Filtering and Summarization of Profanity and Hate Speech in Steam Platform Review Data

Do-Young Kim
Computer Engineering department
Chungbuk National University
Cheongju, South Korea
jjyo0108@chungbuk.ac.kr

Young-Seob Jeong*
Computer Engineering department
Chungbuk National University
Chengju, South Korea
ysjay@chungbuk.ac.kr

***Abstract*—최근 Steam 플랫폼에서 제공되는 게임 리뷰는 소비자들의 구매 결정에 중요한 역할을 하고 있다. 그러나 욕설 및 증오 표현이 포함된 리뷰는 여전히 문제로 남아있다. 본 연구는 한국어 리뷰 데이터를 대상으로 욕설 및 혐오 표현을 효과적으로 필터링하고, 리뷰를 긍정 및 부정으로 나누어 요약하는 서비스를 제안한다. 이를 위해 KcELECTRA 모델을 활용하여 구어체와 비공식적 표현을 효과적으로 처리하며, 한국어 리뷰 데이터를 분석하여 사용자들에게 유용한 정보를 제공한다.**

***Keywords*-Steam; KcELECTRA-small; 게임 리뷰; 욕설; 혐오표현; 필터링; 한국어; 리뷰 요약;**

I. Introduction

최근 몇 년간 Steam은 전 세계적으로 가장 성공적인 게임 유통 플랫폼 중 하나로 자리 잡았으며, 그 사용자 수는 급격히 증가하고 있다. 올해 9월, 동시 접속자 수가 38,367,277명으로 역대 최고치에 달한 기록은 게임 산업 내에서 Steam의 막대한 영향력을 보여준다 [5].

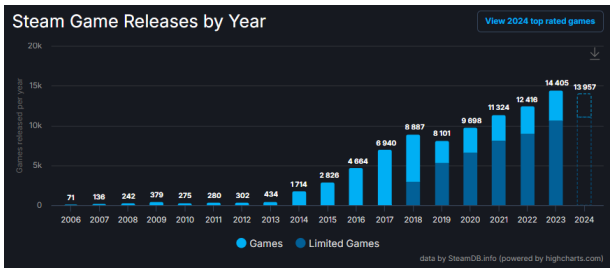

Figure 1. Number of Steam Game Releases by Year

Steam에는 Fig. 1 화면처럼 수많은 게임이 매년 플랫폼에 추가되는데, 이러한 게임들을 구매할 때 소비자들이 많이 참고하는 요소 중 하나는 바로 사용자 리뷰다. 이번 실험을 위한 설문 조사에서도 72.7%가 주로 구매 의사 결정을 위해 게임 리뷰를 읽는다고 답했다.

Steam은 현재 리뷰의 유용성에 따라 도움이 되는 리뷰를 최우선으로 보여주는 옵션을 제공하고 있으며, 유용성은 리뷰의 길이, 플레이 시간, 긍정 및 부정 감정 표현 비율 등을 기준으로 평가된다 [2]. 이 시스템은 소비자가 더 양질의 리뷰를 접하는 데 기여하고 있지만, 욕설과 증오 및 혐오 표현은 여전히 필터링되지 않고 리뷰창에 나타나곤 한다.

적당한 욕설은 구매에 긍정적인 영향을 끼칠 수도 있다고 밝혀진 바 있으나 [4], 온라인상의 증오 발언에 노출되면 우울증과 불안, 자기 의심 및 자신감 수준 등 이용자의 정신적 건강에 악영향을 줄 수 있다는 연구가 보고되고 있다 [1]. 또한, 설문 결과에 따르면, 게임 리뷰에서 욕설과 혐오 표현을 발견했을 때 응답자의 54.6%가 불쾌감을 느꼈으며, 이 중 33.3%는 매우 강한 불쾌감을 경험했다고 응답했다.

이에 따라 본 연구는 욕설과 증오 및 혐오 표현을 필터링하면서도 소비자들이 게임 구매에 참고할 수 있는 유용한 리뷰를 요약 제공하는 서비스를 제안한다.

