 수립	에뛰드하우스
		심리 치료 목적 감정 분석 챗봇 '티티'	트로스트
		면접을 상황 대비 어플리케이션 'IM	워드마인드
		오픈 도메인 인공지능 챗봇 '이루다'	스캐터랩
	학계	감정분석 기반 심리상담 AI 챗봇 시스템에 대한 연구(2021)	안세훈, 정옥란
		감정 분석을 위한 BERT 사전학습모델과 추가 자질 모델의 결합(2021)	이상아, 신호필
		Out-of-Vocabulary 단어에 강건한 병렬 Tri-LSTM 문장 임베딩을 이용한 감정분석	이현영, 강승식

또한, 실효성 있는 연구 결과를 위해 한국어 단발성 대화 데이터셋 200건과 네이버 영화 리뷰 100건을 대상으로 실제 감정 데이터 구축 시뮬레이션 진행을 수행하였다.

연구의 결과 향후 국어원의 감성 분석 연의 도메인 및 주제로 인간 간의 일상 대화 또는 특정 상품·서비스에 대한 소비자 리뷰가 적합할 것으로 보인다. 일상 대화 텍스트는 일반 상식에 대한 대화 혹은 우울증인 트라우마 관련 상담 대화가 감정 데이터를 구축하는 것에 용이할 것으로 판단된다. 상품·서비스에 대한 소비자 리뷰 데이터는 다수의 소비자가 생산한 짧은 길이의 데이터를 수집할 수 있는 분석에 SNS 텍스트를 이용하는 것이 적합하다.

본 연구에서는 전체 데이터 크기가 아닌 클래스별 데이터의 규모를 제안하고자 한다. 각 감정 유형마다 5,000문장을 기준으로 볼 수 있다. 하지만 감정 유형에 따른 등장 빈도 차이가 있으므로 감정 유형별로 동일한 규모의 데이터를 수집하는 것에 어려움이 있을 수 있다. 이러한 불균형을 방지하기 위해 최대, 최소 인스턴스의 감정 유형 간 비율을

사전에 정의하여 제한하는 것이 하나의 방법이 될 수 있다. 이를 위해서는 감정 유형 설정에 대한 깊은 분석을 필요로 한다.

2. 사업의 의의 및 기대 효과

‘2021년 말뭉치 감성 분석 및 연구’ 사업을 통한 기대 효과는 다음과 같다.

□ 감성 분석 말뭉치 구축 기대 효과

- 민간에서 활용 가능한 국가 공공재로서의 말뭉치 확대 구축 및 국어 자원의 활용도와 가치 향상에 기여
- 4차 산업혁명 대비 기반 기술 개발 및 인공지능 기술 개발, 활용을 위한 대규모 말뭉치 구축으로 국어 자원의 활용도와 가치 제고
- 민간 공유를 통해 언어 인공지능 등 관련 산업 활용을 위한 기반을 마련하고 국어 및 국어문화 연구, 국어정책 수립의 기초 자료로 활용

본 사업과 함께 국립국어원에서 추진하고 있는 국어 말뭉치 구축 사업을 통해 인공지능 스피커, 대화형 로봇, 로봇 개인 비서 등 한국어 인공지능의 성능을 향상시킬 것으로 기대되며, 향후 4차 산업혁명 시대의 인공지능 서비스 개발 및 기술 혁신을 위한 중요 자료가 될 전망이다.

<부록1> 말뭉치 감성 분석 지침

속성 기반 감성 분석 지침

<목 차>

1. 속성 기반 감성 분석	01
1) 속성 기반 감성 분석의 개요	01
2) 속성 기반 감성 분석의 대상	01
3) 속성 기반 감성 분석 정보의 유형과 주석 방법	02
2. 도메인별 세부 분석 지침	08
1) 영화 도메인	08
2) 제품 도메인	11
3) 여행 도메인	16

2021 말뭉치 감성 분석 및 연구: 속성 기반 감성 분석 지침

1. 속성 기반 감성 분석

1) 속성 기반 감성 분석 개요³⁸⁾

감성 분석(Sentiment Analysis)이란 언어에서 나타나는 주관성을 포착하여 그것이 내포하는 긍정 또는 부정의 감성을 평가하여 분석하는 작업임. 그중 속성 기반 감성 분석(ABSA, Aspect-Based Sentiment Analysis)은 대상이 가지고 있는 다양한 속성에 대한 감성을 보다 구체적으로 분석함. 예를 들어 “이 영화는 액션이 화려해서 볼 만했어”라는 문장을 단순히 ‘영화’에 대한 긍정 평가가 아닌 영화의 ‘영상·시각 요소’에 대한 긍정으로 분석함으로써 보다 세밀한 화자의 감성을 분석하는 것이 가능함.

2) 속성 기반 감성 분석 대상

(1) 화자의 긍정적 또는 부정적 주관어 드러나는 문장은 분석 대상임. 화자의 긍정 및 부정 감성의 주관성이 드러나는 문장 예시는 아래와 같음.

예 이번 포뇨 역시 히사이시 조의 음악이 첫장면부터 분위기를 압도했다.

영화 <벼랑 위의 포뇨> 리뷰

예 저도 참보는 제품인데 요거 완전 대박이에요 πππ

화장품 ‘돌피너스크림’ 리뷰

예 정말 장점을 열거해보니, 넘나 매력적인 냉장고 맞네요 ^^

전자기기 ‘LGDIOS얼음정수기냉장고’ 리뷰

(2) 단순히 제품의 기능 또는 효능에 대해 언급하더라도 다음 예시와 같이 화자의 주관어 드러나는 감성 표현이 있는 문장이라면 분석 대상으로 봄.

예 천연식물로 추출한 천연재료로 향균, 소취기능, 피부보호기능까지 있어 안심하고 사용하네요~

세정제 ‘라베르샤세제’ 리뷰

예 휠다이얼버튼으로 온오프에 바람세기 조절까지 되는데 정말 손쉽고 간편하더라구요!!!

전자기기 ‘마르세이유에어튜너드라이기’ 리뷰

(3) 주관성이 드러나지 않은 문장(사실의 나열, 타인의 견해 인용, 해쉬 태그 단순 나열 등)은 속성 기반 감성 분석 대상이 아님. 구체적인 예시는 아래와 같음.

※ 단, 문장의 중간에 해쉬 태그가 들어가는 경우는 분석 가능함.

예 영국작가 이완 맥이언의 소설 ‘어톤먼트’를 영화화한 작품. → 사실의 나열

38) 본 연구에서는 「SemEval-2016 Task 5: Aspect-Based Sentiment Analysis」의 속성 기반 감성 분석 방법론을 기반으로 하여 한국어 특성에 맞는 문장 감성 분석을 하였음. 「SemEval-2016 Task 5: Aspect-Based Sentiment Analysis」에서는 상품 및 서비스에 대한 리뷰를 대상으로 7개 도메인, 약 6000문장의 속성 기반 감성 분석이 이루어짐.

영화 <어톤먼트> 리뷰

- 예 시크릿 폴더라고해서 자체 폴더숨김 기능도 탑재하고 있더라고요 → 사실의 나열
전자기기 ‘삼성 노트북 9 Pen’ 리뷰
- 예 올 마미는 이게 무슨 핑크냐고 너 가지고있는거랑 별 차이 모르겠다고 → 타인의 견해 인용
화장품 ‘맥립스틱’ 리뷰
- 예 우리 남편이 애정하는 #우공사 #찜질기 ♥□ → 타인의 견해 인용
전자기기 ‘우공사 찜질기’ 리뷰
- 예 #eclado#에끌라두션크림#간증선크림#케어선크림#간증선크림#인생선크림#수분크림인가선크림인
가#오일0% #육아스타그램#나들이필수품#ad → 해쉬 태그 단순 나열
화장품·세정제 ‘에끌라두션크림’ 리뷰

(4) 단일 문장 안에서 감성이 명확히 드러나지 않거나 긍·부정 판단이 어렵더라도 문서 전체 문맥 상 해당 문장이 감성을 내포하고 있다면 분석 대상으로 간주함.

- 예 어른들은 어른들 나름대로 너무 좀 뻘한 감동이라 애매하다는 눈치들이었구요. 아 이게 감동적이긴 한데 너무 뻘하니깐 감동해주긴 좀 뻘쭙한 느낌?이라고나 할까요. 저 역시도 그런 느낌이었구요.
→ 해당 문장만 고려한다면 긍·부정 감성을 알 수 없지만 문맥 상 부정의 감성임을 인지할 수 있으므로 분석 대상임.

3) 속성 기반 감성 분석 정보의 유형과 주석 방법

(1) 도메인(Domain)

본 연구에서는 웹 말뭉치의 리뷰 문서를 영화, 제품, 여행 총 3가지 유형의 도메인으로 분류하여 분석함. 이는 다시 7개의 세부 도메인으로 나누어지며 문장의 수를 기준으로 하여 주요 도메인 3개, 기타 도메인 4개로 분류하였음. 자세한 내용은 <표 1>과 같음. 이는 ‘2021 말뭉치 감성 분석 및 연구’의 분석 대상 문서 총 2,081개에서 나타나는 개체(entity), 속성(attribute) 등을 경험적으로 분석한 결과에 기초함. 따라서 분석 대상 문서의 구성 방식에 따라 도메인은 더 일반화되거나 세분화될 수 있음.

<표 1> 도메인 구성

도메인	세부 도메인		리뷰 대상
영화	주요 도메인 39)	영화	추격자, 원티드, 과속스캔들 등
제품		화장품·세정제	알파 캐비어 크림, 덴티스테치약 등
	기타 도메인	전자기기	클레온무선칫솔살균기, 미로가습기 등
제품 기타		대디베이비기저귀, 롬페르 크로스백 등	
숙소		시그니엘호텔, 아이엠게스트하우스 등	
여행	기타 도메인	식당	제주돌담집, 월정리갈비밥 등
		관광지	평강랜드, 광안리해수욕장 등

39) 문장 수(데이터의 양)을 기반으로 하여 영화, 화장품·세정제, 전자기기를 주요 도메인으로 상정하였음. 나머지 기타 도메인의 경우 데이터의 양이 충분하지 않아 도메인별 세부 분석 지침에서 제시하는

(2) 개체#속성 쌍(E#A 쌍)

① 개체(Entity)

가. 개체의 개념

개체란 도메인별로 화자의 긍정 또는 부정 감정이 부여되는 대상을 유형화한 것임. 문서 주제에 해당하는 영화 작품 또는 제품과 그와 관련된 구성 요소 등이 개체가 될 수 있음.

나. 개체의 태깅 방식

화자가 긍정 또는 부정의 감정을 가지는 대상을 찾고 이에 해당하는 개체를 태깅함.

예 음악 또한 그 영상에 맞게 최고만을 들려준다.

도메인: 영화 / 개체: 음악·음향

예 발색이 너무 예쁘데~ 지속력과 촉촉함도 대박!!

도메인: 화장품·세정제 / 개체: 본품

예 건조기는 우선 사고불일이네요 ㅎㅎ

도메인: 전자기기 / 개체: 전자기기

② 속성(Attribute)

가. 속성의 개념

‘속성’은 개체가 가지는 다양한 속성 중 사용자가 긍정 혹은 부정의 극성 감정을 가지게 되는 속성을 말함.

나. 속성의 태깅 방식

개체 분류 이후 해당 개체가 가지는 속성 중 하나를 선택함. 감성 분석 정보는 개체와 속성을 묶어 개체#속성 쌍의 형태로 문장에 부착됨. 이때 분석 문장 내에서 개체#속성 쌍 분석의 기준이 될 만한 감성 대상을 찾을 수 없다면 문서의 전체 문맥을 고려하여 판단함. 개체#속성 쌍의 분석 예시는 아래와 같음.

예 마스크킹 레이어링 애플은 메디힐에서 새로 출시한 애플로..애플겝도 편리해서 휴대,보관도 용이하네요!

도메인: 화장품·세정제 / 개체#속성 쌍: 패키지·구성품#편의성

예 김의성과 조우진, 정만식 등의 연기력도 두말할 것 없이 좋았다.

도메인: 영화 / 개체#속성 쌍: 배우#연기력

예 거기에 잡티커버까지!! 완전 대박~~

도메인: 화장품·세정제 / 개체#속성 쌍: 본품#품질

③ 의견 대상 표현(Opinion Target Expression, OTE)

개체#속성 쌍의 예시가 다양하게 나오지 않을 수 있음.

가. OTE의 개념

OTE란 문장 내에서 '개체'로서 나타나는 객관적 표현(명시적 지시 대상)임.

나. OTE의 태깅 방식

ㄱ. 긍부정 감성 평가의 대상이 되는 명사 또는 명사구를 최장 범위로 원문에서 추출함. 이때 개별 분석의 OTE 영역이 중첩될 경우 해당 영역의 중복 추출을 허용함.

