국립국어원 2022-01-44

발 간 등 록 번 호
11-1371028-000929-01

2022년 말뭉치 감정 분석 및 연구

사업 책임자
이 영 희

국립국어원

제 출 문

국립국어원장 귀하

국립국어원과 체결한 연구용역 계약에 따라 ‘2022년 말뭉치 감정 분석 및 연구’에 관한 연구 보고서를 작성하여 제출합니다.

■ 사업기간: 2022년 5월 2일 ~ 2022년 12월 2일

2022년 12월 2일

사업 책임자: 이 영 희((주)버즈메트릭스)

사업 수행자 (주)버즈메트릭스

사업 책임자 이영희

사업 참여자 김수진, 유지현, 이준수, 권주원,
신현주, 김도현, 이진상, 한희재, 서은미

<국문 요약>

2022년 말뭉치 감정 분석 및 연구

본 사업은 4차 산업혁명을 대비하여 인공지능 기술을 개발하고 활용하기 위해 대규모 말뭉치 데이터를 구축하는 것을 목표로 한다. 대규모 말뭉치 데이터 구축을 통해 언어 정보 표준화에 기여하고 참조 기반 자료가 될 수 있는 분석 말뭉치를 확보하여 제공 배포함으로써 국어 자원의 활용도와 가치를 제고하는데 그 목적이 있다. 이에 따른 주요 사업 내용을 요약하면 다음과 같다.

첫째, 말뭉치 감정 및 공격성 분석을 위한 자료를 수집하고 선별하였다. 5어절 이상의 문화콘텐츠 분야 트위터 자료 50,000건을 신규로 수집 및 선별하고, 국립국어원에서 배포한 '2020 일상 대화 말뭉치'에서 5어절 이상의 일상 대화 담화 문서 10,000건을 선별하였다. 트위터 자료의 경우, 사업 참여자를 모집하여 트위터 게시 자료에 대한 저작권 이용 허락 계약을 체결하고, 계약 체결이 완료된 참여자의 트위터 게시 자료를 수집하였다.

둘째, 말뭉치 감정 및 공격성 분석을 위한 방법론 및 지침을 마련하였다. 감정 분석 방법론은 로버트 플루칙(Robert Plutchik)의 감정 체계를 차용하되 8가지 기본 감정(기쁨, 기대, 신뢰, 놀람, 혐오, 공포, 분노, 슬픔)을 감정 분석의 표지로 활용하였다. 8가지 기본 감정 스펙트럼에서 벗어나 어떠한 표지로도 정의될 수 없을 경우 '기타' 표지로 태깅함으로써, 총 9개 감정 표지를 기준으로 분석하였다. 트위터 자료의 경우, 공격성이 드러날 경우 공격성 범주를 파악하여 분석하였다. 이 때 공격성이 향하는 대상(target)인 공격 대상 범주 역시 함께 분석하고 비식별화 처리를 하였다.

셋째, 분석 방법론과 지침에 따라 감정 분석을 진행하고 말뭉치를 구축하였다. 감정 분석을 위해 자료 1건당 분석 인원을 10명으로 정하고, 주석자 간 일치도 기준에 따라 적당한 일치도 이상값인 문서를 채택하고, 채택 문서 내 5명 이상의 분석 결과가 일치한 감정 표지를 채택하여 말뭉치를 구축하였다. 말뭉치는 json 형식으로 변환하여 트위터 50,000건, 일상 대화 10,000건에 대한 감정·공격성 분석 말뭉치를 납품하였다.

주요어: 감정 분석, 감정 분석 말뭉치 구축, 감정 분석 말뭉치 활용, 감정 및 공격성 분석

<Abstract>

2022 Corpus Emotion Analysis and Research

The purpose of this project is to build large-scale corpus data to develop and utilize artificial intelligence technology in preparation for the Fourth Industrial Revolution. The purpose is to enhance the utilization and value of Korean resources by securing and distributing analytical corpus that can contribute to standardizing language information through the construction of large-scale corpus data. The summary of the major projects accordingly is as follows.

First, data for the analysis of corpus emotion and aggression were collected and selected. 50,000 new Twitter data in the field of cultural contents were collected and selected, and 10,000 daily conversation discourse documents with more than 5 words were selected from the '2020 Daily conversation Corpus Construction' distributed by the National Institute of Korean Language. In the case of Twitter data, project participants were recruited to sign a copyright permission contract for Twitter posting materials, and Twitter posting data of participants who completed the contract were collected.

Second, methodologies and guidelines for analyzing corpus emotion and aggression were prepared. The emotion analysis methodology borrowed Robert Plutchik's emotion system, but used eight basic emotions (joy, anticipation, trust, surprise, disgust, fear, anger, and sadness) as a cover for emotion analysis. If it cannot be defined as any sign, away from the eight basic emotion spectra, it was analyzed based on a total of nine emotion signs by tagging them with the 'other' sign. In the case of Twitter data, when aggression was revealed, the aggression category was identified and analyzed. At this time, the target category, which is the target of aggression, was also tagged and anonymized.

Third, emotional analysis was conducted according to the analysis methodology and guidelines and a corpus was constructed. For emotional analysis, the number of analysis personnel per data was set to 10, documents with an appropriate degree of agreement or more were adopted according to the criteria for agreement between annotators, and emotion signs with the results of more than five analysis in the adopted document were adopted to build a corpus. The corpus was converted into the json format and delivered an emotional and aggressive analysis corpus for 50,000 Twitter cases and 10,000 daily conversations.

Key-words: emotion analysis, emotion analysis corpus construction, emotion analysis corpus utilization, emotion and aggression analysis

<개조식 요약문>

2022년 말뭉치 감정 분석 및 연구

1. 연구 목적

○ 4차 산업혁명 대비 기반 기술 개발 및 인공지능 기술 개발·활용을 위한 대규모 말뭉치 구축

○ 대규모 말뭉치 구축을 통해 언어 정보 표준화 및 참조 기반 자료가 될 수 있는 분석 말뭉치 제공·배포

2. 주요 사업 내용

(1) 말뭉치 감정·공격성 분석을 위한 자료 수집 및 선별
- 5어절 이상의 문화콘텐츠 분야 트위터 50,000건 신규 수집 및 선별
  (참여자 모집 및 트위터 게시 자료에 대한 저작권 이용 허락 체결 완료 포함)
- 국립국어원 배포 '2020 일상 대화 말뭉치'에서 5어절 이상 대화 담화 문서 10,000상 선별

(2) 말뭉치 감정·공격성 분석 방법론 및 지침 마련
- 로버트 플루칙(Robert Plutchik)의 감정 체계를 차용하여 9개 감정(기쁨, 기대, 신뢰, 놀람, 혐오, 공포, 분노, 슬픔, 기타) 표지를 기준으로 말뭉치 감정·공격성 지침 수립 및 대상(target) 유형 표지 정의
- 말뭉치 감정·공격성 분석의 대상(target) 비식별화

(3) 분석 방법론 및 지침에 따라 감정 분석 및 말뭉치 구축
- 말뭉치 감정·공격성 분석 지침에 따라 분석 진행
- 트위터 50,000건, 일상 대화 10,000건에 대해 감정·공격성 분석 말뭉치 구축

차 례

차 례

표 차례

그림 차례

제 1 장

서 론

1. 사업 목적

본 사업은 4차 산업혁명을 대비하여 인공지능 기술을 개발하고 활용하기 위해 대규모 말뭉치 데이터를 구축하는 것을 목표로 한다. 대규모 말뭉치 데이터 구축을 통해 언어 정보 표준화에 기여하고 참조 기반 자료가 될 수 있는 분석 말뭉치를 확보하여 제공 배포함으로써 국어 자원의 활용도와 가치를 제고하는 데 그 목적이 있다.

2. 사업 수행 범위

본 사업은 위와 같은 사업의 목적에 따라 다음과 같은 수행 범위로 구성되어 있다.

ㅇ 말뭉치 감정·공격성 분석을 위한 자료 수집 및 선별
- 문화콘텐츠 분야 트위터 자료(5어절 이상 문서) 50,000건을 신규 수집 및 선별
- 국립국어원 배포 말뭉치에서 일상 대화 자료(5어절 이상 담화를 선·후행 맥락과 함께 구성) 10,000건을 선별

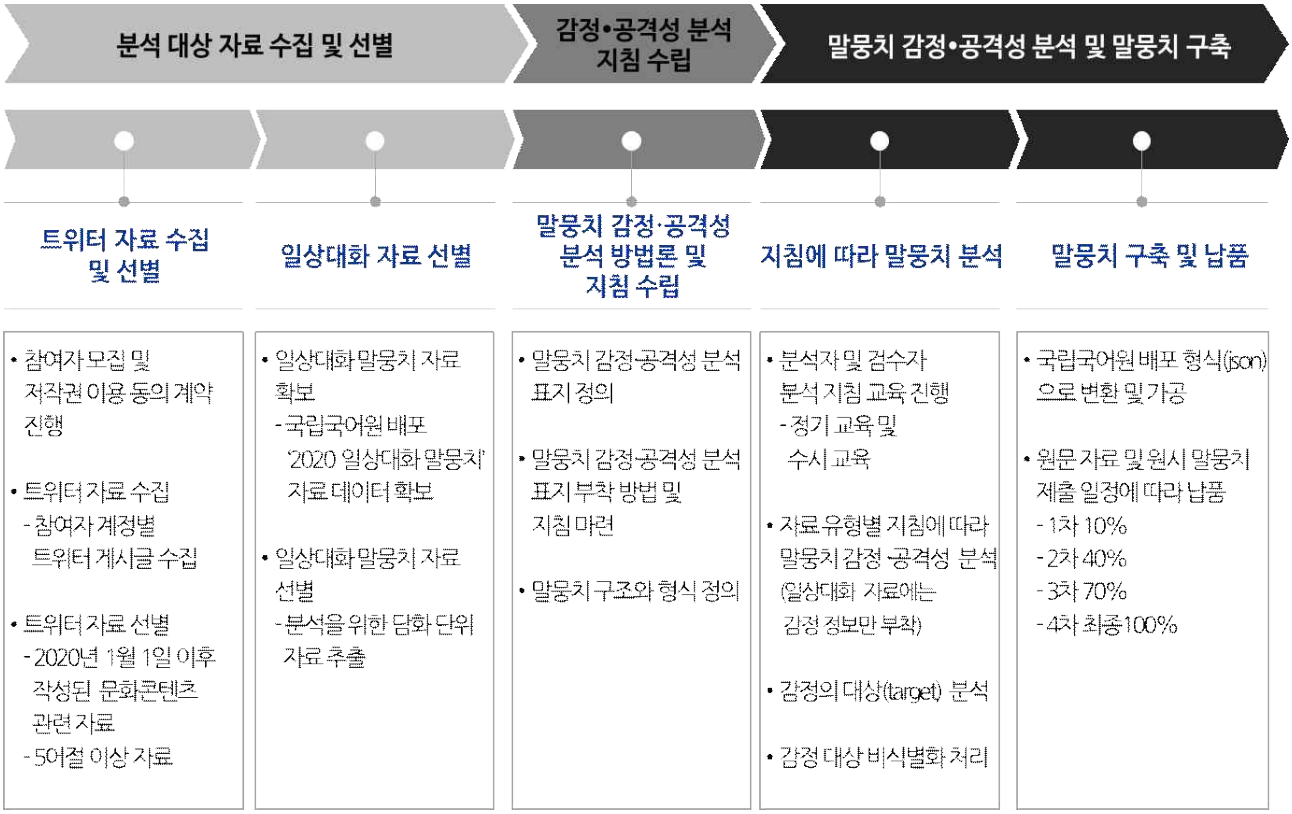
ㅇ 말뭉치 감정·공격성 분석 방법론 및 지침 마련
- 담화 단위 선정 기준 및 정제·가공 지침 수립
- 말뭉치 감정·공격성 분석을 위한 분석 표지 정의
- 감정·공격성 대상이 되는 개인·집단 정보 익명화 방법론 정의
- 말뭉치에 감정·공격성 표지를 부착하기 위한 세부 기준 수립

ㅇ 분석 방법론 및 지침에 따라 자료를 수집·선별·분석하여 말뭉치 구축
- 수립된 방법론과 지침에 따라 말뭉치 감정·공격성 분석 및 지침 보완
- 트위터 50,000건, 일상 대화 10,000건에 대해 감정·공격성을 분석한 말뭉치를 구축하여 최종 납품

3. 사업 수행 절차

본 사업은 트위터와 일상 대화 말뭉치에서 분석 대상 자료를 수집 및 선별하고, 말뭉치 감정·공격성 분석 지침을 수립하여 감정 분석 방법론과 지침에 따라 감정 분석을 진행하고 말뭉치를 구축하는 절차로 이루어졌다. 단계별 주요 수행 내용은 다음과 같다.

<그림 1> 사업 수행 절차



제 2 장

말뭉치
분석 자료
수집 및 선별

1. 트위터 자료 수집 및 선별

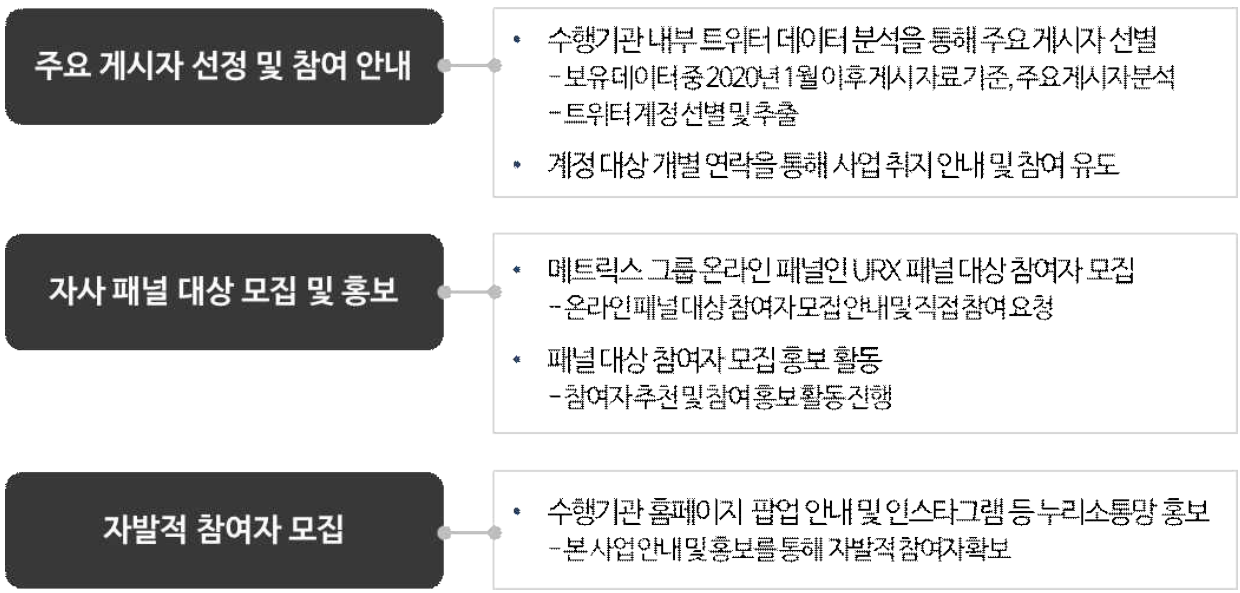
트위터 자료는 국립국어원에서 배포할 수 있도록 저작권 이용 동의 계약 완료가 반드시 필요하다. 이를 위해, 본 사업에 대한 참여자를 모집하고 참여자 대상 저작권 이용 동의 계약을 완료한 자료에 한하여 수집 및 선별 작업을 진행하였다.