항목	원천 데이터
review	'배틀그라운드 너무 재밌어요'
voted_up	True

Table I
An Example of Source Data

연구에서 다루는 원천 데이터는 한국어로 작성된 리뷰 텍스트(문자열)와 해당 리뷰어가 선택한 긍정 또는 부정 라벨(True, False)을 열로 가진 데이터 프레임이며, 이 텍스트는 문장 단위로 나누어지지 않은 원본 데이터를 사용한다.

항목	출력 데이터
positive_summary	'진짜 훌륭한 게임이다.'
negative_summary	'계속 하려면 DLC를 반드시 구매해야 한다.'

Table II
An Example of Output Data

연구의 출력 데이터는 위와 같은 게임의 각 긍정 또는 부정적인 리뷰의 요약문(문자열)이다. 최종 결과물은 GPT API를 통해 생성되며, 이때 출력의 최대 토큰 수를 나타내는 max_tokens 변수는 150으로 설정하였다. 토큰은 텍스트를 구성하는 기본 단위로, 단어의 조각, 공백, 구두점 등이 포함된다. 예를 들어, "Hello, world!"는 3개의 토큰으로 나뉜다: "Hello", ",", "world".

여기서 원천 데이터의 'review'는 해당 플랫폼 리뷰 데이터의 특성상 구어체, 비정형 텍스트, 그리고 소셜 미디어에서 흔히 볼 수 있는 표현들이 다수 포함될 가능성이 높다. 이러한 비정형 데이터를 효과적으로 처리하기 위해 경량

모델로 KcELECTRA-small을 선정하였다. KcELECTRA-small은 한국어에 특화된 사전 학습 언어 모델로, 일반적인 다국어 모델보다 한국어 데이터의 문맥적 이해와 세부적인 표현 처리에서 우수한 성능을 보여준다.

본 연구에서는 Python을 주요 프로그래밍 언어로 사용하며, 데이터 수집을 위해 Steam에서 제공하는 공식 API를 활용한다. 욕설 및 혐오 표현 필터링에서는 직접 입력한 txt 파일 형태의 욕설 및 혐오 표현 데이터셋, K-MHaS 한국어 혐오 표현 데이터셋을 사용한다. 요약 태스크에서는 pororo 라이브러리와 GPT API를 이용해서 요약하는 과정을 거쳐 최종적인 리뷰 요약본을 제공한다.

II. Related Works

기존 연구 중 "Sentiment Analysis of Game Reviews on STEAM using BERT, BiLSTM, and CRF" [3]는 BERT와 BiLSTM을 활용하여 Steam 게임 리뷰에서 감정을 분석한 연구로, 본 연구와 유사하게 BERT 모델을 활용하여 감성 분석을 다루고 있다. 하지만 이 연구는 영어 리뷰에 중점을 두고 있어, 한국어 리뷰를 대상으로 한 연구는 거의 이루어지지 않았다.

또 다른 연구로는 "Aspect-Based Sentiment Analysis of User Created Game Reviews" [6]가 있으며, 이 연구는 게임 리뷰에서 속성 기반 감성 분석을 통해 게임의 다양한 측면을 평가하였다. 그러나 이 연구 역시 영어 리뷰를 대상으로 한 것이 대부분이다. 한국어 리뷰 데이터를 활용한 본 연구는 한국어 사용자들에게 보다 적합한 리뷰 시스템을 제안한다는 점에서 기존 연구와 차별화된다.

기존 제품으로는 Steam Data Suite가 존재하는데, 이는 게임 개발사와 퍼블리셔를 대상으로 하는 분석 도구로, 주로 마케팅 전략을 지원하는 데 중점을 두고 있다. 이 서비스는 사용자 리뷰와 판매 데이터 등을 분석하여 개발사들이 게임의 성과와 시장에서의 반응을 더 잘 이해할 수 있도록 도와준다. 하지만, Steam Data Suite는 기업 회원 전용 서비스로 제공되며, 일반 게임 유저들이 직접 사용할 수 있는 도구는 아니다.

본 연구는 아직 한국어로 작성된 Steam 게임 리뷰를 대상으로 한 연구나 서비스가 부족한 상황에서, 한국어 사용자 리뷰를 필터링하고 요약하는 시스템을 제안한다. 이를 통해 한국어 사용자의 게임 구매 경험을 개선하고, 게임 개발사에게는 중요한 인사이트를 제공할 수 있을 것이다.