예 베리 맘 크림은 보습력이 너무 좋아서 한번만 발라줘도 장시간 촉촉해요 🍷🍷🍷

문장 분석: {본품#품질, 베리 맘 크림}

예 김하늘의 능청스러운 연기는 몇 년전이지만 이미 꽃을 피웠고 강동원은 신인급의 풋풋함이 좋았다.

문장 분석: {배우#연기력, 김하늘}, {배우#연기력, 강동원}

예 다들 셀 수 없이 많은 사람들을 죽인 악당들임에도 불구하고 영화 후반에 갑자기 동료애가 넘쳐나서 함께 힘을 합쳐 싸우게 되는 장면이 인상적이다.

문장 분석: {영상/시각 요소#일반, 다들 셀 수 없이 많은 사람들을 죽인 악당들임에도 불구하고 영화 후반에 갑자기 동료애가 넘쳐나서 함께 힘을 합쳐 싸우게 되는 장면}

예 귀여운 토끼 캐릭터가 악당과 싸우는 장면이 제일 인상적이었다.

문장 분석: {캐릭터#일반, 긍정, 토끼 캐릭터}, {영상·시각 요소#일반, 긍정, 토끼 캐릭터가 악당과 싸우는 장면}

ㄴ. OTE 영역 전후에 인접한 해쉬 태그 등의 기호나 이모티콘은 추출하지 않음(단 OTE 영역 내부에 포함될 경우 분석의 편의를 위해 포함하여 추출함)

예 완전 신기한 #AI카메라 📷

문장 분석: {제품 전체#일반, AI카메라}

예 나 역시 <다크나이트>와 크리스토퍼 놀란의 팬이다.

문장 분석: {영화#일반, 다크나이트}, {감독#일반, 크리스토퍼 놀란}

예 화장이 곱도는 요즘 내 피부 원장님크림으로 유명한 #백아울 #힐링크림 사용해요

문장 분석: {제품 전체#인지도, 긍정, 백아울 #힐링크림}

ㄷ. 감성의 대상이 되는 표현이 문장 내에서 명시적으로 드러나지 않는 경우 'Null' 라벨을 부여함

예 여자보다 더 이쁜 거 같다.

배우 / 문장 분석: {배우#일반, Null}

예 양치후 깔끔하고 입안 가득 상쾌해요

치약 / 문장 분석: {본품#품질, Null}

ㄹ. 문장의 맥락과 의미를 고려하여 개체 유형에 속하는 대상을 OTE로 함.

예 흔들의자도 너무 좋고 그 집에 여러개의 의자 다 갖고 싶었다.

문장 분석: {영상·시각 요소#일반, 긍정, 흔들의자}, {영상·시각 요소#일반, 긍정, 그 집에 여러개의 의자}

예 근데 터치가 적응이 안 돼.

문장 분석: {본품#일반, 부정, 터치}

㉟. 개체를 직접적으로 나타내지 않는 지칭어·지시어 등의 간접 언급은 OTE로 추출하지 않음.

예) 일명 #치약계의샤넬 을 손쉽게 만나볼수있다니 곳곳
문장 분석: {제품 전체#일반, Null}

예) #국내최초 #터키산 #장미오일 로 만들어져서 .. #봄여왕 을 만들어주는 너님 내가 애정한다 진심..
문장 분석: {제품 전체#일반, Null}

예) 아직 이 사람보고 연기를 잘하네 뭐 이런 말 하기는 팬이라도 낮간지럽지만 최근 영화를 쭉 보면 계속 발전하고 있다는 걸 느낄 수 있다.
문장 분석: {배우#연기력, Null}

예) 이 영화는 그야말로 총체적 난관이다
문장 분석: {영화#일반, Null}

㉞. 의미상의 개체와 문장에서 감성의 대상이 되는 표면적인 OTE가 상이할 경우 OTE와 개체가 일치하지 않을 수 있음.

예) 조선 시대의 부패한 정치를 바로잡는 강동원은 보는 이에게 통쾌감을 준다.
문장 분석: {캐릭터#일반, 조선 시대의 부패한 정치를 바로잡는 강동원}
- 의미상 개체 유형: 캐릭터
- 의미상 OTE: 캐릭터의 이름
- 표면상 OTE: 조선 시대의 부패한 정치를 바로잡은 강동원

예) 넘 더운 요즘~ 촉촉하고 산뜻한 #아이소이 #워터풀크림라이트!
문장 분석: {본품#품질, 아이소이 #워터풀크림라이트}
- 의미상 개체 유형: 본품
- 의미상 OTE: 본품 이름
- 표면상 OTE: 아이소이 #워터풀크림라이트

예) #일렉트로룩스 #익스프레서니스트무선주전자 는 디자인이 예뻐서 어디뉘도 이쁨 작렬~!
문장 분석: {본품#디자인, 일렉트로룩스 #익스프레서니스트무선주전자}
- 의미상 개체 유형: 본품
- 의미상 OTE: 본품의 설명
- 표면상 OTE: 일렉트로룩스 #익스프레서니스트무선주전자

㉝. '-것, -음, -기'와 같은 명사절은 OTE로 보지 않음

예) 초호기 몸에 녹색부분이 더 많아지고 야광기능까지 추가된 것이 멋지다.
문장 분석: {캐릭터#연출력, 긍정, Null}

예) 썸머세트라고 해서 세럼 + 선크림 이렇게 패키지로 파는 것 괜찮음
문장 분석: {패키지·구성품#일반, 긍정, Null}

㉜. OTE는 맥락과 의미를 고려하여 최장 범위로 추출하는 것을 원칙으로 함. 단, 최장 범위 중에서 화자의 주관적인 의견에 해당하는 수식 어구는 제외함.

예) 화면 가득히 펼쳐지는 바닷속 풍경에 웅장하면서도 아름다운 오케스트라 음악은 그야말로 환상적.
문장 분석: {음향·음악#연출력, 오케스트라 음악}

예) 유럽에서 대히트치고 한국으로 넘어온 귀한 크림이라고
문장 분석: {제품 전체#인지도, 크림}

④ 의견 극성(Opinion Polarity)

가. 의견 극성의 개념

의견 극성이란 화자가 개체 및 속성에 대해서 나타내는 감성의 극성을 의미함. 의견 극성은 긍정, 중립, 부정의 3단계로 나뉨. 이때 '중립'은 약한 긍정이나 약한 부정을 의미함. 긍정 또는 부정의 감성이 드러나지 않는 주관적 표현은 분석하지 않음.

나. 의견 극성의 태깅 방식

먼저 문장에 개체#속성 쌍 라벨을 부여한 뒤 의견 극성을 판단하여 태깅함. 한 문장에서 동일한 OTE와 개체#속성 쌍에 대해 긍정과 부정이 상충하는 경우는 '중립'으로 태깅함. 분석 태깅 정보를 모두 부착한 분석 예시는 아래와 같음.

- 예) 그야말로 헐리우드 영화 스타일이라고 해야할까.
→ 긍정 또는 부정의 감성이 드러나지 않는 주관적 표현이므로 제외 처리
- 예) 주황색 들어가서 뭔가 더 선크림스럽
→ 긍정 또는 부정의 감성이 드러나지 않는 주관적 표현이므로 제외 처리
- 예) CG 면에서는 흠 잡을 데 없었다.
문장 분석: {영상-시각요소#일반, 긍정, CG 면}
- 예) 영화 보고 나온 뒤 기분만 더러워지는 올해 최악의 영화.
문장 분석: {영화#일반, 부정, 영화}
- 예) 넘 기대는 하지말고 가시면 괜찮아요
문장 분석: {영화#일반, 중립, Null}
- 예) 약간 완전 묶은 크림느낌은 아니고 적당한 크림질감
문장 분석: {본품#일반, 중립, Null}

⑤ Out Of Scope

가. Out Of Scope의 개념

Out Of Scope는 상정하지 않은 개체나 리뷰 대상 외의 타 대상에 대한 감성이 드러나는 문장을 말함

ㄱ. 상정하지 않은 개체에 대한 감성이 드러나는 문장

본 연구에서 상정하지 않은 개체에 대해서 긍부정 감성을 드러내는 문장은 Out Of Scope로 처리한 뒤 의견 극성만 분석함.

- 예) 미셀라 클렌징워터 `퓨레테` 라인 이름 넘 어려운거 아님!?
→ '제품명' 개체가 없으므로 {Out Of Scope, 부정} 처리
- 예) 오히려 이런 상태여서, 영화에 더 심취할 수 있었던 것 같다.다행히 즐지 않고 관람할 수 있었는데, 이 진귀한 경험. 나름대로 좋았다.
→ '관람 경험' 개체가 없으므로 {Out Of Scope, 긍정} 처리
- 예) 그렇다면 역시 45요금제는 하나 쓰잘데기가 없다.
→ '요금제' 개체가 없으므로 {Out Of Scope, 부정} 처리

ㄴ. 리뷰 대상 외의 타 대상에 대한 감성이 드러나는 문장

문서의 리뷰 대상이 아닌 타 대상을 단독으로 언급하며 긍부정 감성을 드러내는 문장은 Out Of Scope로 처리한 뒤 의견 극성만 분석함. 그 외 Out Of Scope 및 문장 제외 처리 관련 세부적인 지침은 아래와 같음.

- 주 리뷰 대상 도메인에 해당되지 않는 대상에 대한 감성 문장은 제외 처리함.
 - ☞ 휴우.. -.; 너무 피곤해서 돌아오는 길에 크리스피 도넛 들려서 더즌 사서 일행에게 반 주고 반박스 들고 왔다.
아아.. 설탕 덩어리..
근데 이거 또 오랜만에 먹으면 맛있다...
의류 세부 도메인 리뷰 문서 中
→ 제품 도메인에 식품은 해당되지 않으므로 해당 문장은 분석하지 않고 제외 처리함
- '타 대상'란 문서의 주 리뷰 대상이 되는 단일 제품 또는 단일 작품 이외의 동일 도메인 내 타 제품, 타 작품을 말함.
 - ☞ 게다가 주인공은 예전에 재미있게 봤던 `불량공주 모모코(원제: 시모츠마 이야기)`에서 후카다 료코와 함께 주연으로 나왔던 츠치야 안나라 하니.
→ 타 대상을 단독으로 언급하며 긍정 평가를 하는 문장이므로 Out Of Scope 처리
문장 분석: {Out Of Scope, 긍정}
- 문장 내에서 타 대상을 언급하더라도 화자의 의견이 해당 문서의 주 리뷰 대상과 관련 있다면 일반 분석함
 - ☞ `화려한 휴가`보다는 좀 더 낮지만 영화 속 내용 전개 부분에 일부 아쉬움이 있었다.
→ 타 대상을 언급했지만 주제 영화에 대한 평가이므로 일반 분석함
문장 분석: {시나리오#구성력, 중립, 영화 속 내용 전개 부분}
 - ☞ `옹박`이나 `똥양공`을 재밌게 봤다면 이 영화도 재밌게 볼 수 있을 것이다.
→ 타 대상을 언급했지만 주제 영화에 대한 평가이므로 일반 분석함
문장 분석: {영화#일반, 긍정, null}
 - ☞ 입이 열릴정도로 강한 치약들은 코가짱 맵고 몸이 안좋은 기분인데 온가족이 다사용할 #치약 으로도 딱 인듯!
→ 타 대상을 언급했지만 주제 제품에 대한 평가이므로 일반 분석함
문장 분석: {제품 전체#일반, 긍정, null}
- 한 문서 내에서 여러 대상을 다루고 있을 경우 가장 비중이 높은 한 가지 대상을 선택하여 분석을 진행함. 3개 이상의 대상을 동일 비중으로 다루고 있는 경우 해당 문서는 분석 대상에서 제외함.
 - ☞ < 2008 최고의 영화 베스트5 >
 1. 다크나이트 (8월, 서울극장)
 - 더 이상의 설명은 생략한다.
 - 아무 말도 필요없다.
모든 면에서 최고였다.
 - 2. 아이언 맨 (4월, CGV 용산)
 - 다크나이트와 정 반대의 느낌으로 최고.
 - 오락영화는 이래야만 한다.

3. 월-E (8월, 메가박스 목동)

- 세기말적인 풍경에서 아름다운 이야기.

로봇에 이렇게 예쁜 감정을 담을 수 있을까.

(생략)

→ 3개 이상의 대상을 동일 비중으로 다루고 있는 문서이므로 제외 처리함

2. 도메인별 분석 지침

1) 영화 도메인

영화 도메인 문서는 작성자가 개별 영화 작품 감상 후 작성한 웹 게시물 리뷰글임. 총 456 문서, 13829문장을 분석의 대상으로 함.