또한 트위터 자료는 5어절 이상의 문화콘텐츠 관련 자료로, 2020년 1월 1일 이후 작성된 문서 5만 건을 대상으로 하였다.

1-1. 참여자 모집 및 선정

트위터 자료는 5만 건을 목표로 하였으나, 문화콘텐츠 자료면서, 5어절 이상인 문서만을 대상으로 하므로, 실제로 더 많은 문서를 확보해야 목표한 5만 건 확보가 가능할 것으로 예상하였다. 목표 건수 대비 최소 5배수 이상의 트위터 자료를 확보하기 위해 최소 200명 이상의 참여자 확보가 필요하였다. 다수의 트위터 자료 참여자 모집을 위해 다음과 같은 다양한 모집 방법을 통해 목표한 문서를 확보하고자 하였다.

<그림 2> 사업 참여자 모집 방법



첫 번째 참여자 모집 방법인 '주요 게시자 선정 및 참여 유도'는 사업 수행 기관의 트위터 데이터를 적극적으로 활용하여 이루어졌다. 내부에서 보유하고 있는 트위터 자료를 검토하여 참여 조건에 부합하는 게시 자료를 보유한 참여자를 찾는 방법으로 진행하였다. 내부 축적된 자료 중, 2020년 1월 1일 이후 작성된 게시 자료 50만 건을 분석하여, 22년 게시글 기준 50건 이상의 문화콘텐츠 관련 게시 자료를 작성한 작성자 계정을 추출하였다. 22년 게시글 기준으로 추출한 이유는 본 사업의 수집 대상이 2020년 1월 1일 이후 게시 자료 기준이며, 현재 활발하게 트위터 활동을 하고 있는 참여자를 선별하기 위함이다. 작성자 계정 선별 과정을 거쳐 사업의 목적과 취지에 대해 안내하고 참여를 유도하는 방식으로 진행하였다.

두 번째로는 사업 수행 기관의 온라인 패널인 URX(메트릭스 그룹 온라인 패널) 회원을 대상으로 참여 안내와 홍보 활동을 진행하였다. URX 패널은 2022년 5월 기준 약 130만 명이 회원으로 가입되어 있어, 사업 참여 유도와 사업에 대한 홍보가 동시에 이루어지기 용이하였다. 또한, 패널 회원 본인이 직접 참여하는 것만으로는 제한된 시간 내에 충분한 모집이 이루어지기 어렵기 때문에, 사업을 홍보하고 주변인을 추천하는 방식을 병행하여 진행하였다.

세 번째 방법은 자발적 참여를 통한 참여자 모집이다. 국립국어원 홈페이지 게시판에 사업 관련 안내 공고문을 게시하였고 사업 수행 기관 홈페이지를 통해 사업 안내를 진행하였으며, 인스타그램에 사업 참여에 대한 홍보 메시지를 게시함으로써 사업의 취지에 공감한 참여자의 자발적 참여를 유도하였다.

1-2. 저작권 이용 허락 계약 체결

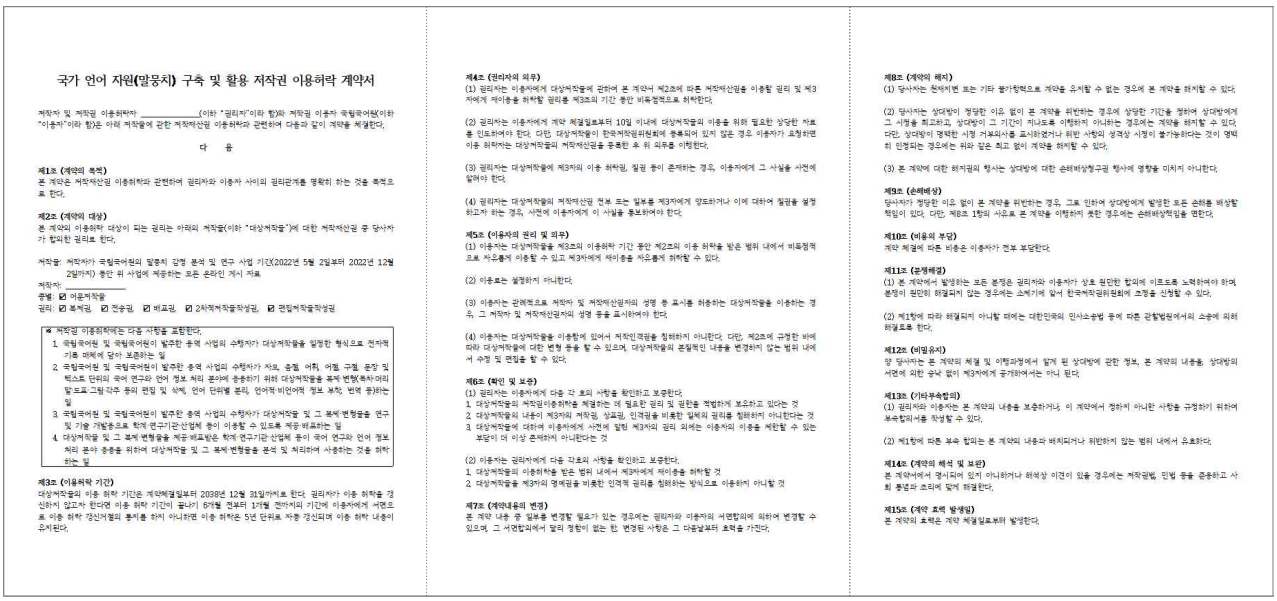
본 사업 참여자는 트위터 자료를 작성한 원문 자료의 저작권자로, 참여자와의 저작권 이용 허락 계약 체결이 필요하다. 특히, 본 사업은 감정 분석 말뭉치를 구축하는 일뿐만 아니라 복제·변형물을 연구 및 기술 개발용으로 학계와 연구기관 및 산업체 등이 이용할 수 있도록 제공 및 배포하는 것을 목적으로 하므로, 저작권자로부터 저작권 이용 허락 계약이 필수적으로 이루어져야 한다. 향후 발생할 수 있는 법률적 분쟁을 최소화하고 민간 활용도를 제고하기 위해 저작권 이용 허락 계약 체결은 본 사업에서 필수적인 과정이다.

저작권 이용 하락 계약서 양식 및 내용은 법률 검토를 거친 후 최종 확정하였으며, 본 사업의 특성에 따라 대상 권리의 내용은 복제권, 전송권, 배포권, 2차적 저작물 작성권, 편집 저작물 작성권을 포함하였다. 저작권 이용 하락 계약에 사용된 계약서의 세부 내용은 다음과 같다.

※ 저작권 이용허락에는 다음 사항을 포함한다.

1. 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 대상저작물을 일정한 형식으로 전자적 기록 매체에 담아 보존하는 일
2. 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 자모, 음절, 어휘, 어절, 구절, 문장 및 텍스트 단위의 국어 연구와 언어 정보 처리 분야에 응용하기 위해 대상저작물을 복제·변형(목차·머리말·도표·그림·각주 등의 편집 및 삭제, 언어 단위별 분리, 언어적·비언어적 정보 부착, 번역 등)하는 일
3. 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 대상저작물 및 그 복제·변형물을 연구 및 기술 개발을 목적으로 학계·연구기관·산업체 등이 이용할 수 있도록 제공·배포하는 일
4. 대상저작물 및 그 복제·변형물을 제공·배포받은 학계·연구기관·산업체 등이 국어 연구와 언어 정보 처리 분야 응용을 위하여 대상저작물 및 그 복제·변형물을 분석 및 처리하여 사용하는 것을 허락하는 일

<그림 3> 저작권 이용 허락 계약서



국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용허락 계약서

저작자 및 저작권 이용허락자 _____(이하 "권리자"이라 함)와 저작권 이용자 국립국어원(이하 "이용자"이라 함)은 아래 저작물에 관한 저작재산권 이용허락과 관련하여 다음과 같이 계약을 체결한다.

다 음

제1조 (계약의 목적)
본 계약은 저작재산권 이용허락과 관련하여 권리자와 이용자 사이의 권리관계를 명확히 하는 것을 목적으로 한다.

제2조 (계약의 대상)
본 계약의 이용허락 대상이 되는 권리는 아래의 저작물(이하 "대상저작물")에 대한 저작재산권 중 당사자가 합의한 권리로 한다.

저작물: 저작자가 국립국어원의 말뭉치 갱신 분석 및 연구 사업 기간(2022년 5월 2일부터 2022년 12월 2일까지) 중인 위 사업에 제공하는 모든 온라인 게시 자료
저작자: _____
종별: ☑ 어문저작물
권리: ☑ 복제권 ☑ 전송권 ☑ 배포권 ☑ 2차적저작물작성권 ☑ 편집저작물작성권

※ 저작권 이용허락에는 다음 사항을 포함한다.
1. 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 대상저작물을 일정한 형식으로 전자적 기록 매체에 담아 보존하는 일
2. 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 자모, 음절, 어휘, 어절, 구절, 문장 및 텍스트 단위의 국어 연구와 언어 정보 처리 분야에 응용하기 위해 대상저작물을 복제·변형(목차·머리말·도표·그림·각주 등의 편집 및 삭제, 언어 단위별 분리, 언어적·비언어적 정보 부착, 번역 등)하는 일
3. 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 대상저작물 및 그 복제·변형물을 연구 및 기술 개발을 목적으로 학계·연구기관·산업체 등이 이용할 수 있도록 제공·배포하는 일
4. 대상저작물 및 그 복제·변형물을 제공·배포받은 학계·연구기관·산업체 등이 국어 연구와 언어 정보 처리 분야 응용을 위하여 대상저작물 및 그 복제·변형물을 분석 및 처리하여 사용하는 것을 허락하는 일

제3조 (이용허락 기간)
대상저작물의 이용 허락 기간은 계약체결일로부터 2038년 12월 31일까지로 한다. 권리자가 이용 허락을 갱신하지 않고자 한다면 이용 허락 기간이 끝나기 6개월 전부터 1개월 전까지의 기간에 이용자에게 서면으로 이용 허락 갱신거절의 통지를 하지 아니하면 이용 허락은 5년 단위로 자동 갱신되며 이용 허락 내용이 유지된다.

제4조 (권리자의 의무)
(1) 권리자는 이용자에게 대상저작물에 관하여 본 계약나 제2조에 따른 저작재산권을 이용할 권리 및 제3자에게 재이용을 허락할 권리를 제2조의 기간 중인 비독점적으로 허락한다.

(2) 권리자는 이용자에게 계약 체결일로부터 10일 이내에 대상저작물의 이용을 위해 필요한 상당한 자료를 인도하여야 한다. 다만, 대상저작물이 한국저작권위원회에 등록되어 있지 않은 경우 이용자가 요청하면 이용 허락하는 대상저작물의 저작재산권을 등록한 후 위 의무를 이행한다.

(3) 권리자는 대상저작물에 제3자의 이용 허락권, 질권 등이 존재하는 경우, 이용자에게 그 사실을 사전에 알려야 한다.

(4) 권리자는 대상저작물의 저작재산권 전부 또는 일부를 제3자에게 양도하거나 이에 대하여 질권을 설정하고자 하는 경우, 사전에 이용자에게 이 사실을 통보하여야 한다.

제5조 (이용자의 권리 및 의무)
(1) 이용자는 대상저작물을 제2조의 이용허락 기간 동안 제2조의 이용 허락을 받은 범위 내에서 비독점적으로 자유롭게 이용할 수 있고 제3자에게 재이용을 자유롭게 허락할 수 있다.

(2) 이용료는 설정하지 아니한다.

(3) 이용자는 관례적으로 저작자 및 저작재산권자의 성명 등 표시를 허용하는 대상저작물을 이용하는 경우, 그 저작자 및 저작재산권자의 성명 등을 표시하여야 한다.

(4) 이용자는 대상저작물을 이용함에 있어서 저작인격권을 침해하지 아니한다. 다만, 제2조에 규정한 바에 따라 대상저작물에 대한 변형 등을 할 수 있으며, 대상저작물의 본질적인 내용을 변경하지 않는 범위 내에서 수정 및 편집을 할 수 있다.

제6조 (확인 및 보증)
(1) 권리자는 이용자에게 다음 각 호의 사항을 확인하고 보증한다.
1. 대상저작물의 저작권이용허락을 체결하는 데 필요한 권리 및 권한을 적법하게 보유하고 있다는 것
2. 대상저작물의 내용이 제3자의 저작권, 상표권, 인격권을 비롯한 일체의 권리를 침해하지 아니한다는 것
3. 대상저작물에 대하여 이용자에게 사전에 알린 제3자의 권리 외에는 이용자의 이용을 제한할 수 있는 보입이 더 이상 존재하지 아니한다는 것

(2) 이용자는 권리자에게 다음 각호의 사항을 확인하고 보증한다.
1. 대상저작물의 이용허락을 받은 범위 내에서 제3자에게 재이용을 허락할 것
2. 대상저작물을 제3자의 명예훼손을 비롯한 인격적 권리를 침해하는 방식으로 이용하지 아니할 것

제7조 (계약내용의 변경)
본 계약 내용 중 일부를 변경할 필요가 있는 경우에는 권리자와 이용자의 서면합의에 의하여 변경할 수 있으며, 그 서면합의에서 달리 정함이 없는 한 변경된 사항은 그 다음날부터 효력을 가진다.

