감염병 대유행 상황에서 사용자의 인식 변화를 반영한 온라인 리뷰의 텍스트 분석을 이용한 빅데이터 기반 레스토랑 추천시스템

2022.05.06.금요일

홍준우(부산대학교 경영학과)
홍태호(부산대학교 경영학과)

2022 한중핵심공동연구사업
부산대학교 교내 워크숍

목 차

1. 서론

- 정보통신기술의 발전과 스마트 기기의 대중화로 **다양한 산업의 온라인에서 온라인 리뷰 형태로 의견공유**가 빈번하게 이루어지고 있음.

- 온라인 리뷰 형태로 공유되는 의견은 **고객관리 및 확장을 위한 필수적 요소로 사용**됨(Eugene and Mary, 1993)

- 다양한 산업의 온라인 리뷰는 **최근 코로나19 감염병의 영향으로 데이터의 다양화를 가속화**시키고 있음(Al-maaitahet al., 2021).

- 소비자의 경험을 기반으로 작성되는 온라인 리뷰는 **입장에 따른 이점을 가짐**
  - 고객: **상품 탐색 및 구매 과정에서 구매 의사결정에 대한 이점**(Cheng and Ho., 2015)
  - 기업: **고객 분석을 통해 매출 증가 원인, 고객 패턴 파악 등에서 이점** (Li et al., 2013).

- 시장의 활성화에 따른 상품의 다양화는 **직관적 선택과 의사결정 능력을 저하**시키고 있어(이진화 등, 2008)
  고객에게 맞춤형 정보를 제공하는 **개인화 추천시스템의 필요성이 대두**되고 있음(Kim et al., 2010)

- 온라인 리뷰의 **상품 속성은 구매 의사결정에 영향**을 미치는데(Vany and Walls, 1996),
  총평점만을 사용한 기존 추천시스템 연구는 상품의 세부항목을 반영하지 못함(윤호민 & 최규완, 2020)

- **정량적 데이터(총평점)만을 이용한 추천시스템은 정확도를 떨어뜨린다는 문제점**이 제기되고 있음 (Jeon and Ahn, 2015).

- 본 연구에서는 정량적 데이터뿐만 아니라 정성적 데이터를 함께 활용한 개인화 추천시스템을 제안함.

2. 문헌연구

[코로나19 팬데믹 전/후의 인식변환]

- 코로나19 팬데믹(pandemic)은 2020년 3월 11일 세계보건기구(WHO)에 의해 선포되었으며, **전 세계적으로 많은 피해를 발생**시키고 있음(WHO, 2022)
  - 인명 피해: 미국의 코로나19 확진자 수 80,598,784명, 사망자 수 986,437명으로

    전 세계적으로 가장 많은 누적 확진자 및 사망자 수를 기록함(중앙방역대책본부 2022년 5월 5일).
  - 경제적 피해: **전세계의 92.9%의 국가가 경제적 피해**를 입었으며, 이는 40.9%의 천연두 팬데믹, 83.3% 대공황 시기 보다 큰 피해 규모임(홍태희, 2020)
  - 정서적 피해: 코로나19로 인한 대표적 정서적 피해로 **코로나블루**가 있으며,

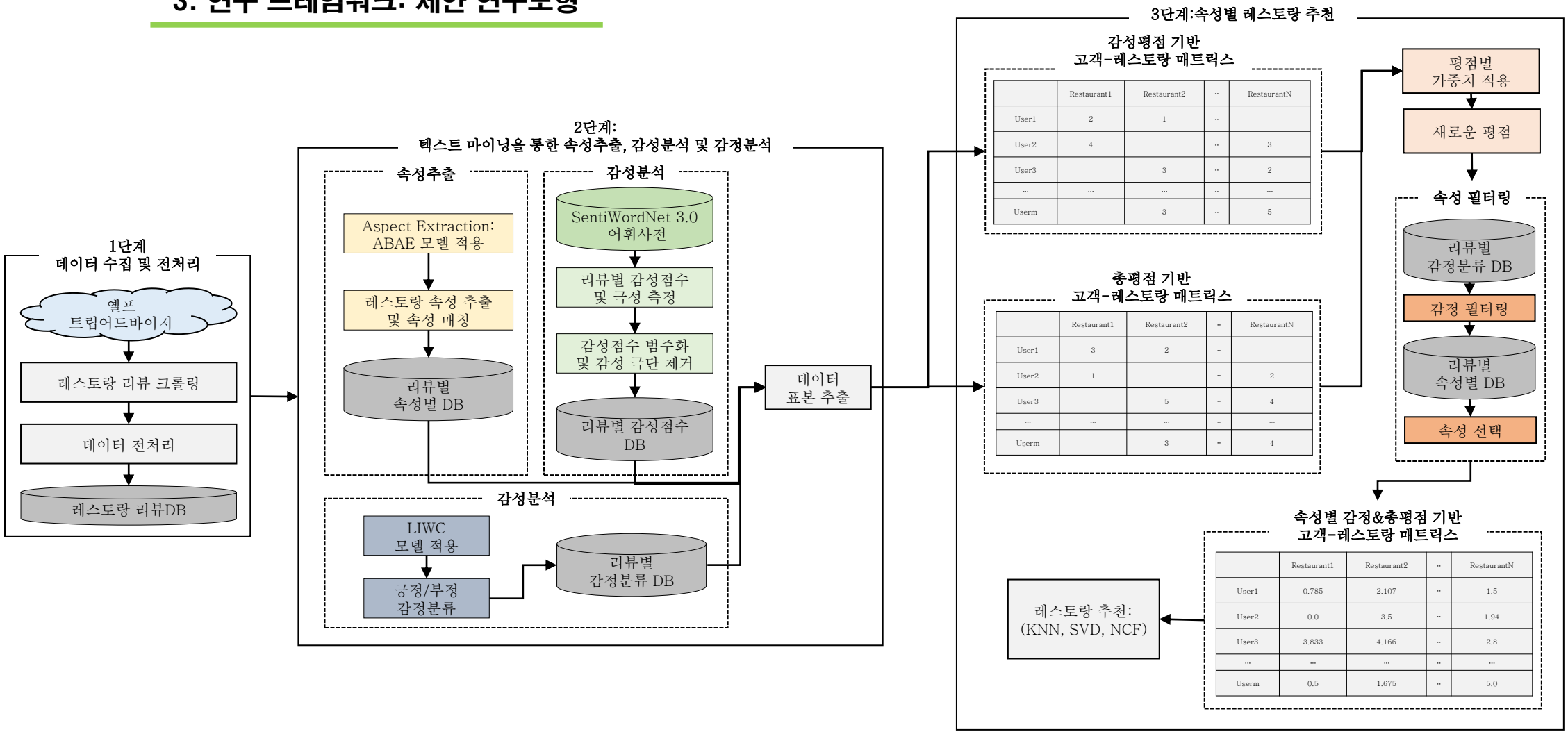
    이는 코로나19 확산 방지 정책으로 **인해 사람들의 사회활동이 위축**되면서 많이 발생함(손헌일 외, 2020)