(1) 개체

<표 2> 영화 도메인의 개체

구 분	내 용
영화	개별 영화 작품 자체
시나리오	대사, 스토리, 전개, 패러디, 유머, 소재, 결말 등 내용 관련 요소
배우	배우, 성우, 캐스팅 관련 요소
캐릭터	등장인물 및 캐릭터 관련 요소
감독	감독 개인 관련 요소
음향·음악	영화 내 음향 효과 및 삽입곡 관련 요소
영상·시각 요소	특정 장면 및 영화의 시각적 요소(세트, 소품, CG, 미술, 액션 등)

(2) 속성

<표 3> 영화 도메인의 속성

구 분	내 용
① 일반	개체 자체
② 구성력	연속성을 가지는 시나리오의 구성(짜임새)
③ 연기력	배우의 연기 평가
④ 인지도	유명도, 인기 등
⑤ 연출력	각 부문 연출 평가

(3) 개체#속성 쌍

<표 4> 영화 도메인의 개체#속성 쌍

	일반	구성력	연기력	인지도	연출력
영화	○			○	○
시나리오	○	○			
배우	○		○	○	
캐릭터	○				○
감독	○			○	○
음향·음악	○			○	○
영상·시각 요소	○				○

(4) 세부 도메인별 예시

<표 5> 영화 도메인 분석 예시

개체	속성	예시	
영화 전체	일반	문장	나쁘다고 욕하진 않겠지만 추천하고 싶지 않은 영화.
		분석	{영화#일반, 부정, 영화}
	인지도	문장	`첫사랑`이라는 소재로 인기몰이를 한 영화죠.
		분석	{영화 전체#일반, 부정, 영화}
	연출력	문장	어쩜 범죄수사물이 이렇게까지 긴장감이 없고 멍멍할 수가 있는지...
		분석	{영화 전체#연출력, 부정, 범죄수사물}
시나리오	일반	문장	스토리 자체는 뻔하지만 나쁘지 않다.
		분석	{시나리오#일반, 중립, 스토리 자체}
	구성력	문장	사건 전개는 긴박함이 스크린에서 눈을 떼게 하지 못한다.
		분석	{시나리오#구성력, 긍정, 사건 전개}
배우	일반	문장	아오이 유우는 정말 청순미 넘치는 마스크였다.
		분석	{배우#일반, 긍정, 아오이 유우}
	연기력	문장	영조 송강호의 연기는 정말이지 엄청나게 멋지다.
		분석	{배우#연기력, 긍정, 영조 송강호의 연기}
	인지도	문장	요즘 핫한 대세배우!!!! 우리 박보검배우님 ㅋㅋㅋㅋ
		분석	{배우#인지도, 긍정, 박보검배우님}
캐릭터	일반	문장	이번 작품의 감초는 역시 `바비`의 연인 `켄`.
		분석	{캐릭터#일반, 긍정, 바비`의 연인 `켄`}
	연출력	문장	역시 요즘 시대의 슈퍼맨이라면 이정도로는 싸워야지!
		분석	{캐릭터#연출력, 긍정, 요즘 시대의 슈퍼맨}
감독	일반	문장	스티브 소머즈 감독은 휴식과 공부가 필요해 보인다.
		분석	{감독#일반, 부정, 스티브 소머즈 감독}
	인지도	문장	감독은 니나가와 미카라고 하는 유명 사진작가라는데...
		분석	{감독#일반, 긍정, 니나가와 미카}
	연출력	문장	그야말로 리들리 스콧 감독의 탁월한 연출이 돋보이는 작품이다.
		분석	{감독#연출력, 긍정, 리들리 스콧 감독의 탁월한 연출}

음향·음악	일반	문장	두 성우가 들려주는 OST는 감동 그 자체다.
		분석	{음향·음악#일반, 긍정, 두 성우가 들려주는 OST}
	인지도	문장	전설의 명곡들이 배경음악으로 나오는 것은 좋았다.
		분석	{음향·음악#인지도, 긍정, 명곡들}
	연출력	문장	특히나 이 영화에서 분위기를 이끌어가는 사운드는 주요포인트.
		분석	{음향·음악#연출력, 긍정, 사운드}
영상·시각 요소	일반	문장	놓치면 안되는 중요한 장면!
		분석	{영상·시각 요소#일반, 긍정, 장면}
	연출력	문장	CG를 대폭 사용한 화려한 풍경은 예술이 따로 없다.
		분석	{영상·시각 요소#연출력, 긍정, 풍경}

※ 세부 분석 기준

- 영화 도메인의 상정하지 않은 개체 예시(Out Of Scope 처리)
 - 예) 뭔가 적합한 제목이 없었으려나.(제목)
 - {Out Of Scope, 부정}
 - 예) 디즈니에서 나오는 3D 애니메이션은 전부 재미있다. (제작사 및 제작 과정)
 - {Out Of Scope, 긍정}
 - 예) 가장 우려했던 `츠나미(つなみ)`를 `해일`이라고 제대로 번역해줘서 다행.(번역)
 - {Out Of Scope, 긍정}
 - 예) 이런 좋은 영화를 미묘한 포스터와 미묘한 카피프라이즈 때문에 망칠 뻔 했다.(홍보)
 - {Out Of Scope, 부정}
- 캐릭터와 배우를 혼용하여 표현하는 문장은 의미 기준으로 분석하는 것을 기본으로 하되 OTE는 텍스트 기준으로 추출함.
 - 예) 주인공인 헤이든 크리스텐슨은 잘생기긴 했지만 어딘가 좀 혼한 얼굴이란 느낌이 든다.
 - 문장 분석: {캐릭터#일반, 중립, 주인공인 헤이든 크리스텐슨}
 - 예) 재미있한테 뭔가 유아인만 또라이로 나오는것 같아서 억지스러운 느낌을 받아서 0.5점 뺐어요.
 - 문장 분석: {캐릭터#연출력, 부정, 유아인}
- 단일 대사, 결말, 소재, 내용 자체 등에 대한 평가는 `시나리오#일반`, 내용의 전개, 흐름, 스토리 구성 등 연속성을 가지는 시나리오 평가는 `시나리오#구성력`으로 분석함.
 - 예) 어떤 건 너무 의도적으로 웃기려고 한 대사라 전혀 우습지가 않았다.
 - 문장 분석: {시나리오#일반, 부정, 대사}
 - 예) 스토리 라인이 너무 빈약하더군.
 - 문장 분석: {시나리오#구성력, 부정, 스토리 라인}
- 주관 표현이 없는 관객 수 언급은 단순 사실이므로 제외 처리함
 - 예) 베테랑 관객수 1200만이 넘었더라고요.
 - 문장 분석: 제외
- 작품의 미개봉 속편에 대한 기대는 분석 대상으로 함
 - 예) 속편이 제작된다면 좀 더 스펙타클한 대결을 기대한다.

문장 분석: {영화#일반, 긍정, 속편}

예) 꼭 좀 속편을 만들어주길.

문장 분석: {영화#일반, 긍정, 속편}

2) 제품 도메인

제품 도메인 문서는 작성자가 제품을 구입 및 사용한 후 작성한 웹 게시물 제품 리뷰글이며 리뷰 대상에 따라 화장품·세정제, 전자기기, 제품 기타 세부 도메인으로 분류됨. 총 1626문서, 19124문장을 분석의 대상으로 함.

(1) 개체

<표 6> 제품 도메인의 개체

구분	내용
① 제품 전체	제품 전체
② 본품 ⁴⁰⁾	특정 장치 및 부품 등 본품 하위 구성 요소
③ 패키지·구성품	제품 패키지 및 구성품(브러쉬, 펌프 등)
④ 브랜드	브랜드 이미지

(2) 속성

<표 7> 제품 도메인의 속성

구분	내용
① 일반	개체 자체
② 가격	제품의 경제적, 마케팅 요소 평가
③ 디자인	형태, 색상 등 제품 외형 디자인 평가
④ 품질	제품의 기능, 성능, 성분, 소재, 효과 등 평가
⑤ 편의성	사용 난이도 및 편의, 사이즈 언급 등 휴대성 평가
⑥ 다양성	제품 구성품의 다양성 및 유형, 색상에 따른 제품 다양성
⑦ 인지도	유명도, 인기, 판매량 등 언급

(3) 개체#속성 쌍

<표 8> 제품 도메인의 개체#속성 쌍

	일반	가격	디자인	품질	편의성	다양성	인지도
제품 전체	○	○	○	○	○		○
본품	○		○	○	○	○	
패키지·구성품	○		○	○	○	○	
브랜드	○	○	○	○			○

(4) 세부 도메인별 예시

40) 본 연구에서 제품 도메인의 '본품' 개체는 제품의 하위 구성 요소를 말한다. 예컨대 제품이 카메라인 경우 카메라는 '제품 전체' 개체, 렌즈는 카메라 본품의 하위 구성 요소이므로 '본품' 개체에 속한다. 렌즈, 충전기, 리모컨과 같이 본품의 용도와 관련하여 사용 시 반드시 필요한 구성 요소는 '본품' 개체에 속한다. 하지만 케이스, 파우치, 설명서, 연동 어플리케이션 등 본품의 기본적 용도로 사용되지 않는 부가 구성품 및 액세서리는 '패키지·구성품' 개체에 속한다.

① 화장품·세정제

<표 9> 화장품·세정제 세부 도메인 분석 예시

개체	속성	예시	
제품 전체	일반	문장	#립오일 까지 넘 맘에 드네염 ^^ ..
		분석	{제품 전체#일반, 긍정, 립오일}
	가격	문장	가격 부담도 없어서 선물하기에도 좋은 #아미솅비누 ^^
		분석	{제품 전체#가격, 긍정, 아미솅비누}
	인지도	문장	승무원밤으로 유명한 제품이에요☺
		분석	{제품 전체#인지도, 긍정, 제품}
본품	일반	문장	벨벳틴트추천 드릴 수밖에 없는 플럼?복숭아향이 난답니다.
		분석	{본품#일반, 긍정, 플럼?복숭아향}
	품질	문장	온천수 성분이 들어서 피부에 순하게 작용 할 것 같아요.
		분석	{본품#품질, 긍정, 온천수 성분}
	편의성	문장	스킨,로션 번거롭게 바를 필요없이 하나로 사용하는 올인원! 제품이에요.
		분석	{본품#편의성, 긍정, 하나로 사용하는 올인원! 제품}
다양성	문장	색상도 총 6가지로 다양하게 고를 수 있는데요!	
	분석	{본품#다양성, 긍정, 색상}	
패키지·구성품	일반	문장	퍼프부터 특별한 #압솔뤼쿠션
		분석	{패키지·구성품#일반, 긍정, 퍼프}
	디자인	문장	로고가 촌스럽지 않고, 세련되게 들어갔어요.
		분석	{패키지·구성품#디자인, 긍정, 로고}
	품질	문장	모공 브러쉬도 있어서 더욱 구석구석 세안력도 좋아§
		분석	{패키지·구성품#품질, 긍정, 세안력}
편의성	문장	펌핑형식으로 되어있어서 사용하기 편해요~~	
	분석	{패키지·구성품#편의성, 긍정, 펌핑형식}	
다양성	문장	4가지 패키지를 내맘대로 골라쓰는 재미까지!	
	분석	{패키지·구성품#다양성, 긍정, 4가지 패키지}	
브랜드	일반	문장	베네퍼트 사랑합니다.
		분석	{브랜드#일반, 긍정, 베네퍼트}
	품질	문장	전성분 모두공개 착한 성분으로 믿고쓸수있는 산다화☺
		분석	{브랜드#품질, 긍정, 산다화}
	인지도	문장	이종석 화장품으로 유명한 #셀레뷰 홈쇼핑 런칭했어요~!
		분석	{브랜드#인지도, 긍정, 셀레뷰}