제8조 (계약의 해지)
(1) 당사자는 천재지변 또는 기타 불가항력으로 계약을 유지할 수 없는 경우에 본 계약을 해지할 수 있다.

(2) 당사자는 상대방이 정당한 이유 없이 본 계약을 위반하는 경우에 상당한 기간을 정하여 상대방에게 그 시정을 최고하고, 상대방이 그 기간이 지나도록 이행하지 아니하는 경우에는 계약을 해지할 수 있다. 다만, 상대방이 명백한 시정 거부의사를 표시하였거나 위반 사항의 성격상 시정이 불가능하다는 것이 명백히 인정되는 경우에는 위와 같은 최고 없이 계약을 해지할 수 있다.

(3) 본 계약에 대한 해지권의 행사는 상대방에 대한 손해배상청구권 행사에 영향을 미치지 아니한다.

제9조 (손해배상)
당사자가 정당한 이유 없이 본 계약을 위반하는 경우, 그로 인하여 상대방에게 발생한 모든 손해를 배상할 책임이 있다. 다만, 제8조 1항의 사유로 본 계약을 이행하지 못한 경우에는 손해배상책임을 면한다.

제10조 (비용의 부담)
계약 체결에 따른 비용은 이용자가 전부 부담한다.

제11조 (분쟁해결)
(1) 본 계약에서 발생하는 모든 분쟁은 권리자와 이용자가 상호 합의한 합의에 이르도록 노력하여야 하며, 분쟁이 합의에 해결되지 않는 경우에는 소송기에 앞서 한국저작권위원회에 조정을 신청할 수 있다.

(2) 제1항에 따라 해결되지 아니할 때에는 대한민국의 민사소송법 등에 따른 관할법원에서의 소송에 의해 해결토록 한다.

제12조 (비밀유지)
양 당사자는 본 계약의 체결 및 이행과정에서 알게 된 상대방에 관한 정보, 본 계약의 내용을 상대방의 서면에 의한 승낙 없이 제3자에게 공개하여서는 아니 된다.

제13조 (기타부속합의)
(1) 권리자와 이용자는 본 계약의 내용을 보충하거나, 이 계약에서 정하지 아니한 사항을 규정하기 위하여 부속합의서를 작성할 수 있다.

(2) 제1항에 따른 부속 합의는 본 계약의 내용과 배치되거나 위반하지 않는 범위 내에서 유효하다.

제14조 (계약의 해석 및 보완)
본 계약서에서 명시되어 있지 아니하거나 해석상 이견이 있을 경우에는 저작권법, 민법 등을 준용하고 사회 통념과 조리에 맞게 해결한다.

제15조 (계약 효력 발생일)
본 계약의 효력은 계약 체결일로부터 발생한다.

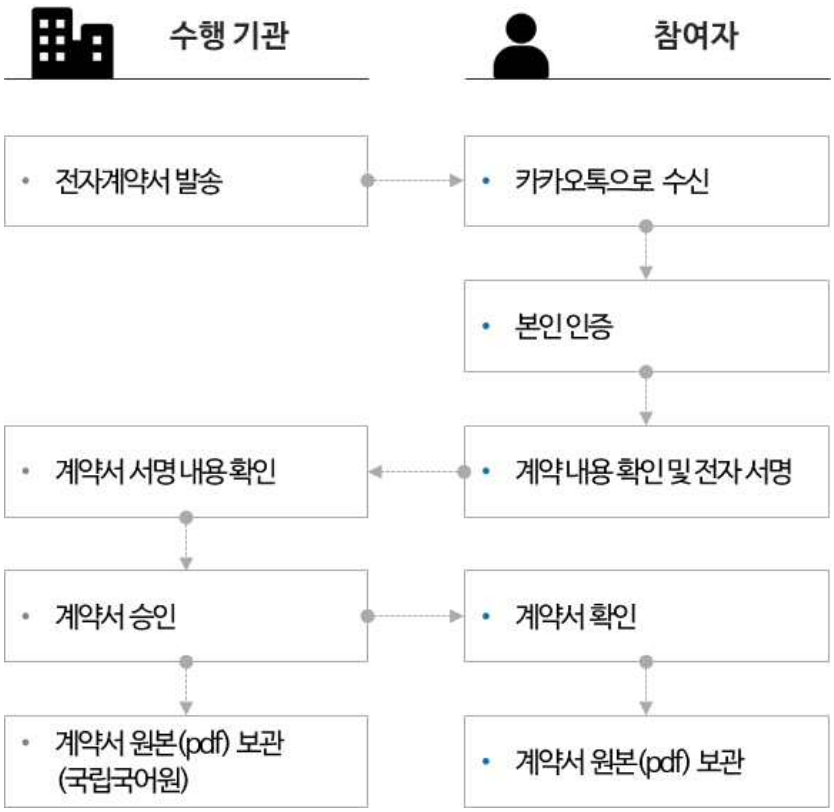
저작권 이용 허락 기간은 계약 체결일부터 2038년 12월 31일까지로 하되, 이용 허락 기간이 끝나기 6개월 전부터 1개월 전까지의 기간에 이용자에게 서면으로 이용 허락 갱신 거절의 통지를 하지 아니하면 이용 허락은 5년 단위로 자동 갱신되며 이용 허락 내용이 유지된다.

계약 진행 과정과 계약서 원본 관리 과정에서 개인정보 유출 등의 문제를 최소화하기 위해, 계약서 서명란에 주소나 연락처, 주민번호 뒷자리 등의 정보는 포함하지 않도록 하였다.

저작권 이용 허락 계약 체결은 전자 계약 시스템을 이용하였다. 전자 계약 시스템을 통해 본인 인증 절차를 거침으로써, 참여자 본인에 대한 인증은 물론, 계약서상 개인정보를 기재해야 하는 번거로움을 최소화할 수 있었다.

사업 참여를 위해 저작권 이용 동의 계약을 완료한 참여자는 총 328명이다.[1)]

<그림 4> 저작권 이용 허락 전자 계약 진행 절차



1) 본 사업을 위해 저작권 이용 동의 계약을 완료한 참여자는 총 336명이나 아래 <표 1>의 자료 선별 기준에 따라 8명의 계약자는 제외하고 최종 328명의 계약자만을 선정하였다.

1-3. 트위터 자료 수집 및 선별

저작권 이용 허락 계약이 완료된 참여자의 트위터 계정을 대상으로 자료를 수집하였다. 자료의 수집은 사업 수행 기관 자체 개발 수집기인 'Buzz Crawler'를 이용하여 수집하였다. 본 사업의 경우 참여자의 계정 또는 게시물 단위로 게시물을 수집하는 방식으로 진행하였다.

트위터 자료의 선별은 5어절 이상의 문화콘텐츠를 기준으로 하였으며, 추가적인 자료 선별 기준 및 정제 기준은 다음과 같다.

<표 1> 트위터 자료 선별 및 정제 기준

선별 대상	세부 기준
문화콘텐츠 관련 문서	영화/드라마/방송, 공연/전시/박람회, 도서/문학, 게임, 캐릭터, 음악/음반/콘서트, 연예인/유명인/팬덤/팬클럽 관련 게시글
포함 대상 기간	2020년 1월 1일 이후 작성 게시글
5어절 이상 문서	5어절 이상으로 구성된 게시글
비문서, 비국문 자료 제외	이미지, 스티커, 사진, 동영상, 파일 링크, 웹 주소, 해시태그로만 구성된 게시 자료 삭제 전문 외국어로 구성된 게시 자료 삭제
중복글, 펌글, 홍보글 제외	중복 게시 자료 삭제, 펌글(기사, 타인이 작성한 게시물 등)로만 구성된 게시 자료, 상업적 광고가 포함된 자료 삭제

실제 수집된 자료는 총 532,373건이었으며, 트위터 자료 선별 및 정제 기준에 따라 최종 선별된 문서는 86,770건으로, 전체 수집 자료의 16%만 유효 문서로 분류되었다. 즉, 5만 건의 유효한 분석 대상 문서를 확보하기 위해서는 최소 10배수 이상의 문서가 필요하였다. 또한 목표한 5만 건 보다 더 많은 문서를 선별한 이유는 감정 분석 결과 일치도가 확보된 문서만 납품 대상 자료가 되기 때문이다. 따라서, 감정 분석을 위한 유효 문서량은 실제 선별 및 정제에 따른 삭제 비율과 일치도에 따른 삭제 비율을 모두 고려해 선정해야 한다.

<표 2> 트위터 선별 자료(예시)

doc_id	본문
ESRW2200000002.16397	아이유 응원봉 샀는데 오늘 오나봄 ㅎㅎㅎ
ESRW2200000002.16398	작가님 ㅠㅠㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ아 내 우울한 일상의 단비...2주 뒤에 프박으로 대리님 뽑으러 가야지 히히
ESRW2200000002.16417	하 오리사카 유타 노래 너무… 너무 좋음
ESRW2200000002.16444	<연애의 온도>랑 <특종: 량첸살인기> 당시에 인상적이었던 한국영화였는데 시리즈는 어떨지 궁금하다.
ESRW2200000002.16510	데뷔 전 사진도 그렇고 레이는 아이돌 화장 안 하면 청순미녀
ESRW2200000002.16935	영화관 오랜만에 왔어,, 연우진 배우님 이따 영접할 생각에 신남

2. 일상 대화 자료 선별

일상 대화는 국립국어원 배포 말뭉치인 ‘2020 일상 대화 말뭉치’에서 선별하였다. ‘2020 일상 대화 말뭉치 구축’ 사업의 일상 대화 말뭉치 총 2,232개 파일에서 감정 담화문 1만 건 확보를 목표로 하였다.

일상 대화 말뭉치는 2인의 발화자가 주고받는 일상적 대화를 녹음하여 전사한 자료이다. 본 사업에서는 일상 대화 말뭉치에서 감정 분석을 위한 감정 담화문을 선별하여 분석 대상 문서로 구축하였다. 감정 담화문이란 감정 분석 대상 문장과 그것의 선행 또는 후행 문장을 맥락의 유기성을 고려해 구성한 5어절 이상의 담화를 의미한다.

2-1. 담화 단위 선정 기준

감정 분석을 위한 감정 담화문 선별의 세부 지침은 다음과 같다.

1) 감정 담화문을 선별하기 위해서는 감정 문장을 찾는 작업이 가장 먼저 이루어져야 한다. 감정 문장이란 발화문에서 ‘발화자의 감정’이 드러나는 발화를 포함하는 문장이다. 감정 문장 선별 담당자는 일상 대화 말뭉치의 발화문을 직접 읽으며 감정 문장을 선별한다. 감정 문장은 색상으로 표시한 문장이다.

<표 3> 일상 대화 발화문 감정 문장 탐색(예시)

utterance_id	발화문
SDRW2000001227.1.1.61	더 도와주는 그런 역할을 하는 거 같아요.
SDRW2000001227.1.1.62	혹시 name1 씨는
SDRW2000001227.1.1.63	영화 장르들 중에서
SDRW2000001227.1.1.64	가장 좋아하는 장르가 있을까요?
SDRW2000001227.1.1.65	**저는 상업 영화를 굉장히 좋아해요.**
SDRW2000001227.1.1.66	누구나 재미있게 볼 수 있는 그런 영화들이요.
SDRW2000001227.1.1.67	이제 너무 심오하거나 그러면 저는 심오한

2) 감정 문장을 찾은 후 담화 맥락을 파악할 수 있는 범위에서 선행 또는 후행 문장을 포함하여 감정 담화문을 선별한다. 이때 감정 담화문 내에서 발화자가 바뀌는 말차례 바꿈의 횟수나 전체 문장 수는 제한하지 않는다.

<표 4> 감정 담화문 발화 선정(예시1)

utterance_id	발화문	선정
SDRW2000001227.1.1.61	더 도와주는 그런 역할을 하는 거 같아요.	
SDRW2000001227.1.1.62	혹시 name1 씨는	○
SDRW2000001227.1.1.63	영화 장르들 중에서	○
SDRW2000001227.1.1.64	가장 좋아하는 장르가 있을까요?	○
SDRW2000001227.1.1.65	**저는 상업 영화를 굉장히 좋아해요.**	○
SDRW2000001227.1.1.66	누구나 재미있게 볼 수 있는 그런 영화들이요.	
SDRW2000001227.1.1.67	이제 너무 심오하거나 그러면 저는 심오한	

3) 감정 담화문은 하나 이상의 감정 문장을 포함할 수 있다. 또한, 하나의 문장만으로도 담화의 맥락을 이해할 수 있다면 단일 문장 담화 구성도 가능하다.

<표 5> 감정 담화문 발화 선정(예시2)

utterance_id	발화문	감정 문장	발화문 구분
SDRW2000000514.1.1.50	그래갖고 사실 내가	1-1	1
SDRW2000000514.1.1.51	비형이 여기 있나 모르겠는지마는		
SDRW2000000514.1.1.52	비형을 별로 안 좋아해.		
SDRW2000000514.1.1.53	조금 무서워하지.	1-2	
SDRW2000001236.1.1.311	그~ 저보다 나이도 많고 계급도 높았던 분인데 그분이 저를 계속 안내를 하시는데	2	2
SDRW2000001236.1.1.312	단 한 번도 저보다 늦게 나온 적이 없었고 숙소에도		
SDRW2000001236.1.1.313	그다음에 버스를 타고 갔을 때 저랑 같이 앉아 있으면 절대 제 무릎에 그분의 무릎이 닿은 적이 한 번도 없었고		
SDRW2000001236.1.1.314	물론 일본 문화가 그렇다고는 하는데 저는 그게 되게 충격이었어요.		

4) 연속하는 발화 중간의 특정 발화를 삭제하여 분석 대상 담화를 선별할 수 없다.

