CLEANSED
FEEDBACK

소프트웨어융합 캡스톤 디자인
최종발표

경희대학교 소프트웨어융합학과
2020105742 한지훈

CONTENTS

01 문제 정의

개요

영화, 드라마, 만화 등 **예술 작품**은 감동과 즐거움을 전달하며 **관객과 소통하는 매개체**

관람 후 관객들은 댓글을 통해 **감상평과 감정**을 자유롭게 표현

이런 공간에는 작품과 작가를 향한 **욕설과 악성 댓글**도 존재

이러한 부정적인 표현은 창작자와 다른 관람객 모두에게 **마음의 상처**를 남기며, 건강한 소통 문화를 해침

관련 기사

암 치료 받는데 "죽어라" 악플… 눈물로 그리는 웹툰

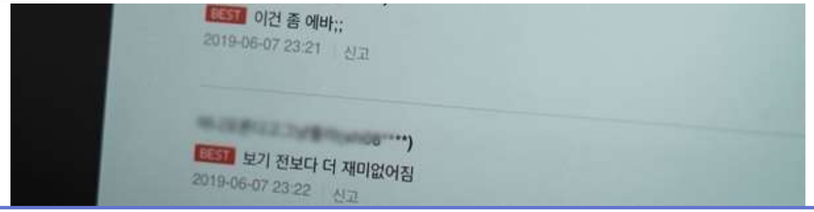
노동강도, 악성댓글 등 정신적 고통 크지만 지원은 미미

신지수 (clickjs)

Today pick

'악플 정글'에 방치된 웹툰… "연재를 중단합니다"

박소영 기자 구독 + 입력 2019.11.30 04:40 수정 2019.11.30 09:29 | 1면



02 기존 연구

A. 비난 분류/제거 연구

EX) K-HATERS: A Hate Speech Detection Corpus in Korean with Target-Specific Ratings

Data example
일본은 접종비용을 내나보네 I guess one should pay for getting vaccinated in Japan.
전세계 어디든 무개념 틀딱들이 젤 문제 Anywhere in the world, idealess dotard is the biggest problem.
기레기야! 백신을 남을 위해서 맞냐? Hey such a presstitute ! Are you getting a vaccine for someone else?

192,158개 한국어 뉴스 댓글에 대해 3단계 혐오 레이블을 적용한 대규모 Corpus 'K-HATERS' 구축

BERT 기반 모델을 통해 정치·종교 등 특정 타겟에 대한 혐오 표현을 효과적으로 분류·제거, implicit hate도 탐지 가능

B. 비난 순화 연구

EX) 거대언어모델과 설명가능 인공지능을 활용한 한국어 유해 댓글 순화 연구

	예시 1	예시 2
Raw Data	저 [redacted] [redacted] 는 좀 지워주세요	와 컨셉아니면 [redacted] 해라 [redacted] 야 [redacted] ㅋㅋㅋ 이게 [redacted] 냐
Our method	**저 분 얼굴은 좀 지워주세요**	**와, 컨셉이 아니면 정말 심각하네요. 이게 정말 사람의 행동인가요?**

KcELECTRA 분류기로 유해 댓글을 탐지한 후, XAI(SHAP) 기법을 사용해 유해성에 기여한 토큰 마스킹

원본 문장 + 마스킹된 문장 + few-shot 예시 프롬프트를 LLM에 입력하여 의미 유지, 유해 표현 순화한 문장 생성

03 기존 연구 대비 차별점

기존 연구와 비교	A. 비난 분류/제거 연구	B. 비난 순화 연구	C. 본인 방식
중립 댓글 Ex) “재밌게 보고 있어요”		원본 제공	
순수 피드백 댓글 Ex) “전개 속도가 좀 빨라요”		원본 제공	
순수 비난 댓글 Ex) “XX 재미없네”	제거	순화 의미 X Ex) “진짜 재미가 없네요”	제거
비난 섞인 피드백 댓글 Ex) “맞춤법 검사기도 안돌리나 ㅋㅋ”	피드백도 사라짐	순화해서 제공 Ex) “맞춤법에 유의해야 할거 같아요”	순화해서 제공 Ex) “맞춤법에 유의해야 할거 같아요”