② 전자기기

<표 10> 전자기기 세부 도메인 분석 예시

개체	속성	예시	
제품 전체	일반	문장	3주간 정말 잘 사용하고 있는 물걸레 로봇청소기 추천
		분석	{제품 전체#일반, 긍정, 물걸레 로봇청소기}
	가격	문장	가성비 좋은 노트북 찾으신다면 레노버 요가북 추천드려요!
		분석	{제품 전체#가격, 긍정, 레노버 요가북}
	디자인	문장	삼성노트북 끄아악 디자인 정말 예술임!!
		분석	{제품 전체#디자인, 긍정, 삼성노트북}
	품질	문장	그치만 요 작은아이가 실력은 상상이상이니 기대해보세요!
		분석	{제품 전체#품질, 긍정, Null}
	편의성	문장	LED 기기에 건전지를 넣어주고 안경처럼 착용만 해주면 끝!
		분석	{제품 전체#편의성, 긍정, LED 기기}
	인지도	문장	#조인성칫솔 로 핫한 #쿨샤전동칫솔 로 바뀌어요!
		분석	{제품 전체#인지도, 긍정, 쿨샤전동칫솔}
본품	일반	문장	구석구석 청소하기 좋은 3in1 콤비네이션 브러쉬가 들어있는데요
		분석	{본품#일반, 긍정, 3in1 콤비네이션 브러쉬}
	디자인	문장	전면은 일반 노트북처럼 깔끔합니다.
		분석	{본품#디자인, 긍정, 전면}
	품질	문장	사운드는 말할 것도 없이 엄지척이애용 ㅎㅎ
		분석	{본품#품질, 긍정, 사운드}
	편의성	문장	조작 버튼도 단순해서 인지가 어려운 저희 엄마도 잘 사용하시더라고요 ㅎㅎ
		분석	{본품#편의성, 긍정, 조작 버튼}
다양성	문장	노즐들이 다양하게 들어있는데용 :)	
	분석	{본품#다양성, 긍정, 노즐들}	
패키지·구성품	일반	문장	새하얀 그램과 잘 어울리는 흰색 마우스.
		분석	{패키지·구성품#일반, 긍정, 흰색 마우스}
	디자인	문장	그래도 이렇게 예쁜 파우치.. 주신다면 땡큐져..
		분석	{패키지·구성품#디자인, 긍정, 파우치}
	품질	문장	특히, 단선도 잘 안 당하는 극강 케이블인 점도 강점.
		분석	{패키지·구성품#품질, 긍정, 케이블}
편의성	문장	가족 여럿이서 쓸땐, 거치대까지 붙여 쓰면 칫솔보관하기 편해요 ㅎㅎ	
	분석	{패키지·구성품#편의성, 긍정, 거치대}	
브랜드	일반	문장	겨울철 더 빛을 발하는 유닉스 제품!!!
		분석	{브랜드#일반, 긍정, 유닉스 제품}
	가격	문장	이어폰이든 충전기든 애플사에서 나온 건 하나같이 비싸.
		분석	{브랜드#가격, 부정, 애플사}
	품질	문장	하이 테크놀로지 기술력의 닥터웰이 자신 있게 추천하는 제품!
		분석	{브랜드#품질, 긍정, 닥터웰}
	인지도	문장	LG는 디스플레이 또한 전세계적으로 유명하죠.
		분석	{브랜드#인지도, 긍정, LG}

③ 제품 기타

<표 11> 제품 기타 세부 도메인 분석 예시

개체	속성	예시	
제품 전체	일반	문장	글라스락 베이비 이유식기 이거 정말 물건이더라구요
		분석	{제품 전체#일반, 긍정, 글라스락 베이비 이유식기}
	가격	문장	에스러빈 마켓 롱패딩이야말로 가성비짱!!
		분석	{제품 전체#가격, 긍정, 에스러빈 마켓 롱패딩}
	디자인	문장	#몽블리안마봉 귀여워서 챙길맛이 남..
		분석	{제품 전체#디자인, 긍정, 몽블리안마봉}
	품질	문장	자율신경을 시원하게 마사지해주는 목어깨안마기 세라핑거봉~
		분석	{제품 전체#품질, 긍정, 목어깨안마기 세라핑거봉}
	편의성	문장	어깨를 덥지 않아서 더욱 편안하고 활동도 자유로워서 너무 편안해요 🐾
		분석	{제품 전체#편의성, 긍정, Null}
인지도	문장	요즘 핫한 #요미차트 아시나요~?!^^	
	분석	{제품 전체#인지도, 긍정, 요미차트}	
본품	일반	문장	손잡이 부분에 거울이 있어서 중간에 한번씩 메이크업 점검하기도 좋은..§
		분석	{본품#일반, 긍정, 손잡이 부분}
	디자인	문장	돋보기 모양의 귀여운 바늘이 시침 빨간화살표는 분침
		분석	{본품#디자인, 긍정, 시침}
	품질	문장	견고하고 튼튼해서 넘 맘에 드는 홈 아이템
		분석	{본품#품질, 긍정, 홈 아이템}
	편의성	문장	지퍼여서 또 세상 편하네 넘이뿌당♥👍👍
		분석	{본품#편의성, 긍정, 지퍼}
다양성	문장	#4가지상쇄함 이라 선택해서 사용하기 좋아	
	분석	{본품#다양성, 긍정, Null}	
패키지·구성품	일반	문장	마마포레스트 처음 사용해보는데~요거 #스타터키트 좋네요!!
		분석	{패키지·구성품#일반, 긍정, 스타터키트}
	디자인	문장	사진에는 안찍혔지만 틈케이스도 이쁜 게임^^
		분석	{패키지·구성품#디자인, 긍정, 틈케이스}
	품질	문장	무엇보다 한장씩 뽑히는 NON POP-UP 기술이 최고네요
		분석	{패키지·구성품#품질, 긍정, 한장씩 뽑히는 NON POP-UP 기술}
	편의성	문장	그물망도 같이있어 목욕 끝난후 정리하기도 쉽습니다~
		분석	{패키지·구성품#편의성, 긍정, 그물망}
다양성	문장	맥포머스 매직월드 특별구성품 역대 최다 구성이라 놓치고 싶지 않아~	
	분석	{패키지·구성품#다양성, 긍정, 맥포머스 매직월드 특별구성품}	
브랜드	일반	문장	넘 만족스러운 #찌웃제작소
		분석	{브랜드#일반, 긍정, 찌웃제작소}
	가격	문장	늘 찾게되는 가격좋고~ 품질좋아~ 늘 만족도 100인 아동복쇼핑몰 제이키즈
		분석	{브랜드#가격, 긍정, 아동복쇼핑몰 제이키즈}
	품질	문장	향기장인 #데메테르
		분석	{브랜드#품질, 긍정, 데메테르}
	인지도	문장	메모리폼매트리스 중 제일 핫한 #사일런나잇
		분석	{브랜드#인지도, 긍정, 사일런나잇}

※ 세부 분석 기준

- 제품 도메인의 상정하지 않은 개체 예시(Out Of Scope 처리)
 - ☞ 플란테시스템 이름이 쓸데없이 어렵구만(제품명)
 - {Out Of Scope, 부정}
 - ‘제품’ 주제와 관련되지 않은 감성 문장이라면 제외
 - ☞ 요즘 기분이 안 좋아서 기분 전환하러 백화점 쇼핑 갔어요.
 - 제외 처리
 - ☞ 그렇다면 역시 45요금제는 하나 쓰잘데기가 없다.(요금제 및 통신사)
 - {Out Of Scope, 부정}
 - ☞ 너무 빠른 배송에 어안이 병병.(배송)
 - {Out Of Scope, 긍정}

- 화장품·세정제, 세척제 등의 ‘향, 색상, 제형, 용량’에 대한 언급은 ‘본품#일반’, ‘발색, 발림성, 성분’ 및 내용물의 사용감에 대한 언급은 ‘본품#품질’로 분류함
 - ☞ 벨벳틴트추천 드릴 수밖에 없는 플럼?복숭아향이 난답니다.
 - 문장 분석: {본품#일반, 긍정, 플럼?복숭아향}
 - ☞ 온천수 성분이 들어서 피부에 순하게 작용 할 것 같아요.
 - 문장 분석: {본품#품질, 긍정, 온천수 성분}

- 패키지 및 용기의 위생과 관련한 언급은 ‘패키지·구성품#일반’으로 분석함.
 - ☞ 게다가 튜브타입으로 여름철에 위생적으로 사용할 수 있습니다 :)
 - 문장 분석: {패키지·구성품#일반, 긍정, 튜브타입}
 - ☞ 위생적인 스포이드형식도 맘에 드네요~
 - 문장 분석: {패키지·구성품#일반, 긍정, 스포이드형식}

- 기능 언급 없는 특정 용도에 대한 적절성 언급은 ‘제품 전체#일반’으로 분류
 - ☞ 요건 #출산선물 로도 넘좋을듯해
 - 문장 분석: {제품 전체#일반, 긍정, Null}
 - ☞ 데일리템으로 딱!□좋아§
 - 문장 분석: {제품 전체#일반, 긍정, Null}

- 색조 화장품의 색상 관련 단순 평가는 ‘본품#일반’으로 분류
 - ※ 단, ‘발색’에 대한 언급은 본품#품질로 분류함
 - ☞ §슈에무라 #아이치크팔레트 : 색상들이 지금 계절에 딱~~
 - 문장 분석: {본품#일반, 긍정, 색상들}
 - ☞ 가볍지만 선명한 발색력이 민낫에 발라도 얼굴이 화사해지는 듯
 - 문장 분석: {본품#품질, 긍정, 발색력}

- 특정 장치 및 부속품의 특징에서 기인한 편의성은 ‘본품#편의성’, 패키지 및 디자인 관련 휴대성 및 편의성 언급은 ‘패키지·구성품#편의성’으로 분석함.
 - ☞ 탐투토워시 하나로 샴푸까지 되니깐 샤워시키기도 간편해~§
 - 문장 분석: {본품#편의성, 긍정, 탐투토워시}
 - ☞ 따로 파우치형은 휴대성도 좋아서 가방에 꼭 넣고다닌다!

문장 분석: {패키지·구성품#편의성, 긍정, 파우치형}

- '본품' 개체의 OTE는 물리적 장치와 기능명 모두 추출 가능함

예) 아까 한 번 잠깐 되더니 지금 또 등록 버튼이 먹통이다.

→ '등록 버튼' OTE 추출 가능

예) 전에 작동 안되었던 자막 검색 후 등록 기능이 똑같이 작동 안 된다!!!

→ '자막 검색 후 등록 기능' OTE 추출 가능

- '소음' 관련 언급은 '품질' 속성으로 고정함

예) 작동 시 소리는 좀 나는 편인데 거슬리는 정도의 소음은 아니고 움직이는 모습이 꼭 귀여운 동물의 발 같기도 하고 그러네요.

문장 분석: {제품 전체#품질, 중립, 소음}

예) 전에 작동 안되었던 자막 검색 후 등록 기능이 똑같이 작동 안 된다!!!

→ '자막 검색 후 등록 기능' OTE 추출 가능

3) 여행 도메인

여행 도메인 문서는 작성자가 여행지를 주제로 하여 작성한 여행 리뷰글이며 리뷰 대상에 따라 숙소, 식당, 관광지 세부 도메인으로 분류됨. 총 139문서, 3869문장을 분석의 대상으로 함.

(1) 개체

<표 12> 여행 도메인의 개체

구분	내용
① 장소 전체	관광, 숙소, 식당 등 장소 전반
② 시설	장소 내 시설 전반
③ 상품	기념품, 음식 등 판매 상품과 투어 서비스 등 관광 상품
④ 서비스	직원의 친절 등 서비스 관련 요소

(2) 속성

<표 13> 여행 도메인의 속성

구분	내용
① 일반	개체 자체
② 가격	가격 및 마케팅 요소 평가
③ 편의성	이용 편의성 평가
④ 심미성	인테리어, 외관 등 미적 요소 평가
⑤ 접근성	위치 및 접근성
⑥ 청결도	청결도 및 위생 상태 언급
⑦ 인지도	유명도, 인기 등 언급

(3) 개체#속성 쌍

<표 14> 여행 도메인의 개체#속성 쌍

	일반	가격	편의성	심미성	접근성	청결도	인지도
장소 전체	○			○	○	○	○
시설	○	○	○	○		○	
상품	○	○					○
서비스	○						

(4) 세부 도메인별 예시

① 숙소

<표 15> 숙소 세부 도메인 분석 예시

개체	속성	예시	
장소 전체	일반	문장	가족들과 함께가기 너무 좋은 #비클래시 #폴빌라펜션 추천해드릴게요 !
		분석	{장소 전체#일반, 긍정, 비클래시 #폴빌라펜션}
	심미성	문장	또한 신축 리조트로 내 외관 모두 깔끔 §
		분석	{장소 전체#심미성, 긍정, 신축 리조트}
	접근성	문장	영흥도와 대부도 가운데 위치해있어 수도권 근접펜션 찾는분들에게 딱~
		분석	{장소 전체#접근성, 긍정, Null}
	청결도	문장	숙소 자체가 무지 깨끗하다.
		분석	{장소 전체#청결도, 긍정, 숙소 자체}
인지도	문장	워낙 평이 좋아서 기대하고 갔는데도 더 좋음 §§	
	분석	{장소 전체#인지도, 긍정, Null}	
시설	일반	문장	넓은 바베큐장도 넘 맘에 듭니다~
		분석	{시설#일반, 긍정, 바베큐장}
	가격	문장	대가족이나 워크샵도 문제없으니 진짜 이만큼 가성비좋은곳도 없죠~
		분석	{시설#가격, 긍정, Null}
	편의성	문장	침대는 깨끗하고 편안했다.
		분석	{시설#편의성, 긍정, 침대}
	심미성	문장	맑은 공기와 탁 트인 바다와 섬이 보이는 아름다운 오션뷰 객실§
		분석	{시설#심미성, 긍정, 오션뷰 객실}
청결도	문장	시설이 너무 깨끗해서 깜놀했어요§	
	분석	{시설#청결도, 긍정, 시설}	
상품	일반	문장	조식은... 별로다
		분석	{상품#일반, 긍정, 조식}
서비스	일반	문장	끝까지 서비스 짱이고 다들 친절하셨다.
		분석	{서비스#일반, 긍정, 서비스}