<표 6> 감정 담화문 발화 선정(예시3)

utterance_id	발화문	선정
SDRW2000001006.1.1.20	이~ 그~ 최초 계약 후 5년이 지나면	
SDRW2000001006.1.1.21	보호 대상에서	
SDRW2000001006.1.1.22	제외된다	
SDRW2000001006.1.1.23	이런 게 아직 국회를 통과하지도 못하고	
SDRW2000001006.1.1.24	계속 계류 중이라고 하네요.	
SDRW2000001006.1.1.25	그~	○
SDRW2000001006.1.1.26	그리고 이~ 그~ 상가건물 임대차보호법뿐만 아니라	○
SDRW2000001006.1.1.27	다른	○
SDRW2000001006.1.1.28	그~ 법들 사설 업체의	○
SDRW2000001006.1.1.29	그~ 강제 집행을 제한하는 경비업법 개정안도	○
SDRW2000001006.1.1.30	아직 국회를 통화하지 못하고 있다고 하니	○
SDRW2000001006.1.1.31	**정말 답답할 뿐입니다.**	○
SDRW2000001006.1.1.32	그러니까 이분이 얼마나 답답했으면	
SDRW2000001006.1.1.33	이렇게 갑자기	

2-2. 담화 단위 선별

감정 담화문 선별 세부 지침의 내용을 기준으로 추출한 일상 대화 말뭉치 감정 담화 선별 문서 예시는 아래와 같다.

<표 7> 일상 대화 말뭉치 선별 감정 담화문(예시)

utterance_id	분석 대상 담화 선별 예시	번호
SDRW2000001221.1.1.310	내가 뭐가 필요한지도 잘 모르겠어.	1
SDRW2000001221.1.1.311	**살다가 보면은 이렇게 가끔씩 아~ 나 이거 사고 싶다 할 때는 있는데**	

utterance_id	분석 대상 담화 선별 예시	번호
SDRW2000001221.1.1.312	막상 생일 때 되면 아~ 잘 모르겠어	
SDRW2000001221.1.1.313	이렇게 돼요.	
SDRW2000001221.1.1.314	근데 그것도 참 답답한 거 같아.	
SDRW2000001220.1.1.132	해리포터는 확실히	
SDRW2000001220.1.1.133	보면 재밌긴 하더라고요.	2
SDRW2000001220.1.1.134	어.너무 이게 숙 빠지더라고.	
SDRW2000001239.1.1.106	우리 대한민국 국민이라고 하면은	
SDRW2000001239.1.1.107	국방의 의무가 있어서	
SDRW2000001239.1.1.108	아무튼 가야 돼잖아요.	
SDRW2000001239.1.1.109	그리고 어떤 부모들이고	
SDRW2000001239.1.1.110	이~ 티브이 보면은 그렇게 면제 핼라고	3
SDRW2000001239.1.1.111	백 있고 뭣 있는 사람들은	
SDRW2000001239.1.1.112	다 면제받아 갖고 어디서	
SDRW2000001239.1.1.113	막 병원에서 어디서 그 진단서를 끊어 갖고 내가지고 면제받아 갖고	
SDRW2000001239.1.1.114	나와서 막 그런 거 보면은 한심스러워요.	
SDRW2000000577.1.1.105	제가 40대가 되면서	
SDRW2000000577.1.1.106	왠지	
SDRW2000000577.1.1.107	나뭇잎을 보면	
SDRW2000000577.1.1.108	떨어지는 제 모습을 보는 거 같아서	
SDRW2000000577.1.1.109	속상하기도 하고	
SDRW2000000577.1.1.110	많이 외롭고	
SDRW2000000577.1.1.111	그랬지만	4
SDRW2000000577.1.1.112	또 한편으로	
SDRW2000000577.1.1.113	열매가 열리는	
SDRW2000000577.1.1.114	계절을 보면	
SDRW2000000577.1.1.115	또	
SDRW2000000577.1.1.116	떨어지는 것만이 아니라 여물고 있구나	
SDRW2000000577.1.1.117	이런 생각도 들면서	
SDRW2000000577.1.1.118	가을이 꼭 슬프지만은 않은 거 같아요.	

제 3 장

말뭉치 감정·공격성 분석 방법론 및 지침

1. 말뭉치 감정 분석 개요

본 사업에서 감정 분석 대상은 일상 대화 담화문과 트위터 자료이다. 일상 대화 담화문은 2020년 국립국어원에서 구축한 일상 대화 말뭉치 자료를 기반으로 하는 문서이다. 일상 대화 말뭉치 자료는 2인 발화자의 15분 분량의 일상 대화를 녹음하여 억양구 단위로 전사한 자료이다. 본 사업에서는 해당 자료에서 감정이 드러난 5어절 이상의 담화를 선별하여 감정 분석 대상 문서를 구성하였다. 감정 분석 작업에서는 일상 대화 담화문 총 1만 건의 감정 표지 부착을 목표로 한다.

누리소통망(SNS)은 다양한 사람들이 제한 없이 자신의 의견과 생각을 자유로이 주고받는 매체임은 물론 개인의 감정까지도 제재 없이 표출할 수 있는 창구의 역할을 한다. 이 때문에 화자(작성자)가 생성한 텍스트에서 감정이 담긴 표현을 포착하여 감정을 분석하기에 매우 용이한 자료라 할 수 있다. 본 사업의 목표는 문화콘텐츠 관련 트위터 게시물 5만 건에 대한 감정 및 공격성을 분석하여 분석 표지를 부착하는 것에 있다.

본 사업에서 정의하는 감정 분석(Emotional analysis)이란 텍스트가 내포하는 정서를 파악하여 다양한 감정 유형으로 분류하는 것이다. 이는 텍스트의 주관성을 세분화한 감정 유형으로 분류함으로써 긍·부정 극성만을 분석한 감성 분석보다 정교화된 분석이라 할 수 있다. 예컨대 '원하던 대학에 합격해서 날아갈 것 같아. 즐거운 학교 생활이 될 거야'라는 텍스트에 기쁨, 기대의 감정 유형 표지를 부착함으로써 단순히 '긍정' 극성으로 분류하는 것에서 한 단계 더 나아가 작성자의 감정을 세밀하게 나타낸다.

텍스트에서 감정은 다양한 언어 표현을 통해 나타난다. 우리는 감정을 표현할 때 '기쁘다' 또는 '슬프다'와 같은 명시적 감정 어휘를 사용하여 자신의 감정을 직접적으로 나타낼 수 있다. 하지만 '콧방귀를 뀌다' 또는 '펄쩍 뛰다' 등의 관용 표현을 사용하거나 '문을 쾅 닫다' 등의 행동 묘사를 통해서도 감정을 간접적으로 전달하기도 한다. 또한 '그는 프로젝트를 훌륭히 해냈다', '걔는 항상 지저분하게 먹어' 등의 문장처럼 주관적 평가를 함으로써 사물이나 인물에 대한 감정을 드러내기도 한다.

그러므로 감정 분석 대상 텍스트는 단순히 감정 어휘를 포함하는 글뿐 아니라 '맥락 속에서 화자(작성자)의 감정을 읽어낼 수 있는 텍스트'가 되어야 한다. 따라서 감정 분석 작업에서는 작업자가 직접 텍스트를 읽고 표면적으로 드러나지 않는 화자의 감정까지 정

밀하게 분석하는 것이 중요하다. 다양한 감정 표현 수단을 사용한 트위터 자료 예시는 다음과 같다.

<표 8> 감정 표현 유형 예시(트위터 자료)

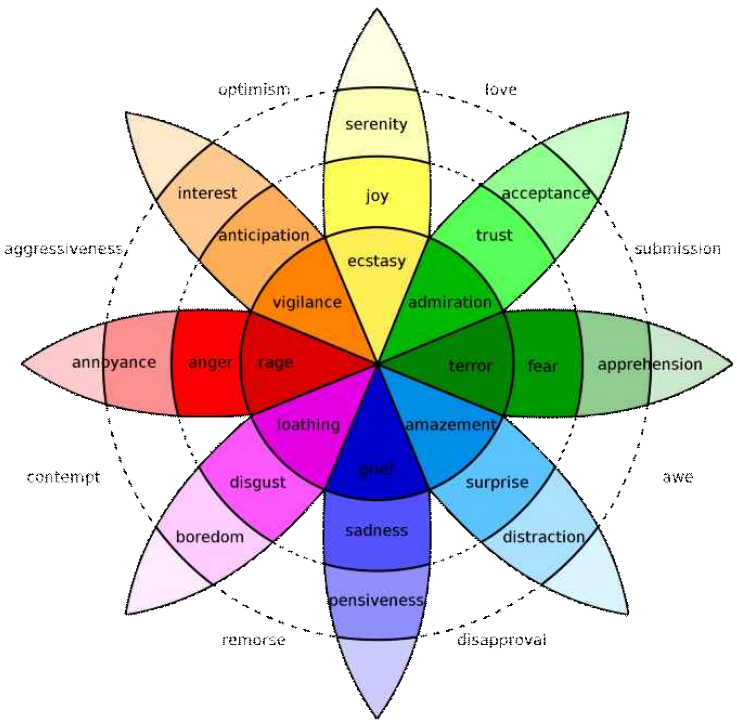
유형	트위터 자료 예시
명시적 감정 어휘 사용	와 나 진짜 드라마 보는데 **<u>넘 빡쳐서 몬살겟어</u>** 박피디 나쁜놈아
관용 표현	오늘 두 명이나 고민하고 있어서 식겁하고 말 걸었다가 선물받아서 맘 놓구 고민하는 차둘이한테 말 걸었는데 이사고민이라 **<u>심장 철렁했네</u>** (>⌓<) 주대 맘을 들었다놨다 들었다놨다 해이...
행동 묘사	혹시 이장면을 말씀하시는 건가요 **<u>저 이장면보고 폰 던졌어요</u>** 다자이 이 (왈왈) 진짜 한대만 치자 제발 이러면서 봤어
주관적 평가	문나이트 다리쪽엔 고대 상형문자가 있는데, '달을 수호하는 전사'라는 말을 만들어서 고대 이집트 문나이트가 되었다고 함. 여기서 오스카 아이작이 수트 입는 장면 나옴!□🌙 **<u>슈트도 클로즈업 한 거 보니까 캐릭터랑 배경 생각해서 어우러지게 잘 만든 것 같아서 감탄 중. 의상팀 존나 배운 변태같애</u>**

2. 말뭉치 감정·공격성 분석 지침

2-1. 말뭉치 감정 분석 표지 정의

인간의 감정을 분류하기 위한 연구는 오랜 시간 동안 다양한 연구자에 의해 이루어져 왔다. 그러나 학계 및 산업계의 보편적 합의를 이끌어 낸 감정 체계는 아직까지 확립되지 않은 상태이다. 이에 본 사업에서는 과업의 효율성을 위하여 분석 대상 텍스트 유형 및 연구 목적에 중점을 두고 트위터 게시물 텍스트의 감정을 분석한 연구 SemEval 2018 Task 1: Affect in Tweets(이하 SemEval 2018)을 참고하여 감정 분석 표지를 구성하였다. SemEval 2018에서는 정서심리학자 로버트 플루칙(Robert Plutchik)이 주장한 기본 감정 및 복합 감정을 선별하여 감정 분석용 표지를 구성하였다. 로버트 플루칙(Robert Plutchik)이 주장한 감정 체계는 아래 그림과 같다.

<그림 5> 플루칙의 감정 수레바퀴(Plutchik's wheel)



※ Plutchik's wheel : 로버트 플루칙(Robert Plutchik)이 제안한 모델 (Plutchik, 1990)
8개의 핵심 감정의 조합으로 발생하는 감정 상태를 표현

로버트 플루칙(Robert Plutchik)은 위 그림과 같이 인간의 기본 감정을 8가지로 분류하였으며 감정 강도의 차이를 두어 각 기본 감정에 대한 스펙트럼을 형성하였고 다수의 감정이 함께 나타나는 것을 복합 감정으로 설정하여 감정 체계를 구성하였다. 본 사업에서는 SemEval 2018과 같이 로버트 플루칙(Robert Plutchik)의 감정 체계를 차용하되 한국어 특성 및 국내 연구 동향을 고려하여 8가지 기본 감정(기쁨, 기대, 신뢰, 놀람, 혐오, 공포, 분노, 슬픔)만을 감정 표지로 활용하고자 한다. 이때 로버트 플루칙(Robert Plutchik)의 감정 수레바퀴에서 상정하는 각 스펙트럼 내 감정은 해당 기본 감정 표지에 포함하도록 한다. 만일 8가지 기본 감정 스펙트럼에서 벗어나 어떠한 표지로도 정의될 수 없을 경우 '기타' 표지로 태깅하도록 한다. 각 감정 분석 표지에 대한 세부 설명은 아래와 같다.

<표 9> 감정 분석 표지

감정 분석 표지	동일 스펙트럼 감정	설 명
기쁨(Joy)	평온, 황홀	어떤 만족감에 의해 느끼는 즐겁고 흥겨운 감정.
기대(Anticipation)	경계, 관심	앞으로 있을 일이나 상황을 미리 짐작함. 또는 그런 내용.
신뢰(Trust)	수용, 감탄/존경	굳게 믿고 의지함.
놀람(Surprise)	놀라움, 부주의/방심	어떤 일이 뜻밖이거나 훌륭하거나 무서워서 신기해하거나 흥분하여 가슴이 뛰는 느낌.
혐오(Disgust)	지루함, 증오	싫어하고 미워함.
공포(Fear)	불안, 두려움	두렵고 무서움.
분노(Anger)	짜증, 격노	분개하여 크게 화를 냄.
슬픔(Sadness)	수심, 비탄	마음이 아프거나 괴로운 느낌.
기타	-	8가지 표지 외 기타 감정. (선후의 감정 정보 없는 궁금함, 의아함. 묘함 등)

인간의 감정은 그 구조와 관계가 복잡다단하여 텍스트의 감정을 단일 표지로 정의할 수 없는 경우가 존재할 수 있다. 예를 들어 복합 감정이 나타나는 경우, 하나의 문서에 다수의 감정이 나타나는 경우, 동일 텍스트 내에서 감정이 변화하여 연속적으로 나타나는 경우가 그러하다. 이러한 문제를 해결하기 위하여 본 연구에서는 복합 감정, 복수 감정, 연속 감정을 분석하기 용이한 감정 복수 태깅이 가능하도록 설정하였다.

트위터의 경우 감정 분석 과정에서 감정이 없다고 판단되는 경우 별도 '비고'란에 기입하도록 하였다. 그러나 감정이 없는 문서의 비중이 지나치게 높은 비중을 차지하지 않도록 하기 위해 감정 없음 비중이 트위터 자료 목표의 40%(2만 건)를 넘지 않도록 하였다. 또한 분석 자료로 활용도를 높이기 위해 기타를 제외한 8개 감정 표지별 최소 1,500건 이상의 문서가 확보될 수 있도록 표지별 목표량을 설정하였다.