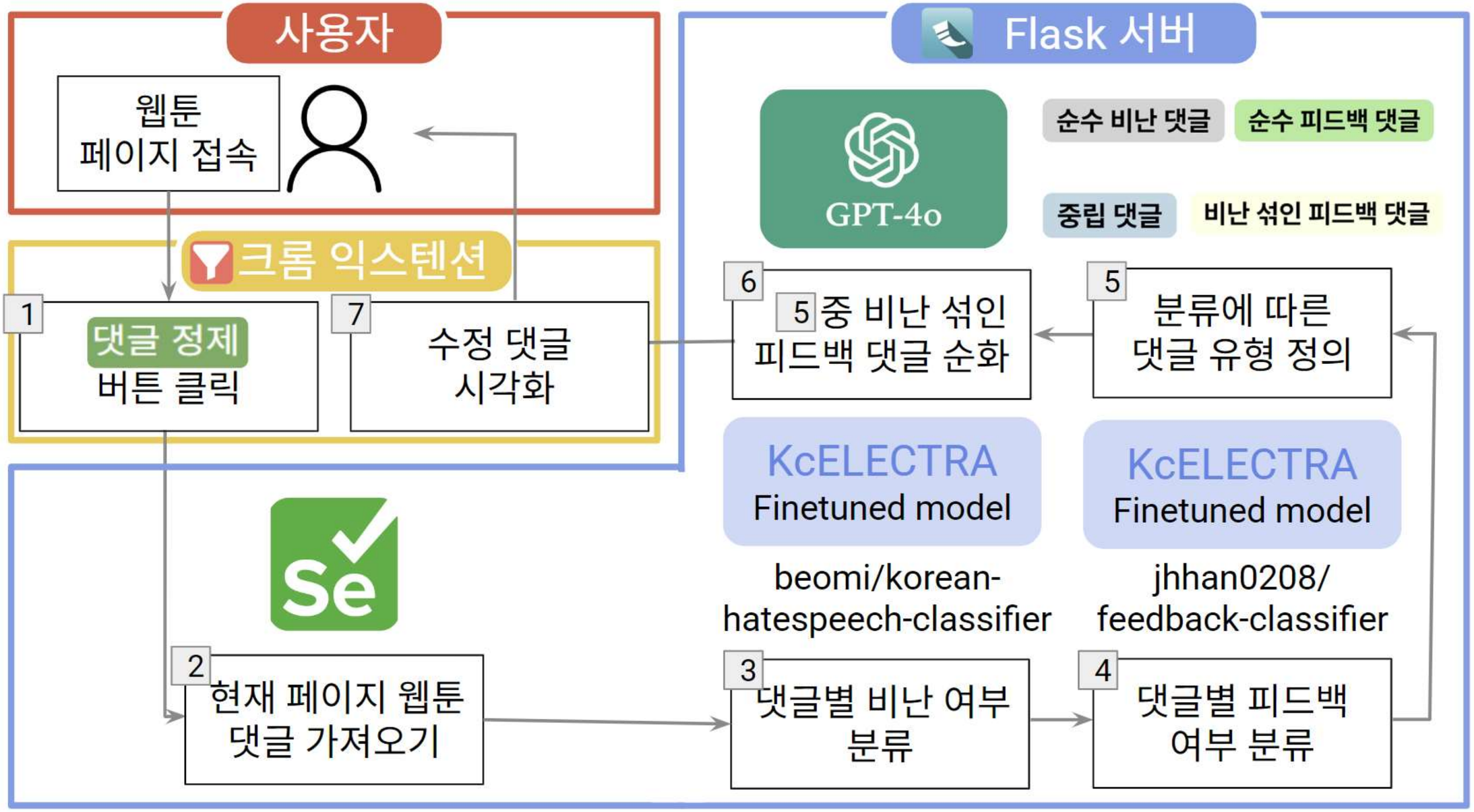
댓글들을 4가지 유형으로 분류

정말 비난만 있는 경우에는 기존의 비난 분류/제거 연구처럼 제거

단 비난이 존재하는 경우에도 피드백이 포함되어 있다면 비난은 순화, 피드백은 살려서 제공

>>> 비난으로 인한 마음의 상처 최소화, 피드백은 받아들여 더 좋은 작품 만드는 데에 기여

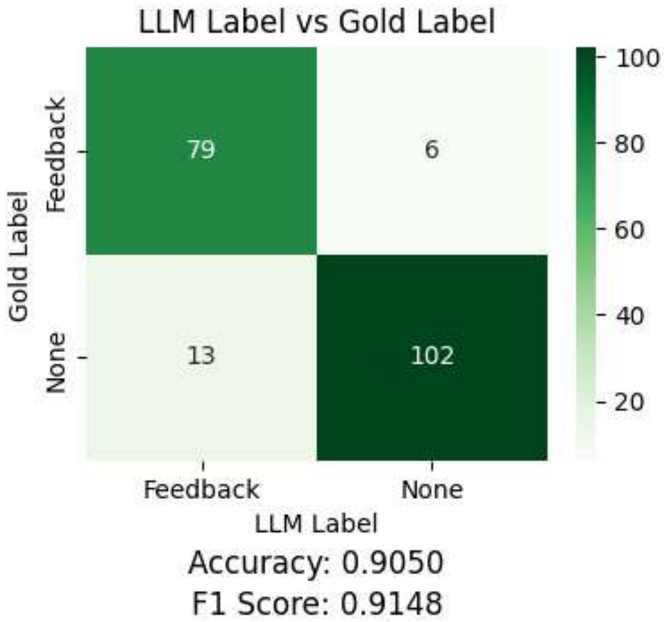
04 파이프라인


사용자
웹툰
페이지 접속
크롬 익스텐션
1
댓글 정제
버튼 클릭
7
수정 댓글
시각화
Flask 서버
GPT-4o
순수 비난 댓글
순수 피드백 댓글
중립 댓글
비난 섞인 피드백 댓글
6
5 중 비난 섞인
피드백 댓글 순화
5
분류에 따른
댓글 유형 정의