② 식당

<표 16> 식당 세부 도메인 분석 예시

개체	속성	예시
----	----	----

장소 전체	일반	문장	저번 여수여행때 들렀던 #고기맛집
		분석	{장소 전체#일반, 긍정, 고기맛집}
	심미성	문장	가게안이 너무예뻐서~ 인생샷찍고 힐링하기좋아요~
		분석	{장소 전체#심미성, 긍정, 가게안}
	청결도	문장	깔끔 친절해서 현금냈더니 상호명 까먹음 ㅠ
		분석	{장소 전체#청결도, 긍정, Null}
인지도	문장	#블라썸1407 . 제주 #맛집 #핫플 로 인정합니다~♥	
	분석	{장소 전체#인지도, 긍정, 블라썸1407}	
시설	일반	문장	화장실 그게 최선이었니....
		분석	{시설#일반, 긍정, 바베큐장}
	편의성	문장	주차공간도 넓으니까 자주 찾게 되네요:)
		분석	{시설#편의성, 긍정, 주차공간}
상품	일반	문장	육즙이 살아있는 맛있는 #돼지갈비 정말 배터지게 먹고왔어요 ㅋㅋ
		분석	{상품#일반, 긍정, 돼지갈비}
	가격	문장	한 끼 식사로는 매우 부담스런 양과 가격
		분석	{상품#가격, 긍정, 가격}
	인지도	문장	가게 이름처럼 레몬파이가 유명한 레몬파이
		분석	{상품#인지도, 긍정, 레몬파이}
서비스	일반	문장	사장님 너무친절
		분석	{서비스#일반, 긍정, 사장님}

③ 관광지

<표 17> 관광지 세부 도메인 분석 예시

개체	속성	예시	
장소 전체	일반	문장	룰루랄라, 내가 좋아하는 #광안리해수욕장.§
		분석	{장소 전체#일반, 긍정, 광안리해수욕장}
	심미성	문장	이렇게 인테리어도 힙했다
		분석	{장소 전체#심미성, 긍정, 인테리어}
	가격	문장	아까 거기보단 비싸지만 그리니치 빌리지에 비교하면 괜찮은 가격...
		분석	{장소 전체#가격, 중립, 가격}
	심미성	문장	이렇게 인테리어도 힙했다
		분석	{장소 전체#심미성, 긍정, 인테리어}
	인지도	문장	파란하늘 파란바다 여긴 #핫플레이스 진짜 넘 좋아
		분석	{장소 전체#접근성, 긍정, 핫플레이스}
상품	일반	문장	하지만 다시 말하지만 시간이 없다면 무조건 투어를 추천함
		분석	{상품#일반, 긍정, 투어}
	가격	문장	왕복 차비만 생각해도 괜찮은 가격이라 생각함
		분석	{상품#가격, 긍정, 가격}
	인지도	문장	지난해 노란꽃 축제 때는 무려 100만 명이 왔다고...
		분석	{상품#인지도, 긍정, 노란꽃 축제}
서비스	일반	문장	사진 찍어보라며 설정 잡아주시는 ~~ >_<
		분석	{상품#인지도, 긍정, Null}

※ 세부 분석 기준

- 여행 도메인의 상정하지 않은 개체 예시(Out Of Scope 처리)
 - ☐ 예 도쿄에 있을동안 날씨가 좋아서 제일쫌아..§(날씨)
→ {Out Of Scope, 긍정}

- 여행 도메인의 경우 여행의 과정을 시간 순서에 따라 나열하고 있는 글이 다수 존재함.
만약 한 문서에 여러 개의 도메인이 겹치는 경우 가장 큰 비중을 차지하는 주 리뷰 대상을 골라 도메인을 선정함.

- 숙소에서 제공하는 조식 및 어매니티에 대한 언급은 '상품' 개체로 분류함
 - ☐ 예 벨벳티트추천 드릴 수밖에 없는 플럼?복숭아향이 난답니다.
문장 분석: {본품#일반, 긍정, 플럼?복숭아향}
 - ☐ 예 온천수 성분이 들어서 피부에 순하게 작용 할 것 같아요.
문장 분석: {본품#품질, 긍정, 온천수 성분}

- 투어 서비스, 행사, 공연, 축제 등은 여행의 체험 상품에 속하므로 '상품' 개체로 분류함
 - ☐ 예 하지만 다시 말하지만 시간이 없다면 무조건 투어를 추천함
문장 분석: {상품#일반, 긍정, 투어}
 - ☐ 예 그치만 그리니치빌리지에서 블루노트와 스몰스 재즈바에 가본 나로선 할렘의 재즈바들의 공연 역시 매우매우 훌륭했다.
문장 분석: {상품#일반, 긍정, 할렘의 재즈바들의 공연}

<부록2> 국내외 감성·감정 분석 현황 자료

1. 국외 감정 분석 말뭉치 리스트

번호	데이터 이름	데이터 종류	데이터 크기	레이블 개수	레이블 유형
1	Cornell movie-review datasets	영화 리뷰	2000개 리뷰	2	pos, neg
2	HCR (Health Card Reform)	SNS(트위터)	2,516개 트윗	5	pos, neg, neu, irrelevant, unsure
3	IMDB (Internet Movie Database)	영화 리뷰	50,000개 리뷰	2	pos, neg
4	MPQA (Multi-Perspective Question Answering)	뉴스 기사	535개 기사	8	pos, neg, both, neu, uncertain-pos, uncertain-neg, uncertain-both, uncertain-neu
5	Obama-MacCain Debate (OMD)	SNS(트위터)	3,238개 트윗	4	pos, neg, mix, other
6	Sanders Twitter	SNS(트위터)	5,512개 트윗	4	pos, neg, neu, irrelevant
7	SemEval-2016 Task 4: Sentiment Analysis in Twitter	SNS(트위터)	30,000개 트윗	3	pos, neg, neu
8	SemEval-2016 Task 5: Aspect-Based Sentiment Analysis	상품/서비스 리뷰	6,000개 문장	3	pos, neg, neu +conflict
9	SS-Tweet (Sentiment Strength)	SNS(트위터)	4,242개 트윗	3	pos, neg, neu
10	SST-2 (Stanford Sentiment Treebank)	영화 리뷰	11,855개 리뷰	2	pos, neg
11	SST-5 (Stanford Sentiment Treebank)	영화 리뷰	11,855개 리뷰	5	neg, somewhat neg, neutral, somewhat pos, pos
12	STS (Stanford Twitter Sentiment corpus)	SNS(트위터)	1.6M개 트윗	3	pos, neg, neu
13	The Dialogue Earth Twitter Corpus	SNS(트위터)	26,110개 트윗	5	pos, neg, neu, not related, can't tell

2. 국외 감정 분석 말뭉치 리스트

번호	데이터 이름	데이터 종류	데이터 크기	레이블 개수	레이블 유형
1	AffectiveText	뉴스기사 제목	1,250개 문장	6	<i>Ekman</i>
2	Blogs	블로그	5,205개 문장	7	<i>Ekman</i> + N/E
3	CrowdFlower	SNS(트위터)	39,740개 트윗	14	<i>Ekman</i> + enthusiasm, fun, hate, neutral, love, boredom, relief, empty
4	DailyDialogs	일상대화	13,118개 문장	7	<i>Ekman</i> + N/E
5	Electoral-Tweets	SNS(트위터)	100,000개 트윗	8	<i>Plutchik</i>
6	EmoBank	일반 텍스트	10,548개 문장	(3)	V-A-D
7	EmoInt	SNS(트위터)	7,097개 트윗	4	fear, joy, sadness, anger (<i>Ekman</i> - disgust, surprise)
8	Emotion-Stimulus	일반 텍스트	1,549개 문장	7	<i>Ekman</i> + shame
9	fb-valence-arousal	SNS(페이스북)	2,895개 포스트	(2)	V-A
10	Grounded-Emotions	SNS(트위터)	2,557개 트윗	2	happy, sad
11	ISEAR	질의응답	7,665개 문장	7	joy, fear, anger, sadness, disgust, shame, guilt
12	SemEval 2018	SNS(트위터)	12,600개 트윗	4	fear, joy, sadness, anger (<i>Ekman</i> - disgust, surprise)
13	SemEval 2019	대화(인간-기계)	38,424개 대화	4	happy, sad, angry, others
14	SSEC	SNS(트위터)	4,868개 트윗	8	<i>Plutchik</i>
15	Tales	동화	15,302개 문장	8	<i>Ekman</i> + angry, disgust
16	TEC	SNS(트위터)	21,051개 트윗	6	<i>Ekman</i>

3. 국내 감성 분석 말뭉치 리스트

번호	데이터 이름	데이터 종류	데이터 크기	레이블 개수	레이블 유형
1	국립국어원 감성 분석 말뭉치 2020	SNS, 블로그	2,081개 문건	5	강한 긍정(2), 긍정(1), 중립(0), 부정(-1), 강한 부정(-2)
2	네이버 영화 리뷰(NSMC)	댓글	200,000개 문장	2	긍정(1), 부정(0)
3	Chatbot_data	대화(사람-챗봇)	11,876 문장 쌍	3	긍정, 부정 중립
4	Toxic comment data	댓글	20,000개 문장	6	toxic, obscene, threat, insult, identity_hate
5	혐오 발언 분류 데이터셋	댓글	9,281개 문장	다중	contain_gender_bias(T/F), bias(gender, other, none), hate(hate, offensive, none)

4. 국내 감성 분석 말뭉치 리스트

번호	데이터 이름	데이터 종류	데이터 크기	레이블 개수	레이블 유형
1	한국어 감성 정보가 포함된 단발성 데이터셋	댓글	38,594개 문장	7	기쁨, 슬픔, 놀람, 분노, 공포, 혐오, 중립
2	한국어 감성 정보가 포함된 연속적 대화 데이터셋	대화(사람-사람)	10,000개 대화(55,627개 문장)	7	행복, 슬픔, 놀람, 분노, 공포, 혐오, 중립
3	감성분류를 위한 대화 음성 데이터셋	대화(음성 발화 및 전사 자료)	4차년도(14,606 문장), 5차년도(29,385 문장)	7	neutral, fear, angry, disgust, sadness, surprise, happiness
4	감성 대화 말뭉치	대화(사람-챗봇)	27만 문장	60	데이터 참고
5	웰니스 대화 스크립트 데이터셋	대화(사람-챗봇)	5,232개 문장	61	데이터 참고

<부록3> 감정 분석 말뭉치 구축 시뮬레이션

1. 한국어 감정 분석 말뭉치 구축 시뮬레이션

1.1. 한국어 단발성 대화 데이터셋

번호	원본id	샘플문장	레이블(E)		레이블(P)	
			주석자 1	주석자 2	주석자 1	주석자 2
1	26824	서면보고 받았다는것이 새빨간 거짓말인거지	disgust		disgust	
2	433	애기들이 좋아하는 맛은 아니였어요ㅋㅋ	disgust		disgust	
3	1999	이러다 감옥에서 못 나오면 어찌	sadness		sadness	
4	3360	자기 남편 술중독에 빠지게해서 집앞 대문앞에서 죽여버렸다.	x		x	
5	38	이제 곧 내년이면 22살인 여대생입니다...요즘 고민이 있는데요...	sadness		sadness	
6	36044	순실이도드나들던 청와대다들어가서 머리채라도잡고 끌고나와야할듯	anger		anger	
7	32263	더러운 박그네의 추악한 진실이 나오는구나	disgust		disgust	
8	18290	계속하는게답일지 아닐지참고민입니다...	sadness		sadness	
9	7258	저는 루리웹의 노예가 된건가요??0	disgust		disgust	
10	5875	살인마 맞네!!!!!!	anger		anger	
11	10436	헤어샵.옷 같은건 무료로 대여해주나요??	x		x	
12	34237	손흥민 밀 다 지잡	disgust		disgust	
13	23583	개미만 죽어나는거였어 ㅋㅋ	x		x	
14	24540	나를 부끄럽게 하네요..	sadness		sadness	
15	4263	저말 진심이였겠죠..?	x		x	
16	7025	김광현 실력에 메이저라니????	surprise		surprise	
17	25504	자나깨나 뒤통수 조심!	x		x	
18	30024	수상소감 너무 좋았어요	happiness		joy	
19	9750	일 안해도 안짤려... 대단하다이..	disgust		disgust	
20	5928	오우!! 유희관보다 빠른데-!!	surprise		surprise	
21	34483	애를 왜 더 힘들게 만들어...시상식 참석가 그렇게 중요한가? 회복하는게 우선이지...	anger		anger	
22	15158	오촌살인사건보고 진짜 분노가치밀어오른다	anger		anger	
23	1462	사람이 만든 바이러스...라면 무섭다 공격 공격	fear		fear	
24	16576	아베.목.땅오면 내가 동상 10000000개.만들어준다..	anger		anger	
25	7081	너무도 당연한 것을 무슨 연구까지..	disgust		disgust	
26	9259	우리나라에서 저랬다간 대통령 물러가라고 쫓발시위했을걸....	x		x	
27	10118	살다살다 이름도 모르고 얼굴 한번 본 적 없는 사람이 부럽다고 생각하게 줄은 몰랐는데.ㅎ0	disgust		disgust	
28	1149	어째야 하나여;; “누구세요?“했다가 진짜 아는 사람이면 무안할거같은데;;0	fear		fear	
29	28518	내가살면서 영화 같은거 극장에서 2번본거 인터스텔라밖에없음...진짜명작임	happiness		joy	
30	15789	어제 온다는 비는 왜 오지 않았고?	anger		anger	
31	14040	고졸 취소가 문제가 아니라...부정입학과 학사관리자들...발본색원하여...엄벌해야	anger		anger	
32	16285	이마트에서 대란이 6900원이던데.. 기사는 판세상인가	disgust		disgust	
33	12505	이길때 혼흔하더니 털리니까 무슨 댓글에 다 악의적인것만 1,2,3등이나	disgust		disgust	
34	14855	4대강 22조는 어디다 쓴거냐?	anger		anger	
35	14548	차라리 같은 소리하고있네!	anger		anger	
36	28152	몸이랑 얼굴 비율이 헐리웃 배우같아ㄷㄷ	surprise		surprise	
37	1167	사랑인지 모르겠다	x		x	
38	21630	다들 경험담이나 어떻게 해야되는지 조언부탁해요ㅠ	x		x	
39	13054	지금퇴근 십알ㅠㅠㅠ	sadness		sadness	
40	35960	육아휴직가는것 좋은데 대체인력충원이 법제화되어있지않아 정작 같은부서 일하는사람들만 죽어난다	disgust		disgust	
41	25941	게임 개발업체 직원들인가? 설마...	x		x	