2-2. 공격성 범주 및 공격 대상 유형 표지 정의

트위터 자료에서 공격성이 드러날 경우 '공격성 범주'를 파악하여 태깅한다. 공격성 범주는 공격성을 촉발하는 원초적인 감정을 공격성 범주로 파악하였다. 감정 표지 중 공격성을 촉발한다고 볼 수 있는 감정은 혐오, 분노 표지로, 공격성이 감지되는 경우 혐오, 분노 감정 중 공격성이 촉발된 감정을 태깅하되 두 가지 감정이 모두 나타나는 경우 두 감정을 모두 태깅하도록 한다.

<표 10> 공격성 범주 표지

공격성 범주 표지	설명	트위터 자료 예시
혐오	타인을 싫어하고 미워하는 감정에서 기인한 공격성	난 불륜로맨스 개극혐인데 사람들 왤케 좋아함 내가 모르는 뭔...공감대가 잇음?;; 영화 급 생각나서 검색햇다가 눈알 튀어나옴
분노	분개하거나 몹시 노여워하는 감정에서 기인한 공격성	가오갤2때 영화관병크가 최강이었는데 영화보기전에 스포하는놈도 있고 디씨캐릭터를 마블캐라면서 이야기하는애도있고 뒷자리 발로치고 커플들 떠들고ㅋㅋㅋㅋㅋ진짜 개빡쳤음 내가 비싼 아이맥스결제하고 너네 병크봐야하냐ㅗㅗ

공격성이 드러날 경우 공격성이 향하는 대상(target)을 고려하여 '공격 대상 범주'를 태깅한다. 공격성이 특정 인물을 향하고 있을 경우 '개인' 표지를, 특정 집단을 향할 때는 '집단' 표지를 부착한다. 개인이나 집단에 속하지 않는 상황, 사건, 구조 등에 대한 공격성을 드러내는 게시글의 경우 '기타' 표지로 태깅하여 공격 대상을 분석한다. 공격 대상

의 실질적 예시는 아래 표와 같다.

<표 11> 공격 대상 범주 표지

공격 대상 표지	설명	공격 대상 예시
개인	익명 또는 실명의 특정 인물	양세형, 기리보이, kim1988 등
집단	특정 단체 및 개인이 아닌 다수의 집단	씨제이, 엑소팬덤, 여고생 등
기타	구조, 상황, 사건, 콘텐츠 등 개인이나 집단에 속하지 않는 대상(비인간)	도서정가제, 코로나, 또 오해영 등

공격성 범주 표지와 공격 대상 범주 표지를 모두 부착한 트위터 공격성 분석 예시는 다음과 같다.

<표 12> 트위터 자료 공격성 분석 예시

번호	트위터 본문	공격 대상 (target)	공격 대상 범주
1	아수라도 평범한 알탕영화인데 대사가 살린 영화자너 대사의 힘은 엄청나 못생긴 **대사** 죽어	대사	기타
2	**철권** 진짜 개쓰레기 게임 **하라다** 미친새끼야 밸런스패치 해라 진짜	철권, 하라다	기타, 개인
3	오늘을 기점으로 이번 시즌 끝날 때 까지 야구 접습니다. **쓰레기팀**에 더이상 감정소모 하고싶지 않네요	쓰레기팀	집단

또한 감정 및 공격성의 대상이 특정 인물 또는 집단일 경우 개인정보의 노출 등을 막기 위해 대상(target) 선정 시 대상(target)의 비식별화 대상 여부를 판단해 비식별화 표지를 부착한다.

2-3. 감정의 대상(target) 분석

분석 대상 텍스트 내에서 화자(작성자)가 느끼는 감정의 대상이 되는 표현을 선별하여 이를 대상(target)으로 정의하여 분석한다. 예컨대 '나는 버섯이 물컹거려서 싫어.'라는 문장이 있다면 '혐오' 감정의 대상 '버섯'이 대상(target)이 된다. 분석 대상 텍스트에서 대상(target)을 선정하는 데는 아래 순서로 규칙을 적용한다.

① 대상(target)의 범위는 지시하거나 의미하는 바가 무엇인지 명확하게 파악할 수 있는 구체 명사 또는 추상 명사로 선정함. 이때 구체 명사나 추상 명사를 수식하는 성분은 대상(target) 선정에서 제외함.

예 나온지 엄청 된 **영화**인데, 어제 잠이안와서 그냥 우연히 다시 봄. 여전히 다시 봐도 감동이고 진짜 대박 슬픔.. → target: 영화

② 문서 내에 대상을 가리키는 명확한 구체 명사 혹은 추상 명사가 없고 '수식 성분 + 의존 명사(-것, -거, -게, -분 등)' 구조의 구문이 감정의 대상이 되는 경우 해당 구문을 대상(target)으로 선정함.

예 그~ 여행지에서 가장 인상 깊었던 건 **사람들이 진짜 친절한 거**였어요.
→ target: 사람들이 진짜 친절한 거

③ 문서 내에 대상을 가리키는 명확한 구체 명사 혹은 추상 명사가 없고 대명사(이거, 저거, 그거 등)가 감정의 대상이 될 경우 해당 표현을 대상(target)으로 선정함.

예 저는 아무래도 **이게** 조금 안타깝긴 하더라고요.
→ target: 이게

④ 동일한 것을 가리키는 추상 명사, 구체 명사, '수식 성분+ 의존 명사' 구문, 대명사 등이 동일 문서 내에 존재할 경우 추상 명사·구체 명사 > 수식 성분 + 의존 명사 > 대명사 순으로 우선 순위를 적용하여 대상(target)을 선정함.

예 영화 마지막에 나오는 **장면** 미쳤더라. 주인공 우는 거 보고 나도 울었음.
→ target: 장면

예 **걔가 지난주에 작업한 거** 봤어? 미친 거 아니야? 그건 좀 아니지.
→ target: 걔가 지난주에 작업한 거

⑤ 같은 대상을 가리키는 동일 수준 명사(또는 완전히 동일한 명사)가 복수로 등장하는 경우 문서 내에서 가장 먼저 등장하는 것을 대상(target)으로 선정함.

예 요즘 내가 보는 **소설**이 있거든. 근데 그 소설 진짜 읽다가 멈출 수가 없더라.

→ target: 소설

⑥ 여러 개의 어절로 구성된 고유 명사의 경우 전체 표현을 대상(target)으로 선정함. 단, 해당 표현이 하나가 아닌 여러 개의 발화(utterance)에 걸쳐 등장한다면 해당 고유 명사의 일부가 포함된 마지막 발화(utterance) 내의 표현만 대상(target)으로 선정함.

예 어제 **슬픔이여 이제**

안녕 듣고 한참 울다가

밤에 잠을 못 잤지.

→ target: 안녕

⑦ 감정에 대한 대상이 문서 내에 없는 경우 대상(target)은 'null'로 처리함.

예 너무 아름다우셔,,, 화면에 잡힐때마다 와,,,, 소리 절로나옴

→ target: null

예 영화속 퇴마당하는 악귀 마냥 울부짖고있음…. 너모 감동적이여

→ target: null

2-4. 감정의 대상(target) 비식별화

본 사업에서는 감정의 대상(target)이 특정 인물 또는 집단일 경우 개인정보의 노출 등을 막기 위해 대상의 비식별화 처리를 진행한다. 대상(target)의 비식별화 처리 여부를 판단하여 비식별화 처리 대상일 경우 해당 대상(target)에 맞는 비식별화 표지를 태깅한다. 태깅한 비식별화 표지는 데이터 출력 시 대상(target) 위치에 마크업되어 나타나며 이를 통해 비식별화 정보의 노출을 막을 수 있다. 비식별화 유형과 그에 따른 비식별화 표지는 다음과 같다.

<표 13> 비식별화 유형 및 표지

비식별화 처리 유형	비식별화 표지	설명
이름	&name&	개인의 실명 (정치인, 연예인 등 공인·유명인 제외, 실존 인물이 아닌 캐릭터·극중 인물 제외)
온라인 계정 (아이디)	&account&	트위터 등 특정 사이트의 온라인 계정
고유 식별 번호 (주민등록번호)	&social-security-num&	개인의 주민등록번호
전화 번호	&tel-num&	휴대폰 번호, 사업장 번호 등
카드 번호	&card-num&	신용카드 번호 등
기타 번호	&num&	비밀 번호 등 기타 비식별화 대상 번호
주소	&address&	동 이하의 상세 주소
출신 및 소속	&affiliation&	개인의 출신 및 소속
기타 비식별화 필요 항목	&others&	위 항목 외 기타 비식별화 대상

제 4 장

말뭉치
감정 분석 및
말뭉치 구축

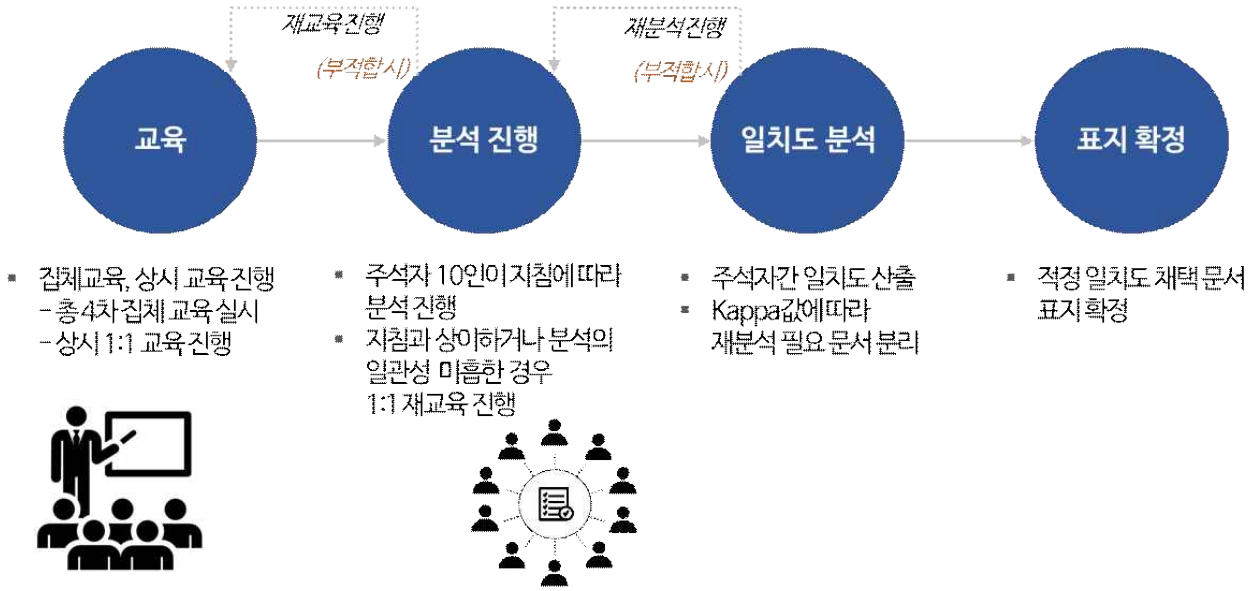
1. 말뭉치 감정·공격성 분석

'말뭉치 감정 분석 지침'을 통해 기준을 명확하게 수립하였다 하더라도 텍스트에서 나타나는 다양한 언어 표현에 대해 감정 분석이 진행되므로 주석자에 따라 감정을 인식하는 유형이 달라지는 문제를 완전히 배제할 수 없다.

이러한 점을 보완하기 위해 선행 연구에서는 감정 분석의 객관성 확보를 위해 다수의 주석자가 감정 분석을 진행하고 응답 수에 따라 표지를 채택하는 방법을 사용하였는데, SemEval 2018 연구에서 7명의 주석자가 분석을 진행하고 2명 이상이 응답한 표지에 대해 채택한 바 있다.

본 사업에서는 선발된 10명의 주석자가 동일한 자료 1건의 문서에 대한 분석을 진행하고, 통계적으로 산출된 일치도를 근거로 '적합 문서'와 '문서별 표지'를 채택하였다. 말뭉치 감정·공격성 분석 진행 절차는 다음과 같다.

<그림 6> 말뭉치 감정·공격성 분석 진행 절차



말뭉치 감정 분석은 3차 이상의 교육을 마친 10명의 주석자가 분석 지침에 따라 분석

및 주석 작업을 진행하였다. 주석자는 4년제 대학 이상 졸업자로 감성 분석, 인공지능 언어 자원 구축 등 유관 실무 경력 2년 이상자로 선발하였다. 또한 10명 중 60%는 언어학 전공자로 구성하여 분석 결과의 품질을 최대한 확보하고자 하였다.

주석자로 선발된 10명을 대상으로 총 4회 정기 집체 교육을 실시하였다. 이 교육은 감정 분석 표지에 대한 이해와 분석 대상 문서인 트위터, 일상 대화 문서에 대한 종합적인 이해도를 높이기 위한 과정으로, 참여자는 최소 2회 이상의 정기 교육을 받도록 하였다. 또한 지침과 다른 감정 분석을 진행하거나 분석의 일관성이 떨어지는 경우 상시로 1:1 교육을 진행하여 분석의 객관성을 확보하고자 하였다.

분석이 완료되면 10명이 분석한 결과에 대한 일치도를 산출하여 일치도 산출값에 따라 적합 문서와 부적합 문서를 분리하였다. 적합 문서는 일치도 산출 기준에 따라 문서별 표지를 확정하였고, 10명이 다시 분석해야 하는 '재분석 진행이 필요한 문서'는 분석 대상으로 삼지 않았다.

2. 일치도 분석 및 표지 확정

2-1. 일치도 분석 개요

10명의 주석자 분석 결과를 취합해야 하는 감정 분석의 특성상 분석의 객관성 확보 및 말뭉치 품질을 확보하기 위해 분석 일치도 산출은 필요하다. 이에 따라 주석자 간 분석 일치도 평가를 위한 통계적 접근과 일치도 관리의 기준 마련이 필요하다.

주석자 간 일치도 산출은 평가자 일치도 분석을 위한 통계적 분석 기법인 카파통계량 중 3명 이상의 복수 평가자 분석에 적합한 일반화 카파통계량을 기반으로 한다. 일치도란 한 표본을 여러 번 반복 측정한 결과가 서로 어느 정도 일치하는가를 알아보는 신뢰도 평가의 척도로서 한 명 혹은 여러 명의 평가자가 한 표본을 평가할 때 일치하는 정도이다. 일치도 분석을 위한 일반화 카파통계량은 r명의 평가자가 n명(또는 건)의 평가 대상을 q개의 범주로 평가한다는 가정을 기준으로 한다. 평가 결과의 범주는 순서를 고려하지 않는 이분형 혹은 명목형이며 평가자는 서로 독립임을 가정하는 Fleiss의 방법을 기반으로 접근하였다.