KcELECTRA
Finetuned model
beomi/korean-
hatespeech-classifier
KcELECTRA
Finetuned model
jhhan0208/
feedback-classifier
Se
2
현재 페이지 웹툰
댓글 가져오기
3
댓글별 비난 여부
분류
4
댓글별 피드백
여부 분류

05 모델 성능 평가

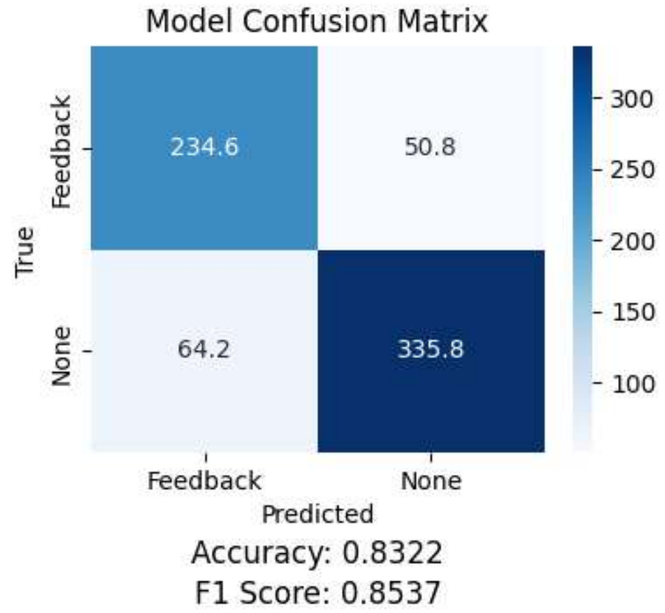
라벨링 성능 평가

1. 라벨링된 댓글들 중 200개 랜덤 샘플링
2. 라벨을 가리고 사람(본인, 동료)이 직접 댓글의 피드백 여부를 판단
3. 본인, 동료 불일치 라벨은 상의해서 결정, Gold Label 생성
4. Gold Label과 LLM Label 얼마나 일치하는지 확인