    또한 이외에도 코로나19 의심환자의 **잠재적 결과에 대한 두려움**, 격리된 환자의 **지루함, 외로움, 분노 등의 감정**을 느낄 때 발생함(Liu et al., 2020)

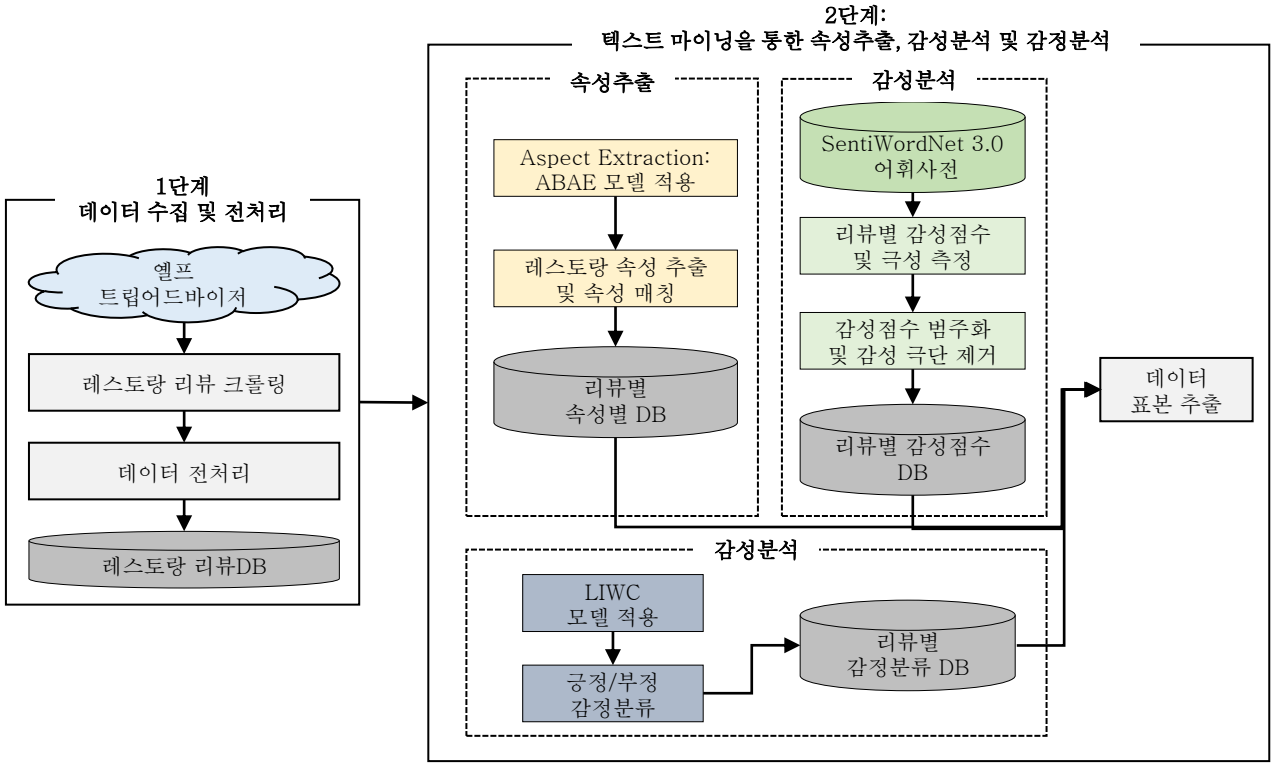
- 최근 국내 및 국외의 국가별 코로나19 방역정책이 완화되는 추세로써 기존의 코로나19 확산방지 정책인 '사회적거리두기', '도시봉쇄', '집합금지', '여행제한 ' 등 의무적 요소에서 **선택적 요소로 변경**되어 코로나19 이전의 모습과 같은 일상으로의 복귀를 준비하고 있음

- 코로나19 팬데믹 전/후, **국가별 정책이 완화되면서 사용자의 인식변화를 반영한 의사결정이 필요**함
  - 코로나19 확산 전/후 온라인 리뷰를 통한 관광객의 감정 및 감성분석에서 행복지수는 상승했으며, 슬픔지수는 하락함(Olga Chernyaeva et al., 2022)
  - 코로나19 감염병의 시기를 텍스트 마이닝을 통해 관광객의 감성을 분석했을 때 코로나19 이후 시기의 감성점수가 증가함(홍준우 & 홍태호, 2021).