번호	원본id	샘플문장	레이블(E)		레이블(P)	
			주석자 1	주석자 2	주석자 1	주석자 2
42	14237	예전에 전도연 기사 댓글들 나나 연기 개뿔하는데 왜나오냐고 그러던게 기억난다 그때 욕들 살벌하던 이제와선..ㄹㄹ	disgust		disgust	
43	25910	언프리티 랩스타4 출연예정 : 제이니, 전지윤	x		x	
44	23578	사실상 일제때 조선인 일본인보다 더 일본인	disgust		disgust	
45	127	갤노트 7 = 폭탄 ???	disgust		disgust	
46	31074	유느님 고생 많으십니다 ㅌㅌ	sadness		sadness	
47	12089	지구상 최악의 종자 째깨들..	anger		anger	
48	17862	우리나라 제품은 가격이 너무 초초초프리미엄이야..	disgust		disgust	
49	24917	두산에서 처럼 잘 하실것 같아요,,	x		anticipation	
50	32814	야구는 투수놀음이란걸 투수들 폭망이면 답이 없지	x		x	
51	18588	지금 죽어도 상관없어..	x		x	
52	18071	너무 힘들게 찍었고 편집도 잘못하는데 하느라 완전 고생 ㅠㅌ	sadness		sadness	
53	1890	처음이라서 지금 완전 불안해 죽겠어요..	fear		fear	
54	4230	데이트 비용 문제 때문에 그러는데..	x		x	
55	13570	봤는데 돈이 너무 아깝워서 속쓰리다ㅌㅌ	sadness		sadness	
56	17221	술은 어찌 마시면 마실 수록 못 끊을까요...	sadness		sadness	
57	5085	분명 속썩일것같아	x		anticipation	
58	27563	지금은 22살 아기는 이제 돌 됐어요	x		x	
59	36934	제발 재검표해서 결과 좀 바꿔봐!	anger		anger	
60	32268	에반게리온이 충전기 뽑으면 5분밖에 작동 못한다는게 이제보면 대단한 기술임	x		x	
61	9438	간지도몰랐었네	x		x	
62	29557	바마형님 고생하셨습니다 ㅠ.ㅠ	sadness		sadness	
63	27634	다시한번 추가추카~~	happiness		joy	
64	13662	이제 마음껏 관리받고 미용실도 가시길아 전에도 뭐 세상 편하게 하셨구나...	disgust		disgust	
65	38105	오타쿠는 가진게 많고 뭐든지 원하는건 얻고본다ㅇ스?	disgust		disgust	
66	9174	딸맞네	x		x	
67	14886	남의 눈에서 피눈물 났으니 그사랑이 얼마나 깊다해도 이미 핏빛으로 얼룩이 고 있지 않을까~~~	disgust		disgust	
68	11555	들통날 거짓말을 뭘 그리 짜증내면서 하셨을까..	disgust		disgust	
69	15202	불법조업안하면다해결되는데 뭐가불만이고	disgust		disgust	
70	14213	저만 헛소리하는 계집들은 배를 갈라야된다	disgust		disgust	
71	23396	솔직히 예원처럼 이목구비 뚜렷한 연예인 요즘 별로 없는듯	x		x	
72	23560	권혁 더 추해지기 전에 삼성으로 복귀해야	disgust		disgust	
73	28768	어른스런 모습 언제나 자랑스러워요	happiness		joy	
74	16485	환장하겠네!!!!	anger		anger	
75	31813	학원에서 어쩌다 보니 어느 여자 번호를 얻었어.	x		x	
76	33960	구단에선 김성근에 대한 더 이상의 투자는 기대말라!!	anger		anger	
77	11102	문제인이 저런 일도 했었어?	x		x	
78	12682	똥물에 튀겨줘.일.년.	anger		anger	
79	34940	대학생이 생각머리하곤	disgust		disgust	
80	1534	앞에 있는 경찰 도끼 들고 있는건가?	x		x	
81	863	아침밥 해야할까요...?	x		x	
82	8817	일상에서도 카톡에서도할 말이 없어요.	x		x	
83	33390	개돼지 소리 들어도 싸네.	disgust		disgust	
84	30232	항상 고마워요	x		x	
85	43	8월이면 둘째가 태어나서 마음이 더 조급한것같아요 ㅠㅌ	sadness		sadness	
86	12853	엄청난 결함인데 당연히 환불과 새제품교환해 주는게 상식 아닌가?	disgust		disgust	
87	29674	백선생님은 우리가 하고 싶어지게 만드는 능력을 가지고 계신듯 합니다.	x		x	
88	28273	특검팀 힘내세요 !!!!	x		anticipation	
89	27488	제가 하고싶었던걸 해주시니 가슴이 뻥 열리는 기분이네요	happiness		joy	

번호	원본id	샘플문장	레이블(E)		레이블(P)	
			주석자 1	주석자 2	주석자 1	주석자 2
90	4498	뭘 또 덮을라고 이런 기사 올리나...	disgust		disgust	
91	2819	초산이라 현재 불안 걱정 긴장으로 하루하루를.	sadness		sadness	
92	25567	역시 박주성 관련 댓글이 1위구만	x		x	
93	33218	지금 대부분 귀경중일텐데 버리지같은 알바들은 가족도 없나보네. 사생아같은 색히들	disgust		disgust	
94	29625	우지오빠 어떻게 노래 그렇게 잘만들고 노래도 잘해요??ㅠㅠ	x		x	
95	35754	있는 사실을 말하라는건데 강요? ㅋㅋㅋㅋ	disgust		disgust	
96	8816	판도라의 상자를 열게 됐는데 제목 그대로..	x		x	
97	23123	재밌는데 어제는 안봐어 권들러리스트봣거든	x		x	
98	18075	이 문장이 머릿속에서 자꾸 맴돌아서 고민..	sadness		sadness	
99	3452	조만간 밀봉한 시체 나온다	x		x	
100	3406	이기혁 기자님 자살당하지 않게 조심하세요 ㄸㄸ	sadness		sadness	
101	33842	오로지 남탓 십자십자		anger		anger
102	3888	추우신데 같이 온기를 나눠야 되지않을까 사료됩니다.		fear		fear
103	27465	손흥민...토티님의 승리...!		happiness		surprise
104	3602	위성진짜던질라나요....		surprise		anticipation
105	25503	낭만닥터 여운이 아직 가시지않았어요		happiness		joy
106	41	집에와서 또 운동하긴했는데지난 3주의 노력이 물거품이 될까봐정말 걱정되요		fear		fear
107	20568	알긴 아는데 못하고있다가 최근 2년동안 그자식이랑 야식을 하도 쳐묵쳐묵해서남는건 살바게 없더라구요...		sadness		sadness
108	6928	특검에서 조대위 동기들만 소환조사하면 ㄹ혜7시간 답나올듯.....		anger		anger
109	33686	태업이라도 하지 않는 한 크보에서 못할 클래스가 아님		happiness		anticipation
110	24187	갑자기 댓글 알바부대가 늘어난것같은 느낌이다		disgust		disgust
111	409	프로필사진에는 아랍인처럼 안생겼던데..		surprise		surprise
112	30867	와~~오늘 정동하 대단했다		surprise		surprise
113	17189	창피스럽다.		disgust		disgust
114	5002	술도 안먹거아무튼 간에 조심해야 된다		fear		fear
115	11770	까고있네 이란한테 개발될것들이		anger		anger
116	21974	걷는 자세가 잘못된건지..		fear		fear
117	18544	약속은 지켜야지 ...힘드네요 ..		sadness		sadness
118	29379	이민기 멋있네.. 진정한 상남자..		happiness		joy
119	33031	축하해요바다누님 결혼까지~		happiness		joy
120	26473	la아리랑 웬만 순풍 거침없이 하이킥 지붕뚫고 하이킥 짧은다리의 역습 모두 김병욱 작품		x		x
121	21343	나쁜놈들 안 그래도 작은 고추가 더 작아졌어.		sadness		sadness
122	34686	윗대가리들이 하도 똥을 싸질러 놔서 아랫 사람들만 힘들구나...		anger		anger
123	37220	그러고보면 예수는 진짜 대단한 사람이지... 말빨로 전세계에 개똥을 전파시켰으니...ㄸㄸ 성공한 북쪽김씨가문 버전이라고 봐야겠네		disgust		disgust
124	24923	뭐하니 제명시켜야지		anger		anger
125	7517	아니지 지금 이렇게 아니잖아!!		anger		anger
126	20253	나 혼자감 ㅠㅠ		sadness		sadness
127	4133	임영민 떨어질 확률크나요?		fear		fear
128	35000	별게 다 기사거리...		surprise		surprise
129	34068	너 만날 생각 없다.		disgust		disgust
130	678	널 면접보기로 했는데...		fear		anticipation
131	32354	대학졸업전 회사 3~4군대에서 스카웃해갑니다.		happiness		joy
132	29641	어제 경기에서 지현오빠 첫타석 홈런...		happiness		joy
133	37940	지금도 각종 방송에서 한수지노래로 되게 많이 사용되는데 ... 요즘시대가 어느땐데 허튼 수작질이냐!!		anger		anger
134	34956	스프링캠프때부터 혹사 때문에 로저스도 갔다는 것을 모르나...		anger		anger