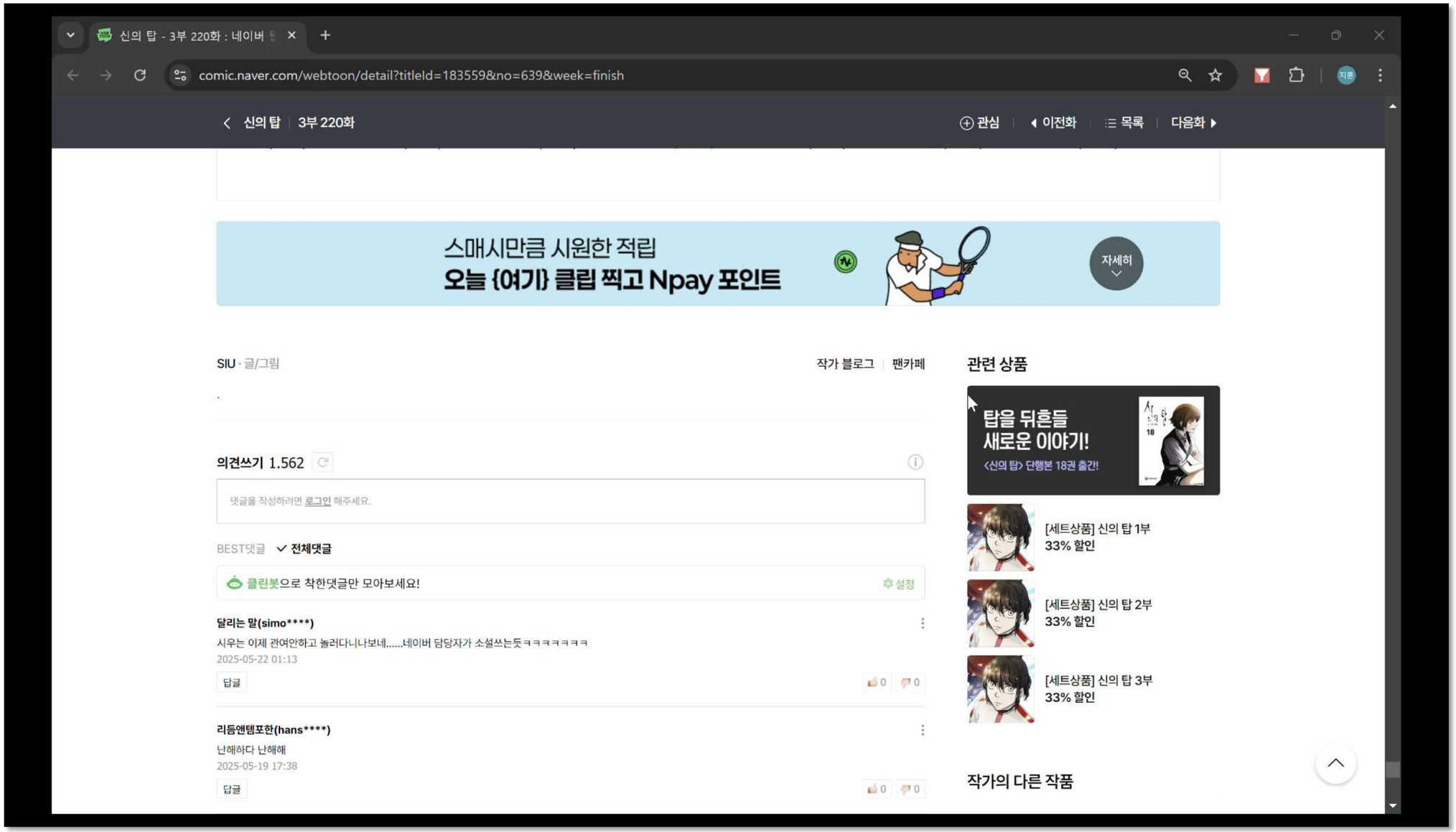


피드백 분류 성능 평가

1. 전체 댓글을 80%의 Train set(2742개)와 20%의 Test set(685개)로 분할
2. 5-Fold Cross Validation 진행
3. 5개 Fold의 평균 지표(Accuracy, F1 Score)로 피드백 분류 모델 성능 확인



06 데모 영상

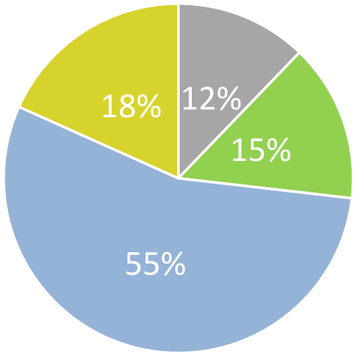


CLEANSED FEEDBACK

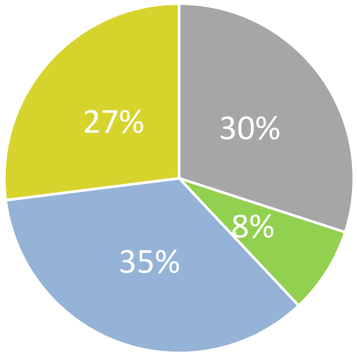
07 결과 댓글 분포 확인

여러 평점대에 대한 댓글 유형별 비중

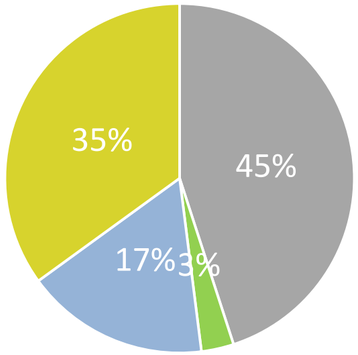
★ 10.0 ~ 8.5



★ 8.5 ~ 7.0



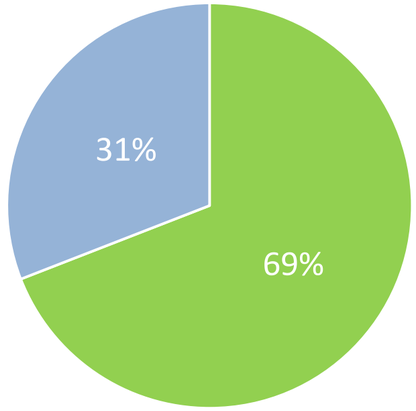
★ 7.0 ~ 5.5



- 순수 비난
- 순수 피드백
- 중립
- 비난 섞인 피드백

**평점이 낮은 웹툰일수록 순수 피드백의 비중은 줄고,
비난 섞인 피드백의 비중이 늘어남을 알 수 있음**

CLEANSED FEEDBACK 적용



- 순수 피드백
- 중립

**서비스를 통해 순수 비난은 제거,
비난 섞인 피드백은
순수 피드백으로 변환**

감사합니다