III. Method

A. Data Collection and Preprocessing

앞서 언급했듯이, 원천 데이터는 Steam API를 이용하여 수집한다. 해당 API는 특정 게임에 대해 APP ID를 제공하면, 그 게임에 대한 리뷰 데이터를 제공하는 형식이다. 이때 반환되는 속성은 최대 17개이나 이 속성을 이용해 예측하는 태스크가 없기 때문에 긍정/부정 분류에 필요한 voted_up 항목과 결과물로 활용할 review 항목만 남기고 나머지는 모두 제거한다. 수집된 데이터는 분석을 위해 여러 단계의 전처리 과정을 거치게 된다. 전처리 단계에서는 외국어, 불필요한 링크 및 해시태그, 일부 제외 이모지 및 기호, null 값 및 빈 문자열, 불용어 제거 등의 작업을 수행하여 데이터를 정제한다. 이러한 과정을 통해 데이터의 품질을 높이고, 모델 학습에 적합한 형태로 변환한다.

B. Model

전처리된 데이터는 이후 욕설 데이터셋, 정규식, 혐오 표현 예측 모델을 이용한 필터링을 통해 욕설 및 혐오 표현이 포함된 리뷰를 제거한다.

여기서 혐오 표현 예측 모델로 KcELECTRA-small-v2022를 사용하며, 파인 튜닝을 위해 K-MHaS 한국어 혐오 표현 데이터셋을 학습시켰다. 모델의 Training Arguments는 대표적으로 몇 가지가 다음과 같이 설정되었다: learning_rate=5e-5, num_train_epochs=3, evaluation_strategy="epoch"

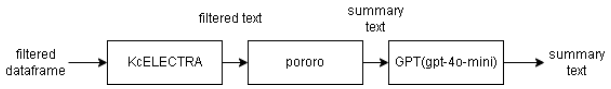

Figure 2. Structure of this Service

본 연구에서 구성한 전체적인 파이프 라인은 Fig. 2과 같다. 전처리와 정규식을 활용한 필터링을 거쳐 생성된 데이터프레임의 리뷰 텍스트는 KcELECTRA 모델을 사용해 욕설이나 혐오 표현의 포함 여부를 판별해 라벨로 도출한다. 욕설 또는 혐오 표현을 포함하고 있다고 판단된 레코드는 제거되며, 제거되지 않은 리뷰들은 join 연산을 통해 하나의 문장으로 구성된다. 이 문장은 매개 변수 final_length를 300으로 설정된 pororo 라이브러리를 통해 300자 이하의 문장으로 요약된다. 다만 전체 리뷰 텍스트가 300자를 넘지 않을 경우, 요약 과정을 거치지 않고 그대로 반환된다. 그리고 요약된 문장은 소셜 미디어 특유의 표현을 포함하고, 맞춤법이 맞지 않을 가능성이 높으므로 GPT API를 이용해 문장을 자연스럽게 다듬고 어미를 통일한다.

IV. Experiments

A. Data

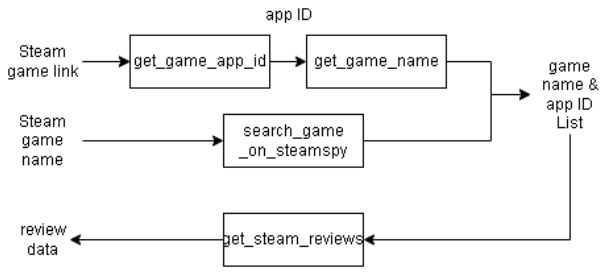

Figure 3. Structure of Data Pipeline

데이터를 받아오기 위한 과정은 Fig. 3 그림과 같다. 사용자가 사이트 검색창에 Steam 게임 상점 페이지 URL을 입력하면, get_game_app_id 함수가 해당 게임의 app ID를 가져온다. 이후, 이 app ID를 get_game_name 함수에 전달하면, Steam API를 통해 해당 게임의 이름과 app ID를

반환된다. URL 입력의 경우 하나의 게임으로 특정되므로 결과는 하나만 존재한다.