<그림 7> 카파통계량(kappa statistic) 산출식

$$\kappa = \frac{P_a - P_e}{1 - P_e}$$

P_a : 관찰된 일치비율
P_e : 우연에 의해 기대되는 일치비율

그러나 일반화 카파통계량은 r명의 평가자가 q개 범주 중 한 가지로 평가한다는 가정을 기준으로 한다. 본 사업에서는 감정 분석 9개 표지에 대한 다중 레이블 값을 인정하므로 복수의 감정 표지가 부착되는 데이터 구조에 대한 산출 방안이 필요하다.

2-2. 주석자 간 일치도 산출 방안

본 사업에서 일치도 산출의 목적은 평가 전체 결과에 대한 것이 아니며, 문서 단위 감정 분석 결과에 대한 주석자 간의 일치도 분석 및 관리를 목표로 한다. 즉, 전체 결과에 대한 일치도가 아닌 문서 1개 단위를 기준으로 일치도 분석값을 도출하도록 하였다.

따라서, 문서 1건을 전체 평가 대상 데이터로 간주하고 9개의 감정 표지를 평가 대상(건)으로 산정하여 k값을 산출한다. 이 경우 9개 감정 표지를 평가 대상자로 간주해 표지별 일치도 값이 산출된다(표지별 Pa 값 도출). 해당 방안을 사용할 경우 다중 레이블값 그대로 일치도 도출이 가능하며, 일치도 산출 데이터와 주석 결과 데이터가 동일하다는 장점이 있다. 문서 단위 일치도 산출값은 Fleiss 카파통계량 산출식을 이용하며, 문서당 각각의 k값이 도출된다. Fleiss 카파통계량 산출식은 아래와 같다.

<그림 8> Fleiss 카파통계량 산출식

$$\kappa = \frac{\frac{1}{nr(r-1)}\left(\sum_{i=1}^{n}\sum_{j=1}^{q} r_{ij}^2 - rn\right) - \sum_{j=1}^{q} p_j^2}{1 - \sum_{j=1}^{q} p_j^2}$$

r = 평가자 수(주석자, 10명)

n = 평가 대상 감정 수(감정 수, 9개)

q = 범주 수(감정별 이분형 값, 2개)

산출식에 따라 실제 10명의 주석자가 분석한 결과를 대상으로 산출한 일치도 산출 결과는 다음과 같다.

<표 14> Fleiss 카파통계량 기준 일치도 산출 결과 예시 (다중 레이블값 기준)

문서	기쁨	기대	신뢰	놀람	혐오	공포	분노	슬픔	기타	k 값
post 1	3	7	0	0	0	0	0	0	0	0.48
post 2	2	0	0	0	8	0	4	0	0	0.47
post 3	0	0	0	0	2	0	8	2	0	0.49
post 4	0	0	0	2	0	0	8	0	0	0.60

문서	기쁨	기대	신뢰	놀람	혐오	공포	분노	슬픔	기타	k 값
post 5	0	0	0	0	0	2	0	8	0	0.60
post 6	0	1	0	0	8	0	2	0	0	0.53
post 7	8	0	4	0	0	0	2	0	0	0.47
post 8	2	0	0	0	0	8	0	0	0	0.60
post 9	0	0	0	0	10	0	0	0	0	1.00
post 10	0	0	0	0	0	0	0	10	0	1.00

2-3. 감정 표지 채택 방안

Landis and Koch(1977)의 일치도 해석 기준에 따라 적당한 일치도(Moderate) 이상에 해당하는 k값 0.4 초과 문서를 채택한다. 감정 표지는 해당 채택 문서 내에서 5명 이상(관찰된 일치 비율 0.4 초과)의 분석 결과가 일치한 표지만을 최종 채택하도록 한다.

<표 15> Landis and Koch(1977)의 Strength of agreement

Kappa value	Strengthof agreement
0.81~1.00	**Almost Perfect(완벽한 일치도)**
0.61~0.80	**Substantial (상당한 일치도)**
0.41~0.60	**Moderate (적당한 일치도)**
0.21~0.40	Fair (어느 정도의 일치도)
0.00 ~ 0.20	Slight (약간의 일치도)
< 0.00	Poor (거의 없는 일치도)

일치도가 0.4를 초과한 문서에서 최종 표지 채택 결과 예시는 다음과 같다.

<표 16> 일치도 분석 결과에 따른 표지 확정 예시

문서	기쁨	기대	신뢰	놀람	혐오	공포	분노	슬픔	기타	확정 표지
post 1	3	**7**	0	0	0	0	0	0	0	기대
post 2	2	0	0	0	**8**	0	4	0	0	혐오
post 3	0	0	0	0	2	0	**8**	2	0	분노
post 4	0	0	0	2	0	0	**8**	0	0	분노
post 5	0	0	0	0	0	2	0	**8**	0	슬픔
post 6	0	1	0	0	**8**	0	2	0	0	혐오
post 7	**8**	0	4	0	0	0	2	0	0	기쁨
post 8	2	0	0	0	0	**8**	0	0	0	공포
post 9	0	0	0	0	**10**	0	0	0	0	혐오
post 10	0	0	0	0	0	0	0	**10**	0	슬픔

3. 말뭉치 감정·공격성 분석 결과

일상 대화 1만 건, 트위터 5만 건에 대한 말뭉치 감정·공격성 분석 진행 결과, 감정 표지별 비중은 다음과 같다.

<표 17> 감정 분석 결과 표지별 비중

감정 분석 표지	일상 대화		트위터		합계	
	분석 표지 수 (건)	비중	분석 표지 수 (건)	비중	분석 표지 수 (건)	비중
기쁨(Joy)	5,710	57.1%	25,256	50.5%	30,966	51.6%
기대(Anticipation)	1,893	18.9%	10,773	21.5%	12,666	21.1%
신뢰(Trust)	414	4.1%	2,646	5.3%	3,060	5.1%
놀람(Surprise)	322	3.2%	3,229	6.5%	3,551	5.9%
혐오(Disgust)	1,296	13.0%	1,810	3.6%	3,106	5.2%
공포(Fear)	521	5.2%	1,015	2.0%	1,536	2.6%
분노(Anger)	144	1.4%	1,979	4.0%	2,123	3.5%
슬픔(Sadness)	1,097	11.0%	3,032	6.1%	4,129	6.9%
감정 없음	-	-	6,801	13.6%	6,801	11.3%

감정 표지는 앞서 설명한 바와 같이 1개 문서 당 여러 개의 표지 부착이 가능하다. 따라서 문서 수에 대한 비중은 전체 문서 내 분석된 표지 수 비중이 된다. 일상 대화와 트위터에 공통적으로 가장 많이 나타난 표지는 '기쁨'이며, 가장 적게 나타난 표지는 '공포'로 나타났다. 또한, 트위터 전체 분석 문서 50,000건 중 '감정 없음'으로 분석된 문서는 6,801건으로 13.6%를 차지하였다.

트위터 대상으로 분석한 공격성의 경우, 전체 5만 건 트위터 문서 중 594건으로 1.2%에 그쳤다. 누리소통망 특성상 타인과의 소통을 목적으로 하므로 공격성을 드러내는 문서를 작성하는 비중이 높지 않았다.

<표 18> 트위터 대상 공격성 분석 결과

공격성	혐오(Disgust)	분노(Anger)
594 건	279	463

4. 말뭉치 구축

트위터 자료의 경우 본 사업을 위해 참여자를 모집하여 구축한 자료이므로 일상 대화 자료와 달리 원시 말뭉치 구축이 필요하다. 따라서 트위터 자료는 국립국어원의 원시 말뭉치 구축 지침을 준용하고자 하였으며, 일상 대화 및 트위터 자료별 JSON 형태의 말뭉치 형식은 다음과 같다.

<표 19> 일상 대화 감정 분석 말뭉치 형식(JSON)

1수준	2수준	3수준	4수준	5수준	6수준	설명
id						말뭉치 파일 ID
metadata						파일의 메타 정보
	title					국립국어원 일상 대화 감정 분석 말뭉치 [파일ID]
	creator					생성자(국립국어원)
	distributor					배포자(국립국어원)
	year					말뭉치 구축 연도(2022)
	category					분류 (구어 > 사적 대화 > 일상 대화)
	annotation_level					분석 층위 (감정 분석)
	sampling					샘플링 방식 (부분 추출-특정 부분 추출)
document						대화 정보
	id					대화 ID
	metadata					문서의 메타 정보
		title				문서 ID
		author				개인 발화자
		publisher				개인 발화 녹음
		date				문서 작성 일시
		topic				주제
		speaker				화자 정보
			id			화자 ID
			age			연령
			occupation			직업
			sex			성별
			birthplace			출생지
			pricipal_residence			주 성장지
			current_residence			현 거주지
			education			학력
		setting				환경 정보
			relation			화자 간 관계
	ea_discourse					감정 담화문 정보
		id				감정 담화문 ID
		sentence				발화 정보
			id			감정 담화문 발화 ID
			form			감정 담화문 발화

1수준	2수준	3수준	4수준	5수준	6수준	설명
			word			어절 정보
				id		어절 ID
				form		어절
				begin		어절의 발화 내 시작 위치
				end		어절의 발화 내 끝 위치
			EA			감정 분석 정보
				id		분석 ID
				target		감정 분석 대상 정보
					id	감정 분석 대상이 나타난 발화 ID
					form	감정 분석 대상
					begin	감정 분석 대상의 발화 내 시작 위치
					end	감정 분석 대상의 발화 내 끝 위치
					emotion	감정 유형

<표 20> 트위터 감정 분석 말뭉치 형식(JSON)

1수준	2수준	3수준	4수준	5수준	6수준	설명
id						말뭉치 파일 ID
metadata						파일의 메타 정보
	title					국립국어원 트위터 감정 분석 말뭉치 [파일ID]
	creator					생성자(국립국어원)
	distributor					배포자(국립국어원)
	year					말뭉치 구축 연도(2022)
	category					분류 (웹 > 누리소통망)
	annotation_level					분석 층위 (감정 분석)
	sampling					본문 전체
document						문서 정보
	id					문서 ID
	metadata					문서의 메타 정보
		title				문서 제목
		author				작성자
		publisher				게시 플랫폼
		date				작성일시, 게시일시
		topic				주제
		crawl_date				크롤링 일시
		url				URL 주소
	sentence					문장 정보
		id				문장 ID
		form				문장
		word				어절 정보
			id			어절 ID
			form			어절
			begin			어절의 문장 내 시작 위치
			end			어절의 문장 내 끝 위치
		EA				감정 분석 정보
			id			분석 ID
			target			감정 분석 대상 정보

1수준	2수준	3수준	4수준	5수준	6수준	설명
				id		감정 분석 대상이 나타난 문장 ID
				form		감정 분석 대상
				begin		감정 분석 대상의 문장 내 시작 위치
				end		감정 분석 대상의 문장 내 끝 위치
				emotion		감정 유형
				aggression		공격성 분석 정보
					type	공격성 유형
					off_target	공격대상 유형

<표 21> 트위터 원시 말뭉치 형식(JSON)

1수준	2수준	3수준	설명
id			말뭉치 파일 ID
metadata			파일의 메타 정보
	title		국립국어원 트위터 원시 말뭉치 [파일ID]
	creator		생성자(국립국어원)
	distributor		배포자(국립국어원)
	year		말뭉치 구축 연도(2022)
	category		분류
	annotation_level		분석 층위 (원시)
	sampling		샘플링 방식(본문 전체)
document			문서 정보
	id		문서 ID
	metadata		문서의 메타 정보
		title	문서 제목
		author	작성자
		publisher	게시 플랫폼
		date	작성일시, 게시일시
		topic	주제(문화/예술)
		crawl_date	크롤링 일시
		url	URL 주소
	paragraph		문단
		id	문단 ID
		form	정제된 형태
		original_form	원문 표기된 그대로의 형태(개인 정보 비식별화 후)

감정 분석 말뭉치 구축 지침에 따라 출력한 말뭉치 형식(JSON) 납품 형태는 다음과 같다.

<그림 9> 일상 대화 감정 분석 말뭉치(JSON) 출력 예시


```
"ea_discourse": [
{
    "id": "SDEA2200000001.1",
    "sentence": [
        {
            "id": "SDRW2000000001.1.1.1",
            "form": "고양이를",
            "word": [
                {
                    "id": 1,
                    "form": "고양이를",
                    "begin": 0,
                    "end": 4
                }
            ]
        },
        {
            "id": "SDRW2000000001.1.1.2",
            "form": "좋아하는데요.",
            "word": [
                {
                    "id": 1,
                    "form": "좋아하는데요.",
                    "begin": 0,
                    "end": 7
                }
            ],
            "EA": [
                {
                    "id": 1,
                    "target": [
                        {
                            "id": "SDRW2000000001.1.1.1",
                            "form": " 고양이",
                            "begin": 0,
                            "end": 3
                        }
                    ],
                    "emotion": {
                        "joy": "True",
                        "anticipation": "False",
                        "trust": "False",
                        "surprise": "False",
                        "disgust": "False",
                        "fear": "False",
                        "anger": "False",
                        "sadness": "False",
                        "etc": "False"
                    }
                }
            ]
        },
                                ~ 생략 ~
```

<그림 10> 트위터 감정 분석 말뭉치(JSON) 출력 예시


```
"document": [
{
    "id": "ESRW2200000001.71976",
    "metadata": {
      "title": "N/A",
      "author": " bogusday",
      "publisher": "twitter",
      "date": "20201210",
      "topic": "문화/예술",
      "crawl_date": "2022062011:00:23",
      "url": "https://twitter.com/bogusday/status/1472454374559/"
    },
   "sentence": [
       {
        "id": "ESRW2200000002.1.1.1",
        "form": "이번 표절 너무 너무 빡친다",
        "word": [
          {
            "id": 1,
            "form": "이번",
            "begin": 0,
            "end": 2,
          },
                                                ~중략~
        ],
         " EA " : [
          {
            "id": 1,
            "target": {
              "id": "ESRW2200000002.1.1.1",
              "form": " 표절",
              "begin": 3,
              "end": 5,
            },
            "emotion": {
                "joy": "False",
                "anticipation": "False",
                "trust": "False",
                "surprise": "False",
                "disgust": "False",
                "fear": "False",
                "anger": "True",
                "sadness": "False",
                "etc": "False"
         },
            "aggression": {
              " type": " anger ",
              "off_target": "etc",
            }
          }
                                                ~ 생략 ~
```

제 5 장

결 론

1. 요약

1-1. 말뭉치 분석 자료 수집 및 선별

말뭉치 분석 자료로 트위터 자료 5만 건, 일상 대화 감정 담화문 1만 건을 수집 및 선별하였다. 트위터 자료는 저작권 이용 허락 계약이 완료된 참여자(총 336명)의 트위터 계정을 대상으로 자료를 수집하였다. 문서는 문화콘텐츠 자료, 5어절 이상인 문서만을 대상으로 선별하였으며, 자료의 수집은 사업 수행 기관 자체 개발 수집기인 'Buzz Crawler'를 이용하였다. 일상 대화는 국립국어원 '2020 일상 대화 말뭉치 구축' 배포 말뭉치에서 선별하였다. '2020 일상 대화 말뭉치 구축' 사업의 일상 대화 말뭉치 총 2,232개 파일에서 감정 담화문 1만 건을 선별하였다.