3. 연구 프레임워크: 제안 연구모형



3. 연구 프레임워크: 단계별 설명(1)



[1단계: 데이터 수집 및 전처리]

① 데이터 수집: “옐프”, “트립어드바이저”의 레스토랑 리뷰 데이터를 수집

② 데이터 전처리: 불필요한 텍스트 데이터 제거 & 변환

[2단계: 텍스트 마이닝을 통한 속성추출, 감성분석 및 감정분석]

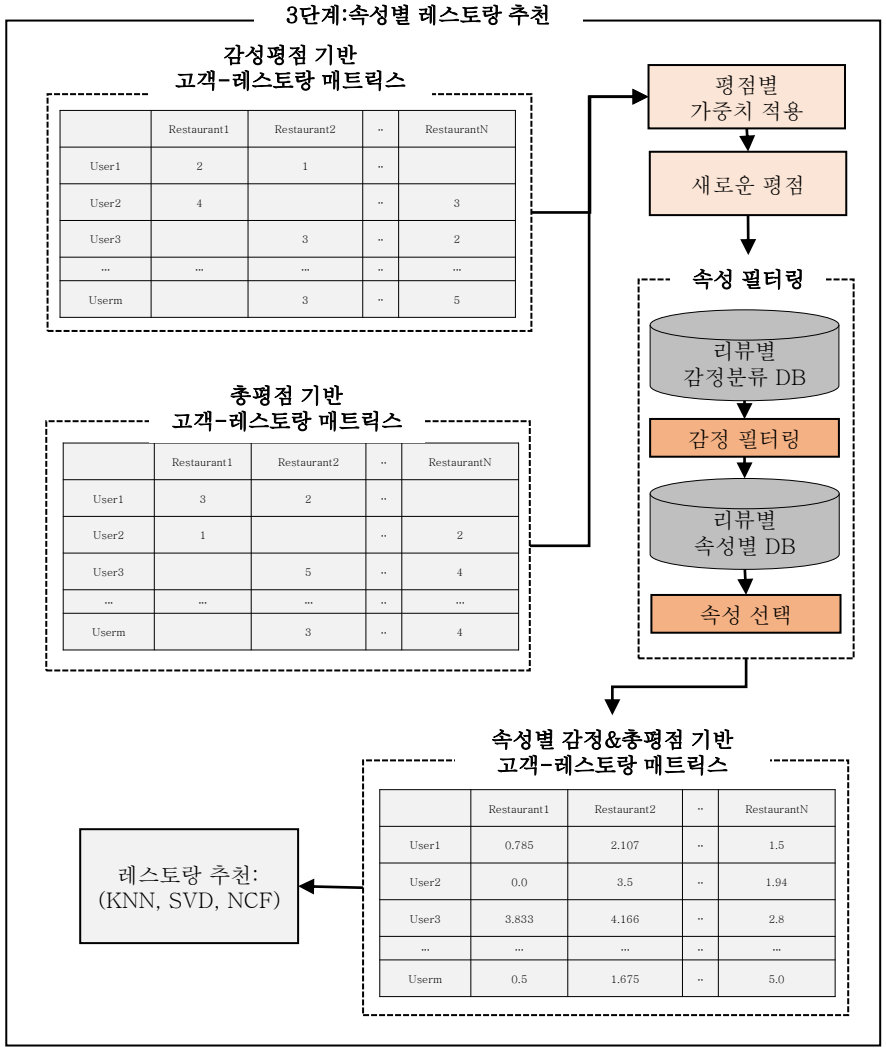
① 감성분석: SentiWordNet3.0 어휘사전을 기반으로 감성점수 및 극성 측정
그리고 군집분석을 통해 감성점수를 범주화

② 속성추출: ABAE 모델을 통해 레스토랑의 4가지 속성을 추출 및 대표 단어 매칭

③ 감정분석: 자연어처리 툴인 LIWC를 통해 감정분석을 하여 긍정/부정 감정을 분류

④ 데이터 표본 추출: 데이터 희소성 검토, 극단치 제거

3. 연구 프레임워크: 단계별 설명(2)



[3단계: 속성별 레스토랑 추천]

① 감성평점 기반 고객-레스토랑 매트릭스

② 총평점 기반 고객-레스토랑 매트릭스

③ 속성별 감정 & 총평점 기반 고객-레스토랑 매트릭스:

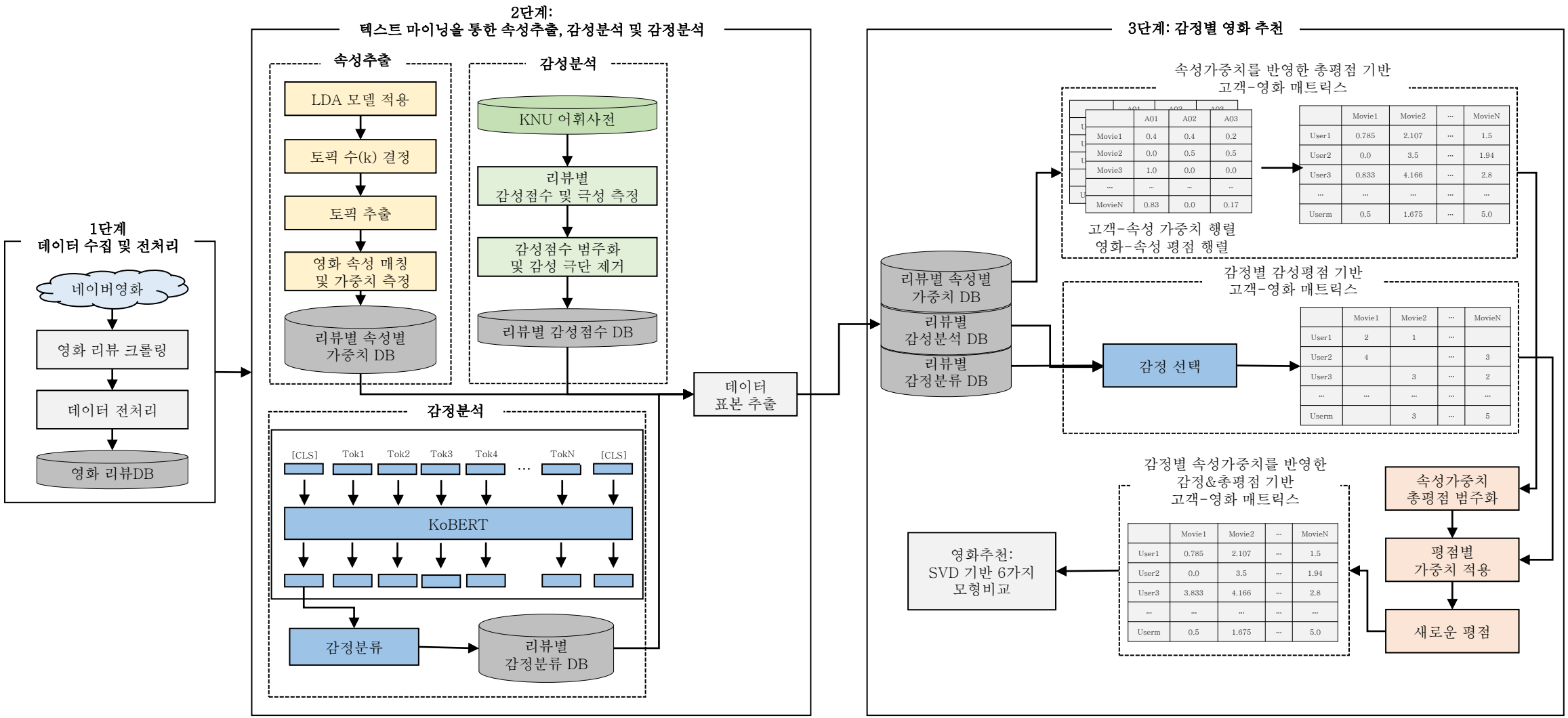
- 두 가지 평점별 가중치를 적용하여 하나의 새로운 평점 생성
- 속성 필터링: 긍정 감정에 대한 속성별 평점을 추출