키워드를 입력할 경우에는 search_game_on_steamspy 함수가 비공식 API인 SteamSpy API를 통해 해당 키워드가 포함된 모든 게임의 app ID와 게임명을 리스트로 반환하게 된다. Steam 공식 API에서는 이러한 기능을 제공하지 않기 때문에, 두 가지 방법을 병행하여 사용했다.

B. Experimental Method

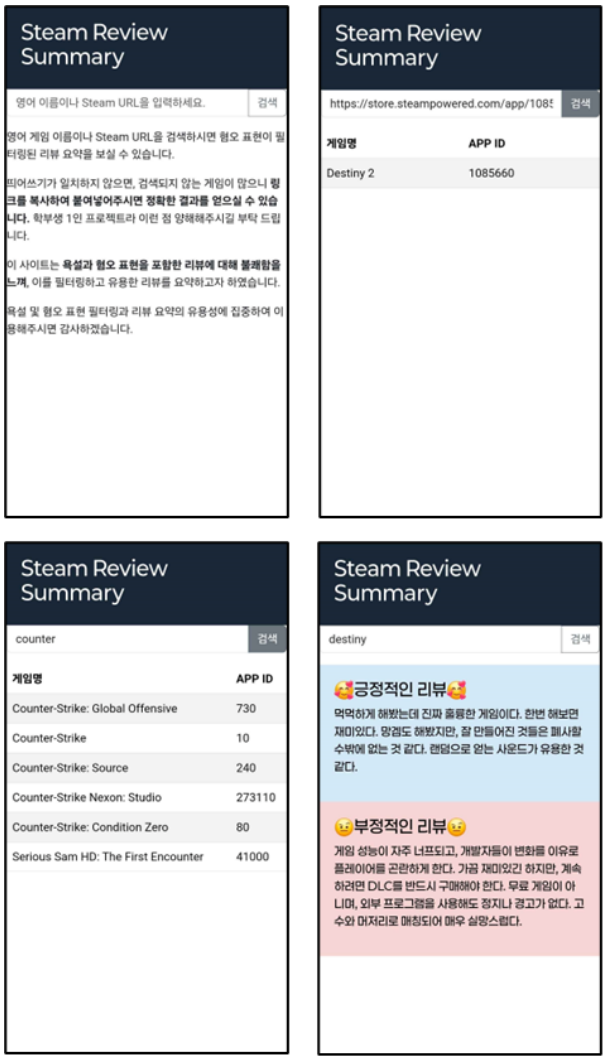

Figure 4. Steam Review Summary Homepage

실험은 웹사이트를 통해 이루어진다. 사용자가 검색창에 Steam URL이나 게임명과 관련된 키워드를 입력하면, 해당되는 게임의 리스트가 반환된다. 그 리스트 중 원하는 게임을 선택할 시 선택된 게임의 app ID가 API에 전달되어 리뷰 데이터를 가져오게 된다. 이 리뷰 데이터는 전처리 과정을 거쳐 욕설과 혐오 표현이 필터링되며, pororo 라이브러리가 필터링된 모든 리뷰 텍스트를 합쳐 300자 이하로 요약한다. 마지막으로 GPT API를 통해 300자 이하의 요약문이 자연스럽게 다듬어지고 어미가 통일된 상태로 반환되어 프론트엔드에 보여진다.

C. Result

이 실험은 두 가지 측면에서 결과를 고려할 수 있다. 첫 번째는 욕설과 혐오 표현의 배제의 관점이고, 두 번째는 리뷰 요약의 관점이다. 결론을 도출하기 위해, 해당 사이트를 이용한 사용자들을 대상으로 설문을 진행하였다.

먼저 첫 번째 측면은 설문에서 '리뷰 요약에서 필터링된 표현(혐오 표현 및 욕설)의 배제가 긍정적으로 느껴지셨습니까?'라는 항목을 통해 측정하였다. 응답자의 약 82%가 필터링된 표현의 배제가 긍정적이었다고 답했다. 주요 이유로는 욕설 및 혐오 표현을 직접 읽지 않아도 되어 좋다는 의견이 가장 많았으며, 비난과 같이 편파적이고 감정적인 리뷰보다 더 객관적인 리뷰를 읽을 수 있어 좋다는 의견이 그 뒤를 이었다.