번호	원본id	샘플문장	레이블(E)		레이블(P)	
			주석자 1	주석자 2	주석자 1	주석자 2
135	33838	그 멍청하년이 나와서 유체이탈만 하겠지		disgust		disgust
136	24531	손석희사장이 현재소장 후임으로 내정 됐나 보네요.		disgust		disgust
137	4011	대학로에 펠리피노들 조심해라...		fear		fear
138	29511	빅뱅은 월드스타라서 인기가요같은 무대가 이젠 어색하다		surprise		surprise
139	21376	저기에 북한 돼지새끼가 앉아서 피흘리고 있어야 하는데..에휴...어린 것이 너무 마음 아프다...ㅠ		sadness		sadness
140	1294	라는 소리들음내가 잘못된거임?		sadness		sadness
141	29385	이거에 그리 기여한바는 없지만 너무너무 축하해요 ㅠ ㅠ ㅠ		happiness		joy
142	31711	영화로 쟁있게 봤던 마다가스카뮤지컬로 다시 태어났어요!		happiness		joy
143	2160	그리고 아따맘마 14권이 일본에서도 아직 만나와서 또 고민0		sadness		anticipation
144	20805	어떡해요 ㅠ ㅠ ㅠ?		fear		fear
145	2429	밀가루 음식 먹어도 괜찮을까요?		fear		fear
146	18252	팔이떨리고 저리고 ㅠ ㅠ		sadness		sadness
147	35230	광우병 촛불들었던 미친연놈들 많을거다.		disgust		disgust
148	24295	니퍼트도 안되고 오승환도 안되지		sadness		sadness
149	24493	이런 인간이 설쳐서 다음에 투표 꼭해야겠다		anger		anger
150	12969	가지수만 많이 넣어서 될 일도 안되게 만든 무능한 민주당.. 오아.. 열 받네		anger		anger
151	32093	어제 김현철이 살렸달ㅋㅋㅋㅋ		happiness		joy
152	5349	설마 1981년 광주 처럼 내일 광화문에 되지는 않겠지		fear		fear
153	24792	두명이나 메자 갈 수 있다고 한 팀 맞나		anger		anger
154	3524	웬지 영국이나 eu랑 입장이 다르나 조만간 중국도 위험하겠는데....		fear		fear
155	120	미국 시민들 불안해서 어떻게 살아...ㄷㄷ		fear		fear
156	26899	삼성증권 말발굽펀드.		anger		anger
157	26128	김영란법으로 폰팔이도 몰같이좀 해야지 않겠나?		disgust		disgust
158	11327	매국단체가 3만명이나됨?		surprise		surprise
159	8681	메이저 5승에 방어를 7점대가 100만달러....		surprise		surprise
160	5560	발정난논이 미쳤나??		anger		anger
161	5633	그래서 헤어진 사람한테도 연락해 ?		surprise		surprise
162	1726	저 귀신갈튼거보면 오늘 밤 혼자못자요 ㅠ ㅠ 0		fear		fear
163	10979	깊신도 짝이 있다는데.설마 ?		surprise		anticipation
164	9320	이렇게 다 되는데 왜 세월호때는 걷는것조차 못하게 한거지 ?		surprise		surprise
165	1314	제 언니의 남편을 제 남편이 뭐라고 부르나요 ?		x		x
166	5265	저이제 학점짜이는건가요..		fear		fear
167	13424	무당일가가 조종하는 나라. 헬조선.		anger		anger
168	37459	정신병자 쿠루에서 똘똘 방승에 나오는 자체가 똘똘이냐 ~		disgust		disgust
169	36801	역시 전라도 죄없는 사람잡는다		disgust		disgust
170	15559	탄핵결사반대!결사반대!결사반대!		anger		anger
171	26195	엘지가아니라 애플이 반사이익볼듯		surprise		anticipation
172	15031	아직도 정신 못차렸구만 쫓겨나봐야 정신차릴거지?		anger		anger
173	27708	그래서 손지환선수도 좋아해요.....		happiness		joy
174	26365	로벤 푹푹한듯 안한듯.ㅎㅎ		happiness		joy
175	17062	오라버니 마음이 아플것 같아서 저의 마음이 더 아프네요...		sadness		sadness
176	3106	카톡차단하지않는다는건무슨 의미일까요??		x		x
177	34016	덱은 언행이랑 정장이랑 안어울려요~		anger		anger
178	33273	이색기 완전 양아치네		anger		anger
179	28462	완전 감동적이네~!		happiness		joy
180	5575	치킨공주 모하니??		surprise		surprise
181	14103	태극기집회는 말도 안꺼내네 편파보도 어느언론에서 그러던데 기사내고 안내고는 자기들이 판단해서 하는거라구		anger		anger

번호	원본id	샘플문장	레이블(E)		레이블(P)	
			주석자 1	주석자 2	주석자 1	주석자 2
182	30621	린드블럼 그동안 고생했어요		happiness		joy
183	13338	조윤선은 일본군 위안부로 김기춘은 탄광으로 보내세요		anger		anger
184	18354	힘들어요휴...		sadness		sadness
185	29689	마음씨가 따뜻하군 헤리b		happiness		joy
186	7952	헬퀴가 포시를 갈줄이야...		surprise		surprise
187	38587	정신병자새끼 밥되게 놔두지!!		disgust		disgust
188	30565	감사 합니다~'		happiness		joy
189	5455	밥먹자 또는 술먹자 할것같기도 하고...		surprise		anticipation
190	24798	1조위엔은 180조원이 아닌가요?		x		x
191	22339	그건 모르겠는데 두산보다 약하다는건 알겠다		anger		anger
192	18361	19일날 가입소 해서 이제 혼련 쳐주 차를 찍고 있는데..		sadness		sadness
193	24374	아무이상없이 제품이상없는거 같으면 계속 써도 무방하지않나..?		x		x
194	31805	그리고 미리 메리크리스마스~		happiness		joy
195	37278	로스쿨도 폐지해라 어떻게 뽑을지 눈에 선하다		anger		anger
196	16877	모르쇠 저런것들은 주리를 틀어야돼		anger		anger
197	16111	수비도 수준급이고 터너 없으면 3루 대안있나?		happiness		trust
198	10460	니 키스때문에 갤럭시 폭발했어		surprise		surprise
199	4240	반팔 바디슈트 입혀도 될까요??		x		x
200	6166	미국 이거 막 하와이 독립한다 알래스카 독립한다 이러는거 아니냐 ㅋㅋㅋㅋ		fear		fear

1.2. 네이버 영화 리뷰

번호	원본id	샘플문장	레이블(E)		레이블(P)	
			주석자 1	주석자 2	주석자 1	주석자 2
1	9976970	아 더빙.. 진짜 짜증나네요 목소리	disgust	anger	disgust	anger
2	3819312	흠...포스터보고 초딩영화줄....오버연기조차 가볍지 않구나	surprise	disgust	surprise	disgust
3	10265843	너무재밌었다그래서보는것을추천한다	happiness	happiness	happiness	joy
4	9045019	교도소 이야기구먼 ..솔직히 재미는 없다..평점 조정	disgust	sadness	disgust	sadness
5	6483659	사이폰페그의 익살스런 연기가 돋보였던 영화!스파이더맨에서 늙어보이기만 했던 커스틴 던스트가 너무나도 이뻐보였다	surprise	happiness	surprise	joy
6	5403919	막 걸음마 떼 3세부터 초등학교 1학년생인 8살용영화.ㅋㅋㅋ...별반개도 아까움.	disgust	disgust	disgust	disgust
7	7797314	원작의 긴장감을 제대로 살려내지 못했다.	disgust	anger	disgust	anger
8	9443947	별 반개도 아깝다 욕 나온다 이웅경 길용우 연기생활이몇년인지..정말 발로해도 그것보단 낮겠다 납치,감금만반복반복..이드라마는 가족도없다 연기 못하는사람만모였네	anger	disgust	anger	disgust
9	7156791	액션이 없는데도 재미 있는 몇안되는 영화	happiness	happiness	joy	joy
10	5912145	왜케 평점이 낮은건데? 꽤 볼만한데.. 헐리우드식 화려함에만 너무 길들여져 있나?	surprise	surprise	surprise	surprise
11	9008700	강인피니트가짱이다.진짜짱이다♥	happiness	happiness	trust	joy
12	10217543	볼때마다 눈물나서 죽겠다90년대의 향수자극!!허진호는 감성절제멜로의 달인이다~	sadness	happiness	sadness	joy
13	5957425	울면서 손 들고 횡단보도 건널때 뛰쳐나올뻔 이범수 연기 드럽게못해	anger	disgust	anger	disgust
14	8628627	답백하고 깔깔해서 좋다. 신문기사로만 보다 보면 자꾸 잊어버린다. 그들도 사람이었다는 것을.	happiness	happiness	trust	joy
15	9864035	취향은 존중한다지만 진짜 내생애 극장에서 본 영화중 가장 노잼 노감동임 스토리도 어거지고 감동도 어거지	disgust	anger	disgust	anger
16	6852435	ㄱ냥 매번 긴장되고 재밌음ㅠㅠ	happiness	happiness	joy	joy

번호	원본id	샘플문장	레이블(E)		레이블(P)	
			주석자 1	주석자 2	주석자 1	주석자 2
17	9143163	참 사람들 웃기게 바스코가 이기면 락스코라고 까고바비가 이기면 아이돌이라고 한다.그냥 까고싶어서 안달난것처럼 보인다	disgust	anger	disgust	anger
18	4891476	굿바이 레닌 표절인것은 이해하는데 왜 뒤로 갈수록 재미없어지냐	disgust	disgust	disgust	disgust
19	7465483	이건 정말 깨알 캐스팅과 질퍽하지않은 산뜻한 내용구성이 잘 버무러진 깨알일드!!♥	happiness	happiness	joy	joy
20	3989148	약탈자를 위한 변명, 이라. 저놈들은 착한놈들 절대 아닌걸요.	disgust	anger	disgust	anger
21	4581211	나름 심오한 뜻도 있는 듯. 그냥 학생이 선생과 놀아나는 영화는 절대 아님	x	surprise	x	joy
22	2718894	보면서 웃지 않는 건 불가능하다	happiness	happiness	joy	joy
23	9705777	재미없다 지루하고. 같은 음식 영화인데도 바베트의 만찬하고 넘 차이남...바베트의 만찬은 이야기도 있고 음식 보는재미도 있는데 ; 이걸 볼게없다 음식도 별로 안나오고, 핀란드 풍경이라도 구경할래는데 그것도 별로 안나옴 --	disgust	anger	disgust	anger
24	471131	절대 평범한 영화가 아닌 수작이라는걸 말씀드립니다.	happiness	surprise	anticipation	surprise
25	8480268	주제는 좋는데 중반부터 지루하다	disgust	sadness	disgust	sadness
26	4254115	다 짤랐을꺼야. 그래서 납득할 수 없었던거야.. 그럴꺼야.. 꼭 그랬던걸꺼야..	disgust	anger	disgust	anger
27	7295746	kl2g 고추를 털어버려야 할텐데	x	anger	x	anger
28	5457633	카밀라벨 발언기	disgust	anger	disgust	anger
29	6091784	재밌는땡	happiness	happiness	joy	joy
30	8322926	센스있는 연출력..탁월한 캐스팅..90년대의 향수.. 그래서 9점..	happiness	happiness	joy	joy
31	6331922	엄포스의 위력을 다시 한번 깨닫게 해준 적.남 꽃검사님도 연기 정말 좋았어요! 완전 명품드라마!	disgust	happiness	joy	joy
32	9005897	줄쓰레기 진부하고말도안됨ㅋㅋ 아..시간아까워	anger	anger	anger	anger
33	164908	재밌는데 별점이 왜이리 낮은고	happiness	surprise	joy	surprise
34	9671124	1%라도 기대했던 내가 죄인입니다 죄인입니다....	disgust	anger	disgust	anger
35	5712231	아직도 이 드라마는 내인생의 최고!	happiness	happiness	joy	joy
36	4726746	패션에 대한 열정! 안나 원투어!	surprise	happiness	anticipation	joy
37	7203432	키이라 나이틀리가 연기하고자 했던건 대체 정신장애일까 틱장애일까	disgust	anger	disgust	anger
38	112724	허허...원작가 정신나간 유령이라... 재미있겠네요!	happiness	happiness	anticipation	joy
39	1105872	포스터는 있어보이는데 관객은 114명이네	x	sadness	x	sadness
40	4805788	이 영화가 왜 이렇게 저평가 받는지 모르겠다	surprise	sadness	surprise	sadness
41	9336496	단순하면서 은은한 매력의 영화	happiness	happiness	joy	joy
42	9994900	'다 알바생인가 내용도 없고 무서운거도 없고 웃긴거도 하나도 없음 완전 별칭겨운 영화.ㅇㅇ내ㅇ시간 넘 아까움 .. 완전 낡임	disgust	anger	disgust	anger
43	6760577	오게두어라! 서리한이 굶주렸다!	x	x	x	x
44	3440864	정말 맘에 들어요. 그래서 또 보고싶는데 또 보는 방법이 없네? >.. T--	sadness	happiness	sadness	joy
45	8548510	윤계문이라는 멋진 배우를 발견하게 됐어요. 소소한 일탈이 잔잔한 미소를 머금게 합니다. 음악은 조금 아쉽네요ㅠㅠ 8점 주고 싶은데 평점 올리고 싶어 10점 줄게요^^	happiness	happiness	joy	joy
46	7719892	평점에속지마시길시간낭비 돈낭비임	disgust	anger	disgust	aner
47	9440670	리얼리티가 뛰어나긴 한데 큰 공감은 안간다. 이민기캐릭터는 정신의학상 분노조절장애 초기 증상일거다. 투하면 사람패고 욕하고 물건 파손하고.. 조금 오바였음. 극 초반엔 신선했는데 가면 갈수록 이민기 정신상태 공감불가.	disgust	anger	disgust	anger
48	7991636	마이너스는 왜없냐 ㅋ 류비 보고 영화수준 딱 알만하더군 ㅋㅋ 북한에서 이런거 만들라고 돈 대주던?	disgust	anger	disgust	anger
49	4679407	난 우리영화를 사랑합니다....^^;	happiness	happiness	joy	trust