④ 개인화 추천시스템: 속성별 특성을 반영한 레스토랑 추천

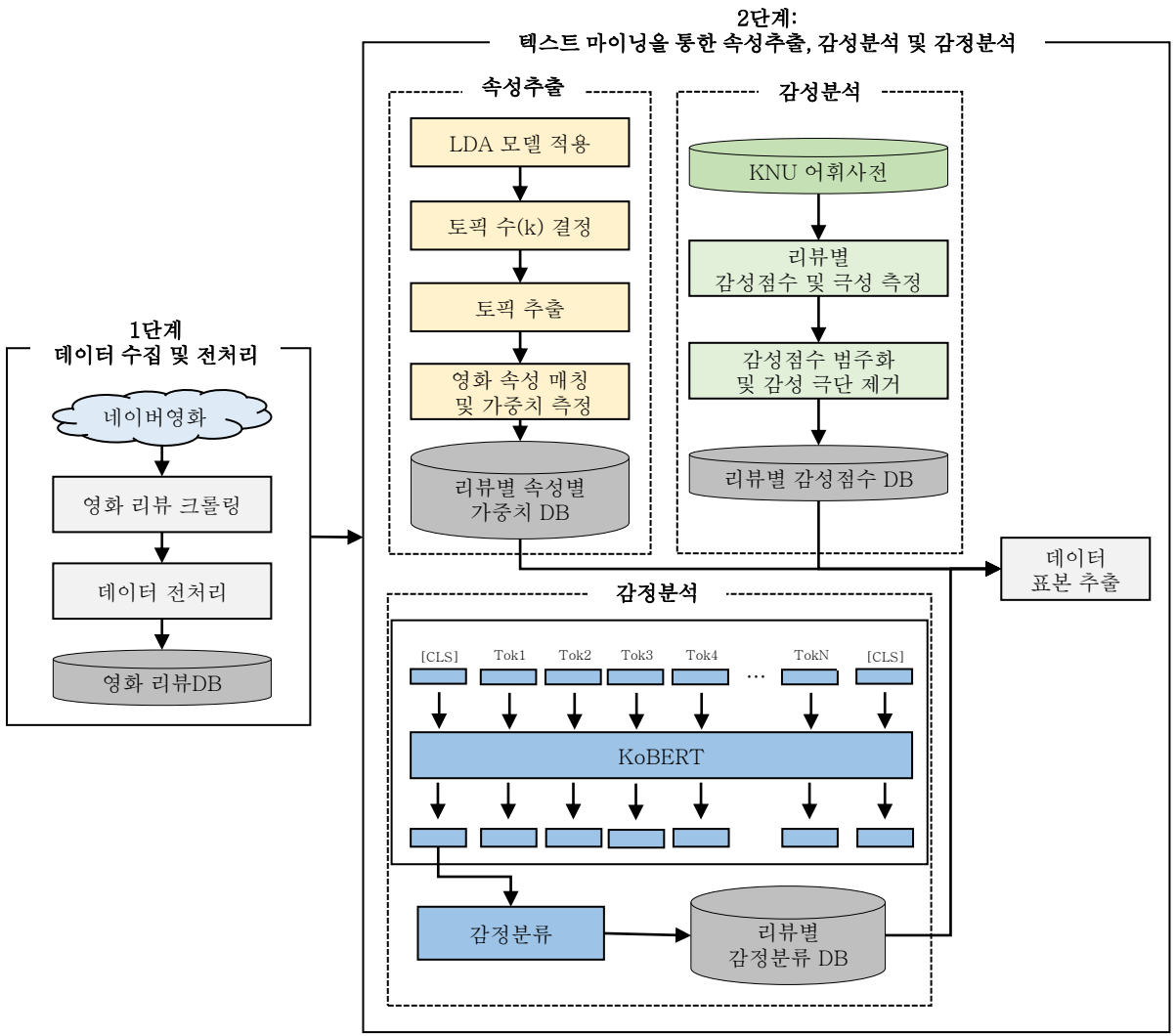
4. 실험 및 실험결과:

개인화 추천시스템: <네이버영화> 중심으로

4. 실험 및 실험결과: <네이버영화> 연구모형



4. 실험 및 실험결과: <네이버영화> 단계별 설명(1)



[1단계: 데이터 수집 및 전처리]

① 데이터 수집: <네이버영화> 제공하는 고객 영화 리뷰 데이터를 수집

② 데이터 전처리: 불필요한 텍스트 데이터 제거 및 변환

[2단계: 텍스트 마이닝을 통한 영화 속성추출, 감성분석 및 감정분석]