1-2. 말뭉치 감정·공격성 분석 방법론 및 지침

본 사업에서는 SemEval 2018과 같이 로버트 플루칙(Robert Plutchik)의 감정 체계를 차용하되, 한국어 특성 및 국내 연구 동향을 고려하여 8가지 기본 감정(기쁨, 기대, 신뢰, 놀람, 혐오, 공포, 분노, 슬픔)만을 감정 표지로 활용하였다. 이때 로버트 플루칙(Robert Plutchik)의 감정의 수레바퀴에서 상정하는 각 스펙트럼 내 감정은 해당 기본 감정 표지에 포함하고, 만일 8가지 기본 감정 스펙트럼에서 벗어나 어떠한 표지로도 정의될 수 없을 경우 '기타' 표지로 태깅하였다.

트위터 자료에서 공격성이 드러날 경우 '공격성 범주'를 파악하여 태깅하였는데 공격성 범주는 공격성을 촉발하는 원초적인 감정을 공격성 범주로 파악하였다. 또한 분석 대상 텍스트 내에서 화자(작성자)가 느끼는 감정의 대상이 되는 표현을 선별하여 이를 대상(target)으로 정의하여 분석하고, 감정의 대상(target)이 특정 인물 또는 집단일 경우 개인정보의 노출 등을 막기 위해 대상의 비식별화 처리를 진행하였다.

1-3. 말뭉치 감정 분석 및 말뭉치 구축

감정 분석은 선발된 10명의 주석자가 동일한 자료 1건의 문서에 대한 분석을 진행하고, 10명이 분석한 결과에 대한 통계적 일치도를 산출하여 일치도 산출값에 따라 적합 문서와

부적합 문서를 분리하고, 적합 문서는 일치도 산출 기준에 따라 문서별 표지를 확정하였다. 이때 카파통계량 중 평가자는 서로 독립임을 가정하는 Fleiss의 방법을 기반으로 산출하였다.

일치도 해석 기준에 따라 적당한 일치도 이상에 해당하는 k값 0.4 초과 문서를 적합 문서로 채택하고, 감정 표지는 채택 해당 문서 내에서 5명 이상(관찰된 일치 비율 0.4 초과)의 분석 결과가 일치한 표지만을 최종 채택하였다. 일상 대화 1만 건, 트위터 5만 건에 대한 말뭉치 감정·공격성 분석 진행 결과 감정 표지별 결과는 다음과 같다.

<표 22> 감정 분석 결과 표지별 비중

감정 분석 표지	일상 대화		트위터		합계	
	분석 표지 수 (건)	비중	분석 표지 수 (건)	비중	분석 표지 수 (건)	비중
기쁨(Joy)	5,710	57.1%	25,256	50.5%	30,966	51.6%
기대(Anticipation)	1,893	18.9%	10,773	21.5%	12,666	21.1%
신뢰(Trust)	414	4.1%	2,646	5.3%	3,060	5.1%
놀람(Surprise)	322	3.2%	3,229	6.5%	3,551	5.9%
혐오(Disgust)	1,296	13.0%	1,810	3.6%	3,106	5.2%
공포(Fear)	521	5.2%	1,015	2.0%	1,536	2.6%
분노(Anger)	144	1.4%	1,979	4.0%	2,123	3.5%
슬픔(Sadness)	1,097	11.0%	3,032	6.1%	4,129	6.9%
감정 없음	-	-	6,801	13.6%	6,801	11.3%

2. 의의 및 기대 효과

‘2022년 말뭉치 감정 분석 및 연구’ 사업은 일상 대화 및 트위터 자료를 토대로 감정 분석 지침을 수립하여 감정·공격성 분석을 진행하였다는 점에서 의의가 있다. 본 사업을 통한 기대 효과는 다음과 같다.

- 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 다양한 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.

참고 문헌

김민선, 송기준, 남충모, 정인선(2012), A Study on Comparison of Generalized Kappa Statistics in Agreement Analysis, 한국통계학회.

Mohammad, S. (2016), 9 - Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text. In H. L. Meiselman (Ed.), Emotion Measurement (pp. 201-237). Woodhead Publishing.

Mohammad, S. (2018), SemEval-2018 Task 1: Affect in Tweets. In Proceedings of The 12th International Workshop on Semantic Evaluation (pp. 1-17). Association for Computational Linguistics.

Zampieri, R. (2019), SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In Proceedings of the 13th International Workshop on Semantic Evaluation(pp. 75-86). Association for Computational Linguistics.

부록

[붙임 1] 국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용 허락 계약서

국가 언어 자원(말뭉치) 구축 및 활용 저작권 이용허락 계약서

저작자 및 저작권 이용허락자 _____ (이하 “권리자”이라 함)와 저작권 이용자 국립국어원(이하 “이용자”이라 함)은 아래 저작물에 관한 저작재산권 이용허락과 관련하여 다음과 같이 계약을 체결한다.

다 음

제1조 (계약의 목적)

본 계약은 저작재산권 이용허락과 관련하여 권리자와 이용자 사이의 권리관계를 명확히 하는 것을 목적으로 한다.

제2조 (계약의 대상)

본 계약의 이용허락 대상이 되는 권리는 아래의 저작물(이하 “대상저작물”)에 대한 저작재산권 중 당사자가 합의한 권리로 한다.

저작물: 저작자가 국립국어원의 2022년 말뭉치 감정 분석 및 연구 사업 기간(2022년 5월 2일부터 2022년 12월 2일까지) 동안 위 사업에 제공하는 모든 온라인 게시 자료
저작자: _____
종별: ☑ 어문저작물
권리: ☑ 복제권, ☑ 전송권, ☑ 배포권, ☑ 2차적저작물작성권, ☑ 편집저작물작성권

※ 저작권 이용허락에는 다음 사항을 포함한다.

1. 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 대상저작물을 일정한 형식으로 전자적 기록 매체에 담아 보존하는 일
2. 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 자모, 음절, 어휘, 어절, 구절, 문장 및 텍스트 단위의 국어 연구와 언어 정보 처리 분야에 응용하기 위해 대상저작물을 복제·변형(목차·머리말·도표·그림·각주 등의 편집 및 삭제, 언어 단위별 분리, 언어적·비언어적 정보 부착, 번역 등)하는 일
3. 국립국어원 및 국립국어원이 발주한 용역 사업의 수행자가 대상저작물 및 그 복제·변형물을 연구 및 기술 개발용으로 학계·연구기관·산업체 등이 이용할 수 있도록 제공·배포하는 일
4. 대상저작물 및 그 복제·변형물을 제공·배포받은 학계·연구기관·산업체 등이 국어 연구와 언어 정보 처리 분야 응용을 위하여 대상저작물 및 그 복제·변형물을 분석 및 처리하여 사용하는 것을 허락하는 일

제3조 (이용허락 기간)

대상저작물의 이용 허락 기간은 계약체결일부터 2038년 12월 31일까지로 한다. 권리자가 이용 허락을 갱신하지 않고자 한다면 이용 허락 기간이 끝나기 6개월 전부터 1개월 전까지의 기간에 이용자에게 서면으로 이용 허락 갱신거절의 통지를 하지 아니하면 이용 허락은 5년 단위로 자동 갱신되며 이용 허락 내용이 유지된다.

제4조 (권리자의 의무)

(1) 권리자는 이용자에게 대상저작물에 관하여 본 계약서 제2조에 따른 저작재산권을 이용할 권리 및 제3자에게 재이용을 허락할 권리를 제3조의 기간 동안 비독점적으로 허락한다.

(2) 권리자는 이용자에게 계약 체결일로부터 10일 이내에 대상저작물의 이용을 위해 필요한 상당한 자료를 인도하여야 한다. 다만, 대상저작물이 한국저작권위원회에 등록되어 있지 않은 경우 이용자가 요청하면 이용 허락자는 대상저작물의 저작재산권을 등록한 후 위 의무를 이행한다.

(3) 권리자는 대상저작물에 제3자의 이용 허락권, 질권 등이 존재하는 경우, 이용자에게 그 사실을 사전에 알려야 한다.

(4) 권리자는 대상저작물의 저작재산권 전부 또는 일부를 제3자에게 양도하거나 이에 대하여 질권을 설정하고자 하는 경우, 사전에 이용자에게 이 사실을 통보하여야 한다.

제5조 (이용자의 권리 및 의무)

(1) 이용자는 대상저작물을 제3조의 이용허락 기간 동안 제2조의 이용 허락을 받은 범위 내에서 비독점적으로 자유롭게 이용할 수 있고 제3자에게 재이용을 자유롭게 허락할 수 있다.

(2) 이용료는 설정하지 아니한다.

(3) 이용자는 관례적으로 저작자 및 저작재산권자의 성명 등 표시를 허용하는 대상저작물을 이용하는 경우, 그 저작자 및 저작재산권자의 성명 등을 표시하여야 한다.

(4) 이용자는 대상저작물을 이용함에 있어서 저작인격권을 침해하지 아니한다. 다만, 제2조에 규정한 바에 따라 대상저작물에 대한 변형 등을 할 수 있으며, 대상저작물의 본질적인 내용을 변경하지 않는 범위 내에서 수정 및 편집을 할 수 있다.

제6조 (확인 및 보증)

(1) 권리자는 이용자에게 다음 각 호의 사항을 확인하고 보증한다.

1. 대상저작물의 저작권이용허락을 체결하는 데 필요한 권리 및 권한을 적법하게 보유하고 있다는 것
2. 대상저작물의 내용이 제3자의 저작권, 상표권, 인격권을 비롯한 일체의 권리를 침해하지 아니한다는 것
3. 대상저작물에 대하여 이용자에게 사전에 알린 제3자의 권리 외에는 이용자의 이용을 제한할 수 있는 부담이 더 이상 존재하지 아니한다는 것

(2) 이용자는 권리자에게 다음 각호의 사항을 확인하고 보증한다.

1. 대상저작물의 이용허락을 받은 범위 내에서 제3자에게 재이용을 허락할 것
2. 대상저작물을 제3자의 명예권을 비롯한 인격적 권리를 침해하는 방식으로 이용하지 아니할 것

제7조 (계약내용의 변경)

본 계약 내용 중 일부를 변경할 필요가 있는 경우에는 권리자와 이용자의 서면합의에 의하여 변경할 수 있으며, 그 서면합의에서 달리 정함이 없는 한, 변경된 사항은 그 다음날부터 효력을 가진다.

제8조 (계약의 해지)

(1) 당사자는 천재지변 또는 기타 불가항력으로 계약을 유지할 수 없는 경우에 본 계약을 해지할 수 있다.

(2) 당사자는 상대방이 정당한 이유 없이 본 계약을 위반하는 경우에 상당한 기간을 정하여 상대방에게 그 시정을 최고하고, 상대방이 그 기간이 지나도록 이행하지 아니하는 경우에는 계약을 해지할 수 있다. 다만, 상대방이 명백한 시정 거부의사를 표시하였거나 위반 사항의 성격상 시정이 불가능하다는 것이 명백히 인정되는 경우에는 위와 같은 최고 없이 계약을 해지할 수 있다.

(3) 본 계약에 대한 해지권의 행사는 상대방에 대한 손해배상청구권 행사에 영향을 미치지 아니한다.

제9조 (손해배상)
당사자가 정당한 이유 없이 본 계약을 위반하는 경우, 그로 인하여 상대방에게 발생한 모든 손해를 배상할 책임이 있다. 다만, 제8조 1항의 사유로 본 계약을 이행하지 못한 경우에는 손해배상책임을 면한다.

제10조 (비용의 부담)
계약 체결에 따른 비용은 이용자가 전부 부담한다.

제11조 (분쟁해결)
(1) 본 계약에서 발생하는 모든 분쟁은 권리자와 이용자가 상호 원만한 합의에 이르도록 노력하여야 하며, 분쟁이 원만히 해결되지 않는 경우에는 소제기에 앞서 한국저작권위원회에 조정을 신청할 수 있다.

(2) 제1항에 따라 해결되지 아니할 때에는 대한민국의 민사소송법 등에 따른 관할법원에서의 소송에 의해 해결토록 한다.

제12조 (비밀유지)
양 당사자는 본 계약의 체결 및 이행과정에서 알게 된 상대방에 관한 정보, 본 계약의 내용을, 상대방의 서면에 의한 승낙 없이 제3자에게 공개하여서는 아니 된다.

제13조 (기타부속합의)
(1) 권리자와 이용자는 본 계약의 내용을 보충하거나, 이 계약에서 정하지 아니한 사항을 규정하기 위하여 부속합의서를 작성할 수 있다.

(2) 제1항에 따른 부속 합의는 본 계약의 내용과 배치되거나 위반하지 않는 범위 내에서 유효하다.

제14조 (계약의 해석 및 보완)
본 계약서에서 명시되어 있지 아니하거나 해석상 이견이 있을 경우에는 저작권법, 민법 등을 준용하고 사회 통념과 조리에 맞게 해결한다.

제15조 (계약 효력 발생일)
본 계약의 효력은 계약 체결일로부터 발생한다.