번호	원본id	샘플문장	레이블(E)		레이블(P)	
			주석자 1	주석자 2	주석자 1	주석자 2
50	9837767	데너리스 타르 가르엔...나도 용의주인이 되고 싶다...누이랑,근친상간이나 하고 다닐지라도,소설 속에선 제일 멋진 늑이 자이메 라니스터였는데,드라마속에선,드래곤(용)이 제일 멋지네(웃음)감독님 토르-2 다크 월드는 말아 잡수셨을지라도,기본 선방은 했음	x	happiness	x	joy
51	9063648	영화가 사람의 영혼을 어루만져 줄 수도 있군요 거친 세상사를 잠시 잊고 동화같은 영화에 행복했네요	happiness	happiness	joy	joy
52	8272095	야 세르게이! 작은고추의 매운맛을 보여주마! 포퓰저그 콩진호가 간다	x	happiness	x	joy
53	2345905	이렇게 가슴시리게 본 드라마가 또 있을까? 감동 그 자체!	happiness	happiness	joy	joy
54	7865630	난또 저 꼬마애가 무슨 원한이 깊길래... 했더니 OO 그냥 혼자 나대다 OO걸 어찌라고.	disgust	anger	disgust	anger
55	7207064	재미있어요	happiness	happiness	joy	joy
56	5719655	전 좋아요	happiness	happiness	joy	joy
57	7246040	너무 충격적이었다. 기분을 완전히 폭 꺼지게 하는 느낌... 활력이라고는 하나도 없는 너무나도 무거운...지독하고 차갑고 무자비하다. 그저 일본인들의 상상력은 정말 대단한거 같다는 생각이 든다.	disgust	surprise	disgust	surprise
58	717775	심심한영화.	x	sadness	x	sadness
59	8317483	백봉기 언제나오나요?	x	x	anticipation	x
60	1031725	보는내내 그대로 들어맞는 예측 카리스마 없는 악역	sadness	disgust	sadness	disgust
61	3993146	불알이 나와서 당황...아무튼 영화가 중간에 끝나는 느낌	surprise	anger	surprise	anger
62	2196616	평범함속에 녹아든 평범한 일상. 조금 미미한게 흠.	sadness	sadness	sadness	sadness
63	8203798	보던거라 계속보고있는데 전개도 느리고 주인공인 은희는 한두컷 나오면서 소극적인모습에 짜증이 TT 맨날 언제끝나나 기대만하고있어요 전개좀 빨리빨리 TT	disgust	anger	disgust	anger
64	2332588	사랑하고싶게하는,가슴속온감정을헤집어놓는영화에 오정말최고.	happiness	happiness	joy	joy
65	10084753	많은 사람들이 이 다크를 보고 우리나라 슬픈 현대사의 한 단면에 대해 깊이 생각하고 사회하고 바로 잡기 위해 노력했으면 합니다. 말로만 듣던 보도연맹, 그 민간인 학살이 이정도 일 줄이야. 이건 명백한 살인입니다. 살인자들은 어디있나요?	disgust	anger	disgust	anger
66	8518645	예전 작품 캐릭터, 에피소드 재탕 삼탕 사골우려먹듯 우리고 내용은 산으로 가고 시청률은 아예안나오고 이제 70회중반인데 120부작이라니 ...	disgust	anger	disgust	anger
67	7956793	김남길의 백점짜리 연기력과 초반 몰입도에도 불구하고 지루하고 손에진 스스	disgust	disgust	disgust	disgust
68	3996917	재밌네 비슷한 영화를 안보신 분들한테는 재밌을 듯	happiness	happiness	anticipation	joy
69	8128006	노래실력으로 뽑는게 맞냐? 박시환이 mama나가면 진짜 망신이다	disgust	disgust	disgust	disgust
70	2192046	아 일본영화 다이런건가?? 유치하다	disgust	anger	disgust	anger
71	7250961	이틀만에 다 봤어요 재밌어요 근데 차 안에 물건 넣어 조작하려고 하면 차 안이 열려있던지 집 안이 활짝 열려서 아무나 들어간단던가 문자를 조작하려고하면 비번이 안 걸려있고 ㅋㅋㅋ 그런 건 억지스러웠는데 그래도 내용 자체는 좋았어요	happiness	happiness	joy	joy
72	9179128	재밌네요 달팽이가 빨라서 더 재밌었어요	happiness	happiness	joy	joy
73	9048450	어설픈 전개 어이없는 결말	disgust	anger	disgust	anger
74	8387244	부패한 로마노프 왕조를 기리는 빛같은 영화... 온몸으로 항거했던 러시아 민중들이 그저 폭도냐	anger	anger	anger	anger
75	9803156	내용전개는 무난한 편이었구 잘 보았습니다 ^^	happiness	happiness	joy	joy
76	2135015	매우 실망.....	disgust	sadness	disgust	sadness
77	7884025	한국영화 흥행코드: 갈등-갈등-계-에속 갈등-화해-감동- 평점 10점 남발- 흥행 뻥하지 뭐...	disgust	sadness	disgust	sadness
78	9801316	아햏햏 아햏햏 아햏햏.	x	happiness	x	joy

번호	원본id	샘플문장	레이블(E)		레이블(P)	
			주석자 1	주석자 2	주석자 1	주석자 2
79	2009382	뭐냐.. 시작하고 3분만에 나왔다. 리플릿 사진 보며 불안하더니만..	anger	anger	anger	anger
80	10112591	단연 최고라고 할수있지	happiness	happiness	joy	joy
81	7280935	감독이 럽먹고 영화를 만들었나보다.. 관객에게 뭘 말하는지도 모르겠고, 엉망진창 개진창이다.	anger	anger	anger	anger
82	5971506	이건 뭐냐? 우리매냐? ;;;	disgust	anger	disgust	anger
83	9875922	정말 쓰레기영화입니다	anger	disgust	anger	disgust
84	9850710	진정 위대한 영화 최고임	happiness	happiness	joy	joy
85	1927486	별루 였다..	disgust	anger	disgust	anger
86	5023820	내일이 기대되는 ‘	happiness	happiness	anticipation	anticipation
87	1859380	근데 조미가 막문위 좋아한건가요??	x	x	x	x
88	1635552	ㅋㅋㅋ 진짜 골감..ㅋㅋ 눈 부라릴때 쓰러짐..ㅋㅋ	x	happiness	x	joy
89	7852366	성룡영화중 최악인듯 ㅋㅋ	disgust	disgust	disgust	disgust
90	9480003	골때리네ㅋㅋㅋㅋ 걸스데이 이혜리 잘 되라!	happiness	happiness	joy	joy
91	1797212	서기가이빠서	happiness	happiness	joy	joy
92	6374262	완전 재밌어요ㅋㅋㅋㅋ백인공주귀여움ㅋㅋㅋㅋㅋㅋ	happiness	happiness	joy	joy
93	1607129	인상적인 영화였다	happiness	happiness	joy	joy
94	8763660	어내 스트와 셀레스틴 완전 강추예요~ 정말 재밌습니다^^	happiness	happiness	joy	joy
95	9361974	재미있는영화입니다.	happiness	happiness	joy	joy
96	7928957	클라라볼라고화신본거아닌데	disgust	anger	disgust	anger
97	10250221	진짜 보면서 너무 슬펐던 영화다	sadness	sadness	sadness	sadness
98	6122458	설정이 재밌고 새로운 에피소드 내에서 메인 스토리도 차차 나오는게 재밌음	happiness	happiness	joy	joy
99	10044377	신카이 마코토의 작화와, 미유와 하나카나가 연기를 잘해줘서 더대박이었다.	happiness	happiness	joy	joy
100	6158844	재미없음 진심 1이월나 캐스팅두못한듯	disgust	anger	disgust	anger

참고 문헌

- 안세훈, 정옥란. "감정분석 기반 심리상담 AI 챗봇 시스템에 대한 연구." (2021).
- 이상아, 신호필. "감정 분석을 위한 BERT 사전학습모델과 추가 자질 모델의 결합." 정보과학 회논문지 48.7 (2021): 815-824.
- 이현영, 강승식. "Out-of-Vocabulary 단어에 강건한 병렬 Tri-LSTM 문장 임베딩을 이용한 감정분석." 스마트미디어저널 10 (2021): 16-24.
- Ekman, Paul. "Expression and the nature of emotion." Approaches to emotion 3.19 (1984): 344.
- Ekman, Paul, Richard J. Davidson, and Wallace V. Friesen. "The Duchenne smile: emotional expression and brain physiology: II." Journal of personality and social psychology 58.2 (1990): 342.
- Ekman, Paul. "An argument for basic emotions." Cognition & emotion 6.3-4 (1992): 169-200.
- Garcia-Garcia, Jose Maria, Victor MR Penichet, and Maria D. Lozano. "Emotion detection: a technology review." Proceedings of the XVIII international conference on human computer interaction. 2017.
- Huang, Chenyang, Amine Trabelsi, and Osmar R. Zaïane. "Ana at semeval-2019 task 3: Contextual emotion detection in conversations through hierarchical lstms and bert." arXiv preprint arXiv:1904.00132 (2019).
- Kim, Evgeny, and Roman Klinger. "A survey on sentiment and emotion analysis for computational literary studies." arXiv preprint arXiv:1808.03137 (2018).
- Lai, Yuni, et al. "Fine-grained emotion classification of Chinese microblogs based on graph convolution networks." World Wide Web 23.5 (2020): 2771-2787.
- Mohammad, Saif, et al. "SemEval-2018 Task 1: Affect in Tweets." Proceedings of The 12th International Workshop on Semantic Evaluation (2018): 1-17.
- Russell, James A. "A circumplex model of affect." Journal of personality and social psychology 39.6 (1980): 1161.
- Pang, Bo, et al. "Thumbs up? Sentiment Classification using Machine Learning Techniques." In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (2002): 79-86.
- Plutchik, Robert. The emotions. University Press of America, 1991.
- Wang, Xiangyu, and Chengqing Zong. "Distributed representations of emotion categories in emotion space." Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021.

<Abstract>

Sentiment analysis corpus research and construction in 2021

The purpose of this project is to build large-scale corpus data for the development and utilization of artificial intelligence technology in preparation for the 4th Industrial Revolution and to propose the future direction of the emotion and sentiment analysis corpus construction project conducted by the National Institute of Korean Language (NIKL). As a follow-up project to the 'Sentiment analysis corpus research and construction' in 2020, this project intends to supplement the limitations of the corpus of 2020 and to improve the annotation quality of the dataset. In addition, one purpose of this project is to suggest future directions of constructing emotion detection corpora of NIKL in terms of support related research and industrial communities. The achievements of this project are summarized as follows.

First, We developed the guidelines for the Korean aspect-based sentiment analysis corpus construction. We classified domains in which a corpus would be more suitable for analysis, and construct entity labels and attribute labels corresponding to each domain in order to enable more precise sentiment analysis.

Second, we modified and enhanced our sentiment analysis workbench tool from 2020 to re-analyze the sentiment analysis corpus. Furthermore, we additionally collected and obtained test document data for key domains and carried out the same analysis procedures to verify that these test data would be apt for evaluating AI performance.

Third, we examined the latest trends in the field of sentiment analysis by surveying the domestic as well as international constructions and utilizations of sentiment analysis corpora in both academia and industry. Moreover,

through interviews with experts in AI, NLP (Natural Language Processing), and data mining fields, we gained insights into trends in each field and sought advice on the future direction of the NIKL's sentiment analysis corpus construction project.

Fourth, we present the future direction of the NIKL's Emotion Detection corpus construction project. We investigated data types, sizes, and formats that would be practically useful in academia and industry through domestic and international trend surveys and expert interviews, and we propose a detailed criteria for future emotion and sentiment analysis corpus construction projects.

Key words: Sentiment analysis, Emotion detection, Sentiment analysis corpus construction, Aspect based sentiment analysis, Future direction of Sentiment and Emotion analysis corpora

<기획·연구>

국립국어원 이승재 학예연구관(언어정보과장)
국립국어원 김소희 학예연구사
국립국어원 이민주 연구원

<사업 참여자>

사업 책임자 이영희 (㈜버즈메트릭스 이사)
사업 참여자 김수진 (㈜버즈메트릭스 부장)
정종서 (㈜버즈메트릭스 과장)
신현주 (㈜버즈메트릭스 부장)
유지현 (㈜버즈메트릭스 주임)
박나경 (㈜버즈메트릭스 대리)
추승아 (㈜버즈메트릭스 주임)
옥승희 (㈜버즈메트릭스 대리)
서은미 (㈜버즈메트릭스 대리)
김도현 (㈜버즈메트릭스 과장)
이진상 (㈜버즈메트릭스 주임)

발행인: 국립국어원장

발행처: 국립국어원

서울시 강서구 금남화로 154

전화 02-2669-9775, 전송 02-2669-9727

인쇄일: 2021년 12월 06일

발행일: 2021년 12월 06일

인 쇄: (주)타라그래픽스

※ “이 책은 국립국어원의 용역비로 수행한 ‘2021 말뭉치 감성 분석 및 연구’
사업의 결과물을 발간한 것입니다.”