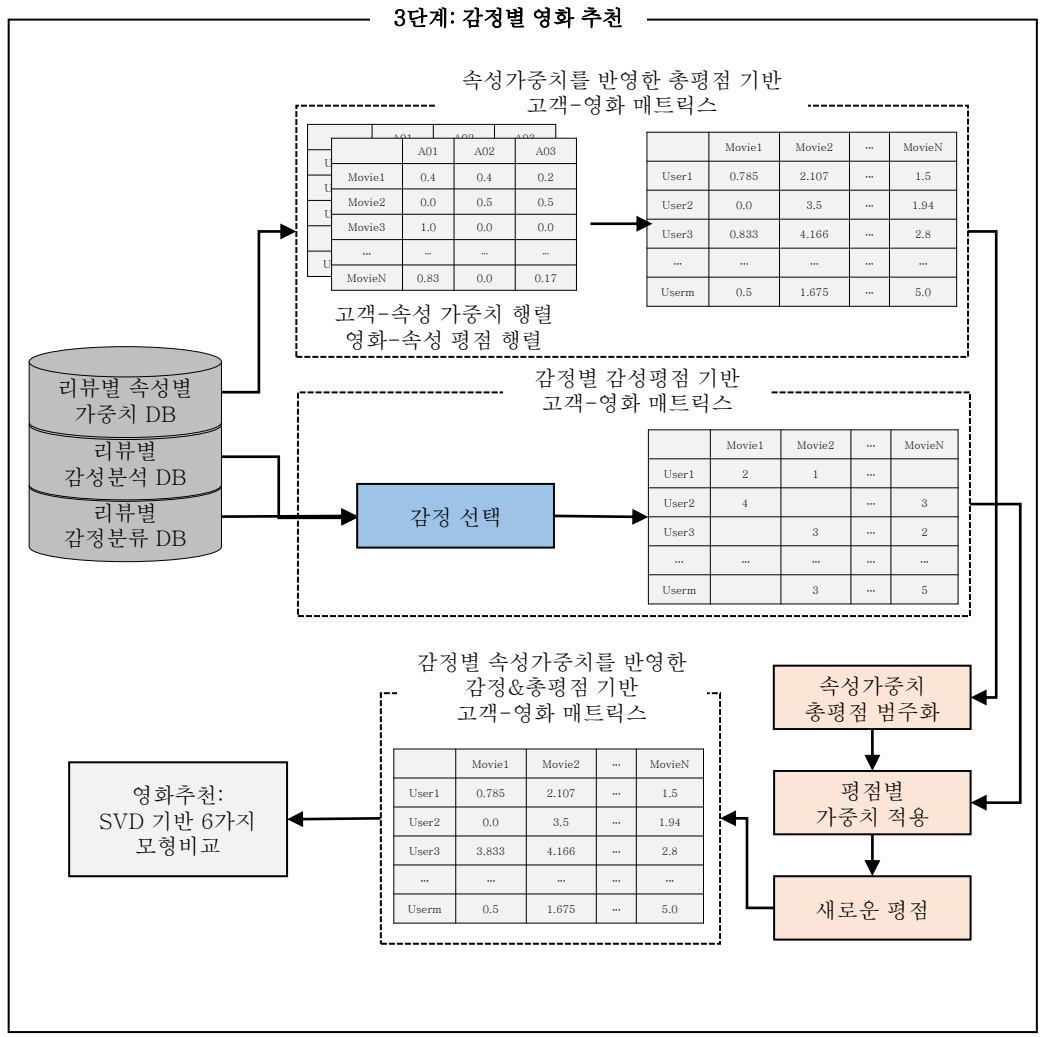
① 감성분석: KNU 어휘사전을 기반으로 감성점수 및 감성 극성 측정
그리고 군집분석을 통한 감성점수의 극단치 제거, 감성점수 범주화

② 속성추출: LDA를 통해 추출한 단어를 3가지 영화 속성으로 네이밍, 대표단어 정의
그리고 리뷰별 속성 가중치 측정

③ 감정분석: BERT 종류의 한국어 버전인 KoBERT를 사용하여 7가지 감정으로 다중분류

④ 데이터 표본 추출: 데이터 희소성 검토

4. 실험 및 실험결과: <네이버영화> 단계별 설명(2)

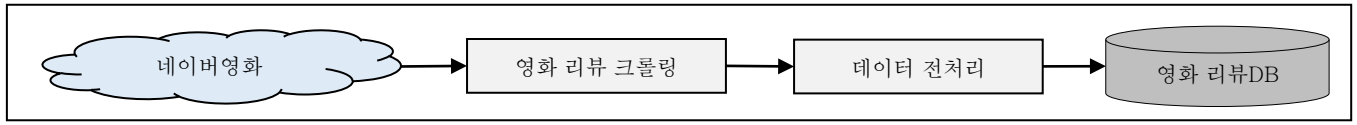


[3단계: 감정별 영화 추천]

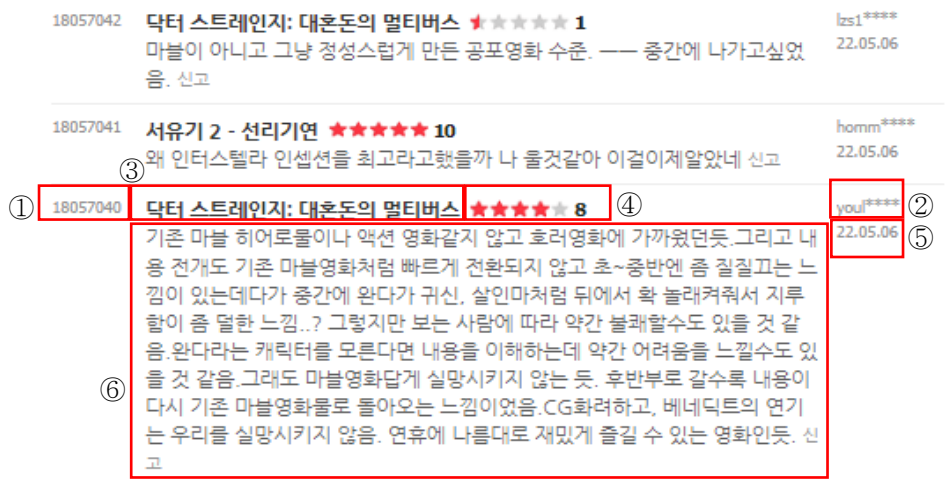
- 평점별 가중치를 적용한 새로운 평점:
  - **속성가중치 반영한 총평점 기반 고객-영화 매트릭스**: 실수형태를 정수형태로 범주화
    - **고객-속성 가중치 행렬**: 고객별 속성의 합이 0 혹은 1이 되도록 구성
      ※ 참고: 속성의 합이 0인 경우는 리뷰에 속성가중치가 존재하지 않는 경우에 해당함.
    - **영화-속성 총평점 행렬**: 영화별 속성의 합이 영화의 평균 총평점이 되도록 구성
  - **감정별 감성평점 기반 고객-영화 매트릭스**
- 감정별 속성가중치를 반영한 감성 & 총평점 기반 고객-영화 매트릭스
- 감정별 개인화 영화 추천