2022년 월 일

권리자 :	이용자 :
성명 (인)	성명 국립국어원장 (인)
주민등록번호(앞자리만)	주소 서울특별시 강서구 금낭화로 154

부록

[붙임 2] 감정·공격성 분석 지침

2022 말뭉치 감정 분석 및 연구: 감정 및 공격성 분석 지침

1. 감정 분석

1) 감정 분석 개요

본 사업에서 정의하는 감정 분석(Emotional analysis)이란 텍스트가 내포하는 정서를 파악하여 다양한 감정 유형으로 분류하는 것이다. 이는 텍스트의 주관성을 세분화한 감정 유형으로 분류함으로써 긍정 또는 부정 극성만을 분석한 감성 분석보다 정교화된 분석이라 할 수 있다. 예컨대 '원하던 대학에 합격해서 날아갈 것 같아. 즐거운 학교 생활이 될 거야'라는 텍스트에 기쁨, 기대의 감정 유형 표지를 부착함으로써 단순히 '긍정' 극성으로 분류하는 것에서 한 단계 더 나아가 작성자의 감정을 세밀하게 나타낸다.

2) 감정 텍스트 유형

텍스트에서 감정은 다양한 언어 표현을 통해 나타난다. 우리는 감정을 표현할 때 '기쁘다' 또는 '슬프다'와 같은 명시적 감정 어휘를 사용하여 자신의 감정을 직접적으로 나타낼 수 있다. 하지만 '콧방귀를 뀌다' 또는 '펄쩍 뛰다' 등의 관용 표현을 사용하거나 '문을 쾅 닫다' 등의 행동 묘사를 통해서도 감정을 간접적으로 전달하기도 한다. 또한 '그는 프로젝트를 훌륭히 해냈다', '걔는 항상 지저분하게 먹어' 등의 문장처럼 주관적 평가를 함으로써 사물이나 인물에 대한 감정을 드러내기도 한다. 때문에 감정 분석 대상 텍스트는 단순히 감정 어휘를 포함하는 글뿐 아니라 '맥락 속에서 화자(작성자)의 감정을 읽어낼 수 있는 텍스트'가 되어야 한다. 따라서 감정 분석 작업에서는 작업자가 직접 텍스트를 읽고 표면적으로 드러나지 않는 화자의 감정까지 정밀하게 분석하는 것이 중요하다. 다양한 감정 표현 수단을 사용한 트위터 게시글 예시는 아래와 같다.

유형	트위터 게시글 예시	
명시적 감정 어휘 사용	1	와 나 진짜 드라마 보는데 **넘 빡쳐서 몬살겟어** 박피디 나쁜놈아
관용 표현	2	오늘 두 명이나 고민하고 있어서 식겁하고 말 걸었다가 선물받아서 맘 놓구 고민하는 차둘이한테 말 걸었는데 이사고민이라 **심장 철렁했네** (ꉂꉂ) 주대 맘을 들었다놨다 들었다놨다 해이...
행동 묘사	3	혹시 이장면을 말씀하시는 건가요 **저 이장면보고 폰 던졌어요** 다자이 이 (왈왈) 진짜 한대만 치자 제발 이러면서 봤어

유형	트위터 게시글 예시	
주관적 평가	4	문나이트 다리쪽엔 고대 상형문자가 있는데, '달을 수호하는 전사'라는 말을 만들어서 고대 이집트 문나이트가 되었다고 함. 여기서 오스카 아이작이 수트 입는 장면 나옴!□◗ **슈트도 클로즈업 한 거 보니까 캐릭터랑 배경 생각해서 어우러지게 잘 만든 것 같아서 감탄 중. 의상팀 존나 배운 변태같애**

<표 1> 감정 표현 유형 예시(트위터 게시글)

3) 감정 분석 대상

(1) 일상 대화 담화문

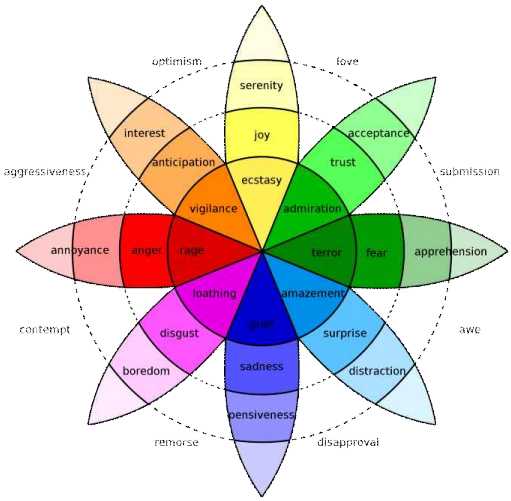
일상 대화 담화문은 2020년 국립국어원에서 구축한 일상 대화 말뭉치 자료를 기반으로 하는 문서이다. 일상 대화 말뭉치 자료는 2인 발화자의 15분 분량의 일상 대화를 녹음하여 억양구 단위로 전사한 자료이다. 본 사업에서는 해당 자료에서 감정이 드러난 5어절 이상의 담화를 선별하여 감정 분석 대상 문서를 구성하였다. 감정 분석 작업에서는 일상 대화 담화문 총 1만 건의 감정 표지 부착을 목표로 한다.

(2) 트위터 게시물

누리소통망(SNS)은 다양한 사람들이 제한 없이 자신의 의견과 생각을 자유로이 주고받는 매체임은 물론 개인의 감정까지도 제재 없이 표출할 수 있는 창구의 역할을 한다. 이 때문에 화자(작성자)가 생성한 텍스트에서 감정이 담긴 표현을 포착하여 감정을 분석하기에 매우 용이한 자료라 할 수 있다. 본 사업의 목표는 문화콘텐츠 관련 트위터 게시물 5만 건에 대한 감정 및 공격성을 분석하여 분석 표지를 부착하는 것에 있다.

4) 감정 분석 표지

인간의 감정을 분류하기 위한 연구는 긴 시간 동안 다양한 연구자에 의해 이루어져 왔다. 그러나 학계 및 산업계의 보편적 합의를 이끌어낸 감정 체계는 아직까지 확립되지 않은 상태이다. 이에 본 연구에서는 과업의 효율성을 위하여 분석 대상 텍스트 유형 및 연구 목적에 중점을 두고 트위터 게시물 텍스트의 감정을 분석한 연구 SemEval 2018 Task 1: Affect in Tweets(이하 SemEval 2018)을 참고하여 감정 분석 표지를 구성하였다. SemEval 2018에서는 정서심리학자 로버트 플루칙(Robert Plutchik)이 주장한 기본 감정 및 복합 감정을 선별하여 감정 분석용 표지를 구성하였으며, 감정 체계는 아래 그림과 같다.



[그림 1] 플루칙의 감정 수레바퀴
(Robert Plutchik's Wheel of Emotions)

로버트 플루칙(Robert Plutchik)은 위 그림과 같이 인간의 기본 감정을 8가지로 분류하였으며 감정 강도의 차이를 두어 각 기본 감정에 대한 스펙트럼을 형성하였고 다수의 감정이 함께 나타나는 것을 복합 감정으로 설정하여 감정 체계를 구성하였다. 본 연구에서는 SemEval 2018과 같이 로버트 플루칙(Robert Plutchik)의 감정 체계를 차용하되 한국어 특성 및 국내 연구 동향을 고려하여 8가지 기본 감정(기쁨, 기대, 신뢰, 놀람, 혐오, 공포, 분노, 슬픔)만을 감정 표지로 활용하고자 한다. 이때 감정의 수레바퀴에서 상정하는 각 스펙트럼 내 감정은 해당 기본 감정 표지에 포함하도록 한다. 만일 8가지 기본 감정 스펙트럼에서 벗어나 어떠한 표지로도 정의될 수 없을 경우 '기타' 표지로 태깅하도록 한다. 각 감정 분석 표지에 대한 세부 설명은 아래와 같다.

감정 분석 표지 (동일 스펙트럼 감정)	**설 명**
기쁨 (평온, 황홀 포함)	어떤 만족감에 의해 느끼는 즐겁고 흥겨운 감정.
기대 (경계, 관심 포함)	앞으로 있을 일이나 상황을 미리 짐작함. 또는 그런 내용.
신뢰 (수용, 감탄/존경 포함)	굳게 믿고 의지함.
놀람 (놀라움, 부주의/방심)	어떤 일이 뜻밖이거나 훌륭하거나 무서워서 신기해하거나 흥분하여 가슴이 뛰는 느낌.
혐오 (지루함, 증오 포함)	싫어하고 미워함.
공포 (불안, 두려움 포함)	두렵고 무서움.
분노 (짜증, 격노 포함)	분개하여 크게 화를 냄.
슬픔 (수심, 비탄 포함)	마음이 아프거나 괴로운 느낌.
기타	8가지 표지 외 기타 감정.

<표 2> 감정 분석 표지

인간의 감정은 그 구조와 관계가 복잡다단하여 텍스트의 감정을 단일 표지로 정의할 수 없는 경우가 존재할 수 있다. 예를 들어 복합 감정이 나타나는 경우, 하나의 문서에 다수의 감정이 나타나는 경우, 동일 텍스트 내에서 감정이 변화하여 연속적으로 나타나는 경우가 그러하다. 이러한 문제를 해결하기 위하여 본 연구에서는 복합 감정, 복수 감정, 연속 감정을 분석하기 용이한 감정 복수 태깅이 가능하도록 설정하였다.

5) 세부 감정 분류(참고 사항)

발화자 본인의 감정 및 주관적 의견을 대상으로 한다. 타인이 느끼는 감정에 대한 언급은 분석의 대상이 아니다. 이에 주의하여 분석하며, 감정 판단이 어려울 경우 아래의 '세부 감정 분류'와 '복합 감정 분류표'를 참고하되, 절대적 기준은 아니므로 감정 분류를 위한 참고 자료로만 사용한다.

감정 표지	설명 및 감정 분류 예시
기쁨	행복감, 즐거움, 재미있음, 신남, 감동, 평온, 황홀, 통쾌함, 설렘, 반가움, 성취감, 뿌듯함, 만족감, 자랑스러움, 후련함, 안도, 다행스러움
기대	예상, 경계, 조심성, 관심, 흥미로움, 바람, 갈망, 부러움
신뢰	수용, 감탄, 존경, 믿음
놀람	놀라움, 부주의, 방심, 당황함, 황당함, 어이없음, 신기함
혐오	역겨움, 무관심, 지루함, 따분함, 심심함, 증오, 미움, 싫어함
공포	불안, 두려움, 섬뜩함, 겁남, 무서움, 우려, 걱정, 불안, 초조함
분노	짜증, 격노, 화남, 약오름, 답답함
슬픔	수심, 비탄, 안타까움, 속상함, 연민, 측은함, 불쌍함, 아쉬움, 허전함, 허무함, 허탈함, 비참함, 서운함, 서러움, 외로움, 쓸쓸함, 우울함, 좌절, 절망, 후회, 미안함, 실망, 수치심, 부끄러움, 민망함, 창피함
기타	8가지 표지에 속하지 않는 기타 감정

<표 3> 감정 표지 및 설명

복합 감정	설명 및 예시
격려	기쁨, 신뢰
사랑	
감사(고마움)	
굴복	신뢰, 공포
순종	
경악	공포, 놀람
경외	
난감함	놀람, 슬픔
죄책감	슬픔, 혐오
회한	
경멸	혐오, 분노
공격성	분노, 기대(예상)
낙관	기쁨, 기대(예상)
희망	
응원	
그리움	기대, 슬픔
심란함	슬픔, 공포
착잡함	
혼란스러움	

<표 4> 복합 감정 분류표

2. 공격성 분석

1) 공격 대상 범주 분석

게시글에서 공격성이 드러날 경우 작업자는 공격성이 향하는 대상(target)을 고려하여 '공격 대상 범주'를 태깅한다. 공격성이 특정 인물을 향하고 있을 경우 '개인' 표지를, 특정 집단을 향할 때는 '집단' 표지를 부착한다. 개인이나 집단에 속하지 않는 상황, 사건, 구조 등에 대한 공격성을 드러내는 게시글의 경우 '기타' 표지로 태깅하여 공격 대상을 분석한다. 공격 대상의 실질적 예시는 아래 표와 같다.

공격 대상 표지	설 명	공격 대상 예시
개인	익명 또는 실명의 특정 인물	양세형, 기리보이, kim1988 등
집단	특정 단체 및 개인이 아닌 다수의 집단	씨제이, 엑소팬덤, 여고생 등
기타	구조, 상황, 사건, 콘텐츠 등 개인이나 집단에 속하지 않는 대상(비인간)	도서정가제, 코로나, 또 오해영 등

<표 5> 공격 대상 범주 표지

앞서 서술한 공격성 범주 표지와 공격 대상 범주 표지를 모두 부착한 트위터 공격성 분석의 결과물 예시는 아래와 같다.

구분	트위터 본문	공격 대상 (target)	공격 대상 범주
1	아수라도 평범한 알탕영화인데 대사가 살린 영화자너 대사의 힘은 엄청나 못생긴 **대사** 죽어	대사	기타
2	**철권** 진짜 개쓰레기 게임 **하라다** 미친새끼야 밸런스패치 해라 진짜 @TEKKEN	철권, 하라다	기타, 개인
3	오늘을 기점으로 이번 시즌 끝날 때 까지 야구 접습니다. **쓰레기팀**에 더이상 감정소모 하고싶지 않네요 https://t.co/1TswBC6OHu	쓰레기팀	집단

<표 6> 트위터 게시물 공격성 분석 예시

<기획·연구>

국립국어원 강미영 언어정보과장
국립국어원 이보라미 학예연구관
국립국어원 서샛별 학예연구사
국립국어원 서혜진 연구원

<사업 참여자>

사업 책임자 이영희 (버즈메트릭스)
사업 참여자 김수진 (버즈메트릭스)
유지현 (버즈메트릭스)
이준수 (버즈메트릭스)
권주원 (버즈메트릭스)
신현주 (버즈메트릭스)
김도현 (버즈메트릭스)
이진상 (버즈메트릭스)
한희재 (버즈메트릭스)
서은미 (버즈메트릭스)

발행인: 국립국어원장
발행처: 국립국어원
서울시 강서구 금낭화로 154
전화 02-2669-9775, 전송 02-2669-9727
인쇄일: 2022년 12월 2일
발행일: 2022년 12월 2일
인 쇄: ㈜타라그래픽스

※ 이 책은 국립국어원의 용역비로 수행한 '2022년 말뭉치 감정 분석 및 연구' 사업의 결과물을 발간한 것입니다.