4. 실험 및 실험결과: 1단계 데이터 수집

- 데이터 수집 및 전처리:
  ① 데이터 수집:
  - 국내 검색포털 시장에서 네이버는 76.7%의 점유율을 갖고 있으며,
    PC뿐만 아니라 모바일 서비스 측면에서도 높은 우위를 선점하고 있음(심형섭 외, 2015).
  - 2022년 2월 추정순이용자는 26,812,367명으로 국내 최고 포털사이트로 인정받고 있음(Nielsen 2022년 2월).
  - <네이버영화> 페이지에서 **웹 크롤링**을 통해 **데이터(리뷰 고유번호, 고객ID, 영화ID, 총평점, 리뷰 작성일자, 고객 리뷰 등)을 수집**



<그림 1>. 1단계: 데이터 수집 및 전처리



<그림 2>. <네이버영화> 웹 페이지의 고객 평점 리뷰 데이터 샘플

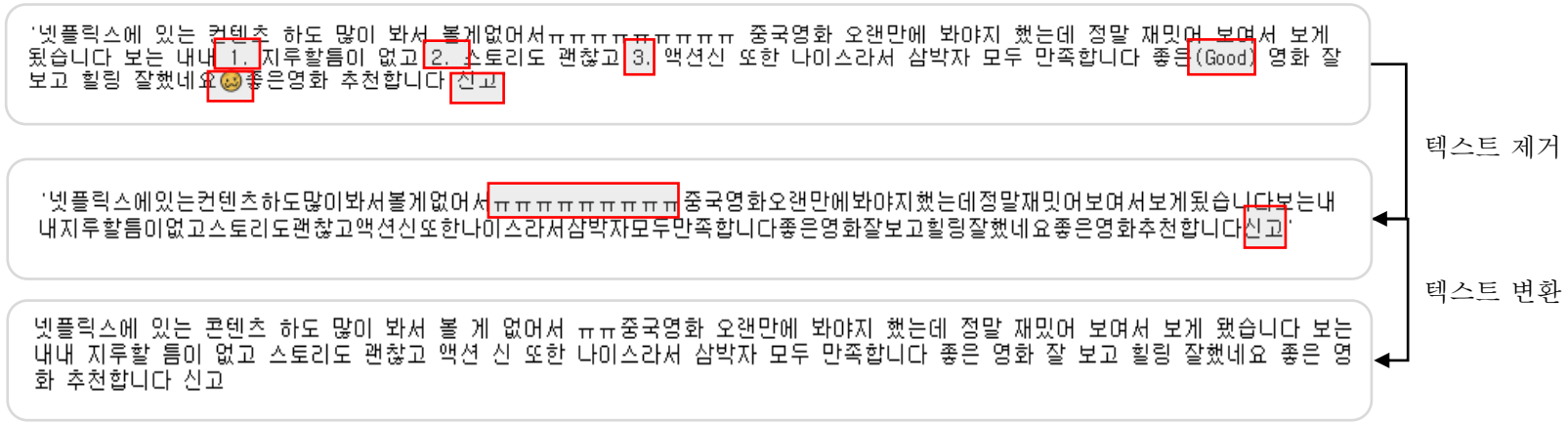
4. 실험 및 실험결과: 1단계 데이터 전처리

② 데이터 전처리:

- 총평점 범주화: 1부터 10사이의 **총평점을 5개로 범주화**
- **텍스트 제거**: 웹사이트 링크, 특수문자, 한국어 불용어, 공백, 한글을 제외한 모든 글자(숫자, 영어 등) 등

※ 한국어 불용어: URL, 웹사이트, 텍스트 및 문서 등"에 대한 키워드 분석을 다루는 RANK NL에서 제공함.

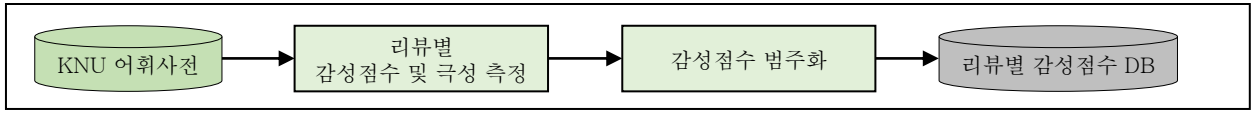
- **텍스트 변환**: 띄어쓰기, 맞춤법 검사기, 반복문자(이모티콘, 자음, 모음), 외래어



<그림 3>. 데이터 전처리: 텍스트 제거 및 변환 예제

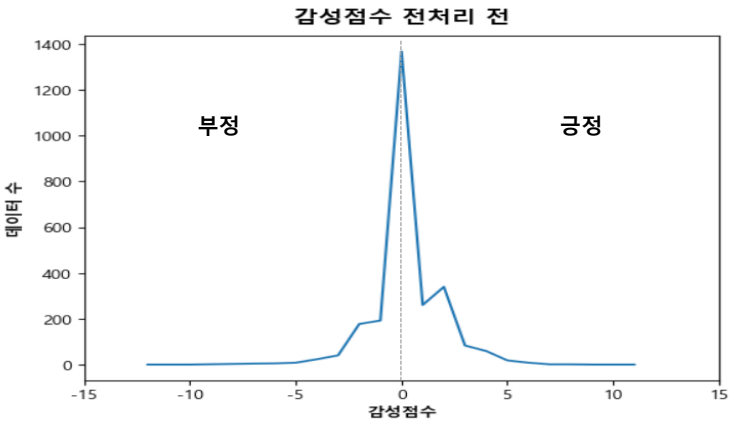
4. 실험 및 실험결과: 2단계 감성분석

- 텍스트 마이닝을 통한 속성추출, 감성분석, 감정분석:
  - ① 감성분석:
    - **감성점수 & 감성극성 추출**: 어휘 기반(**KNU감성 사전**)을 통해 **명사, 형용사, 동사**를 기반으로 측정
    - 감성점수 범주화:
      - 초기값: 감성점수는 -12부터 +11로 분포하며, **군집분석 & 시각화를 통해 감정점수의 양극단 값 38개 제거**(점수별 10개 미만)
      - 범주화: (감정 극단값 제거 후) 감성점수는 -4부터 +5로 조정되었으며, 군집분석을 통해 **5개로 범주화**



<그림 4>. 2단계: 감성분석(Sentiment Analysis)

-12	1
-10	1
-7	5
-6	6
-5	9
-4	24
-3	41
-2	178
-1	193
0	1367
1	261
2	340
3	84
4	60
5	19
6	9
7	2
8	2
9	1
10	1
11	1



<그림 5> 초기 감성점수 분포

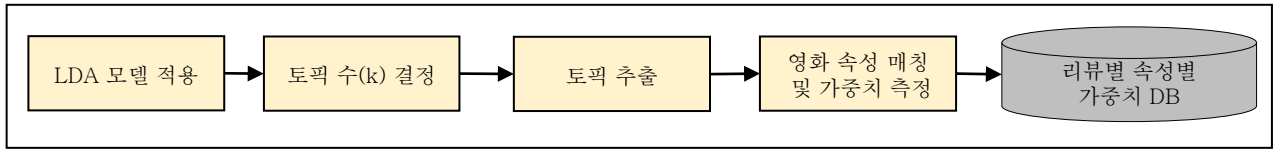
구축 평점	감성값 분포
1	-4, -3, -2
2	-1
3	0
4	1, 2
5	3, 4, 5

<표 1> 감성점수 범주화

4. 실험 및 실험결과: 2단계 속성추출

② 속성추출:

- 토픽모델링(LDA)으로 토픽을 추출했으며 혼잡도(Perlexity), 일관성(coherence) 성과지표를 통해 **최적의 토픽 수(K) 20**을 측정함
- 영화 리뷰를 **핵심적, 주변적, 정서적 속성으로 분류**했으며(Neelameghamand Jain, 1999)
  이를 기반으로 **영화 속성별 대표단어를 매칭**함.
- **리뷰별 3가지 속성에 대한 가중치를 측정**함.



<그림 6>. 2단계: 속성추출: LDA(Latent Dirichlet Allocation)

구분	속성	키워드
영화	핵심적	영화, 생각, 연기, 스토리, 사람, 배우, 아버지, 인물, 감독, 선거, 여배우, 빨갱이, 정부, 버스, 액션, 미화, 대통령, 점주
	주변적	공백, 실화, 시간, 장면, 오류, 분위기, 연출
	정서적	상황, 생각, 사랑, 공포, 이상하게, 상상, 지금, 과거, 기억, 긴장감, 완성, 실망, 감동, 울해, 감동, 최고, 매우, 사실, 지루하고, 감동, 생각, 아깝다, 자극, 충격, 얼마나, 최고, 감동, 좋네요

<표 3>. 영화 속성 및 속성별 대표단어

영화 속성	내용
핵심적	영화 내용과 직접적으로 연관된 키워드와 관련된 속성 ex. 스토리, 연기, 출연배우 등
주변적	영화의 전반적 흐름과 관련된 속성. ex. 배경, 의상, 배경음액, 특수효과 등
정서적	영화 시청에 대해 고객이 느끼는 감정과 관련된 속성 ex. 감동과 슬픔, 재미 등

<표 2>. 영화 속성별 개념 및 예제

속성 구분	핵심적	주변적	정서적
리뷰 수	933	340	350

<표 4>. 속성별 리뷰 수

4. 실험 및 실험결과: 2단계 감정분석

③ 감정분석:

- KoBERT를 통한 다중 감정분류: **7가지(행복, 놀람, 중립, 공포, 분노, 슬픔, 혐오)**
- 학습 데이터: AI Hub에서 제공하는 "한국어 단발성 대화 데이터"

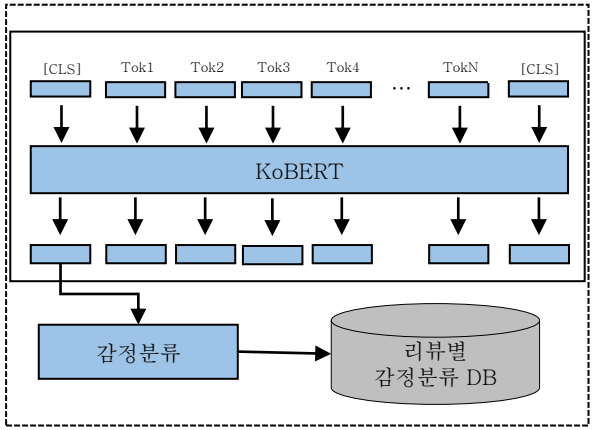
※ SKT Brain에서 **공개한 BERT 종류의 기계번역 모델로서 KoBERT는 한국어 성능의 한계를 극복**하고자 개발되었음
기존의 BERT 모델에서 한국어 데이터를 추가로 학습시켜 한국어 데이터에 대해 높은 정확도를 낼 수 있는 모델임
* **한국어 위키에서 5백만개의 문장과 54백만개의 단어를 학습시킨 모델**

총평점	행복(996)	놀람(151)	중립(519)	
	슬픔(311)	혐오(168)	분노(51)	공포(49)

<표 5>. 감정분류 데이터 수

구분	SNS 글 및 온라인 댓글
성격	7개 감정: 기쁨, 슬픔, 놀람, 분노, 공포, 혐오, 중립
출처	AI Hub, 한국어 단발성 대화 데이터
특징	기존 한국어 텍스트 데이터는 이진분류(감성) 수준을 크게 벗어나지 못하지만 본 데이터는 다중분류(감정)에 대한 데이터 특성을 가짐
데이터 수	38,594

<표 6>. KoBERT 학습 데이터



<그림 7>. 2단계: 감정분석
(Emotion Analysis)

4. 실험 및 실험결과: 2단계 데이터 표본 추출

④ 데이터 표본 추출: 데이터 희소성 문제를 해결하기 위한 과정

- 데이터 표본 추출(Sampling): **유저별 10개 이상** 리뷰 데이터
- 극단치 제거: **총평점 & 감성평점의 극단치** 322개 제거

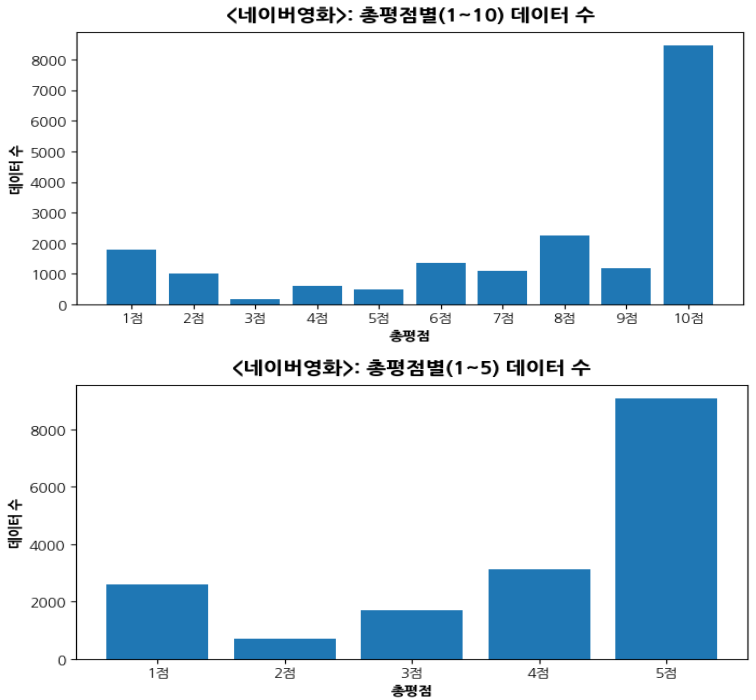
ex. 총평점(4점) & 감성평점(5점) → 유지
총평점(4점) & 감성평점(2점) → 제거

수집 기간			2022.02.14 ~ 2022.03.01(17일)		
리뷰 수			17,261		
영화 수			2,808		
고객 수			9,510		
총평점 (1~10)	1(1,788)	2(1,001)	3(165)	4(599)	5(480)
	6(1,356)	7(1,104)	8(2,245)	9(1,176)	10(8,476)
총평점(1~5)	1(2,612)	2(711)	3(1,706)	4(3,141)	5(9,091)

<표 7>. 영화 리뷰 데이터 수

구분	데이터 필터링	
	이전	이후
영화	2,808	**1,025**
고객	9,510	**165**
데이터 수	17,261	**2,245**

<표 8>. 영화 리뷰 데이터 수: 데이터 필터링



<그림 8>. 영화 리뷰: 총평점별 분포(범주화 전/후)

4. 실험 및 실험결과: 3단계 감정별 영화추천

- 감정별 영화 추천:
  - ① **속성가중치 반영한 총평점 기반 고객-영화 매트릭스**:
    ※ 기존의 실수 형태의 데이터를 정수 형태로 범주화
    - 고객-속성 가중치 행렬: 고객별 속성의 합이 0 혹은 1이 되도록 구성
    - 영화-속성 총평점 행렬: 영화별 속성의 합이 영화의 평균 총평점이 되도록 구성
  - ② **감정별 감성평점 기반 고객-영화 매트릭스**

  - ③ **평점별 가중치를 적용한 평점 구축:**
    두 가지 평점별 가중치를 **0.05 단위로 RMSE를 측정**했을 때
    최적의 성과를 도출하는 지점을 찾아서 해당 지점의 수치를 가중치로 반영
    ※ 속성가중치 총평점(0.05) & 감정평점(0.95)로 적용하여 새로운 평점 생성

  - ③ **감정별 속성가중치를 반영한 감성 & 총평점 기반 고객-영화 매트릭스**

  - ④ 개인화 추천시스템: 감정별 영화추천

4. 실험 및 실험결과: 제안 가능한 서비스 형태

※ 개인화 추천시스템: 감정별 영화 추천

- 추천대상(User): 8045
- 감정필터: **놀란**
- 시청하지 않은 영화 목록: 20개(총 영화목록 186개)
- 추천 영화(Movie): 605, 522, 453, 628, 659

User ID	Movie ID	평균값	예측값
User 8045	605	1.02	1.617
	522	1.19	1.593
	453	1.17	1.566
	628	1.07	1.542
	659	1.0	1.532

<표 9>. 개인화 추천시스템:
고객(8045)에 대한 영화 추천결과

<그림 9> 개인화 추천시스템: 감정별 영화 추천

5. 결론

[학술적 시사점]

① 정성적 데이터인 온라인 리뷰를 통해 **텍스트가 갖는 레스토랑 속성, 감정 및 감성을 정량적 데이터로 변환**하여 추천시스템에 적용함.

② 기존의 추천시스템 연구는 정량적 데이터만을 이용하지만
본 연구에서는 **정량적 데이터와 정성적 데이터를 함께 이용한 개인화 추천시스템**을 제안함.

③ 텍스트의 속성을 추출할 때 주로 활용되는 LDA는 짧은 리뷰에서 속성 분포를 추정하기 어렵다는 한계를 갖고 있어
본 연구에서는 **ABAE 모델을 적용하여 구체적인 속성을 추출**함.

[실무적 시사점]

① 정성적 데이터인 온라인 리뷰를 활용하면서 **데이터의 손실을 예방**했으며,
이를 활용하여 다양한 측면을 고려한 추천 결과를 제시함.

② **정성적 & 정량적 데이터를 함께 사용**하여 하나의 평점을 만들고 이를 속성 및 감정 값과 결합하여
하나의 개인화 추천시스템을 구현했을 때 **고객의 취향에 맞는 레스토랑 추천결과를 제시**함.

감사합니다