두 번째 측면인 요약 관점에서는 4개의 질문과 1개의 후속 질문을 통해 결과를 수집했다. '제공된 리뷰 요약은 이해하기 쉬웠습니까?'라는 질문에서 36.4%는 매우 그렇다, 27.3%는 그렇다, 9.1%는 보통, 27.3%는 그렇지 않다고 답변하였다. 그렇지 않다고 답한 응답자를 대상으로 어떤 점이 이해하기 어려웠는지 묻자, 문장의 완성도가 사람이 작성한 것보다 떨어진다고 느꼈으며, 특정 형태소가 반복되는 문제로 인해 문장이 잘 이해되지 않는 경우가 있었다는 답변이 주를 이뤘다.

'제공된 긍정 또는 부정 리뷰 요약이 실제 사용자 리뷰와 관련이 있다고 느끼셨습니까?'라는 질문에서는 72.7%가 그렇다, 9.1%가 그렇지 않다고 대답했다. '리뷰 요약 결과물이 게임 구매 또는 플레이 의사 결정에 도움이 되었습니까?'라는 질문에는 90.9%가 그렇다고 답변했다. 이 중 부정적인 리뷰 요약이 더 유용하다는 의견이 많았는데, 이는 악의적인 비판이 필터링되어 게임의 단점이 잘 드러난다는 이유에서였다. 반면, 긍정적인 리뷰 요약이 도움이 된다고 답한 경우는 내용이 구체적이어서 좋았다는 의견이 있었다. 또한, 요약문의 신뢰도와 관련해서는 54.5%가 신뢰할 수 있다고 답변했으며, 45.5%가 보통이라고 답변했다.

마지막으로, 본 서비스와 같은 요약형 리뷰 플랫폼이 전체 게임 시장이나 커뮤니티에 긍정적인 영향을 미칠 것이라고 생각하냐는 질문에 90.9%가 그렇다고 답했다. 앞으로 해당 플랫폼을 게임 정보를 확인하는 데 사용할 의향이 있냐는 질문에도 72.7%가 그렇다고 답했다.

V. Conclusion

A. Limitations

이 연구의 한계점으로는 크게 네 가지 항목이 있다. 첫 번째, pororo 라이브러리를 통한 요약 과정에서 입력 데이터를 하나의 텍스트로 합쳤으며, 이를 1024자씩 끊어 요약하였다. 개발 당시에는 입력 데이터 크기가 이렇게까지 클 줄 모르고 summarizer 함수를 사용해 한 번에 요약하려 했으나, 오류와 성능 이슈로 인해 1024자 단위로 나누어 처리할 수밖에 없었다. 이로 인해 문맥이 끊어지는 문제가 발생하여 요약의 일관성이 저하되는 경향이 있었다.

두 번째, DB가 존재하지 않아 리뷰 요약 시마다 매번 API 호출과 요약 태스크 스크립트 실행이 이루어져 로딩 시간이 길다. 이 문제로 인해 게임 검색 후 리뷰문을 보기까지 약 1분 이상이 소요되었으며, 이는 사용자들이 가장 큰 불편함을 느낀 부분이었다.

세 번째, 사이트의 키워드 검색 기능이 제한적으로 동작하여 사용자에게 불편함을 줄 수 있다. 키워드 검색에서 사설 API를 사용하다 보니, Steam 사이트에 존재하는 게임이 검색되지 않는 문제가 발생했다. 이를 보완하기 위해 링크 검색 기능을 추가했으나, 두 가지 방법이 동일한 결과를 반환하지 않는 데서 오는 근본적인 불편함이 있었다. 또한, 사용자 설문 결과에 따르면, 검색창에서 오타가 발생해도 자동으로 수정하여 예측된 결과를 보여주는 기능이 필요하다는 의견이 있었다. 따라서 검색 환경 개선이 필요한 상황이다.

네 번째, 긍정적인 리뷰에도 일부 부정적인 내용이 포함되는 경우가 있다. 처음 연구를 시작할 때는 리뷰 감성 예측 기능을 추가할 계획이었으나, Steam API를 통해 제공되는 데이터에 이미 사용자가 남긴 긍정 또는 부정 평가가 포함되어 있어 해당 기능을 제외하였다. 그러나 본 연구에서 한 가지 간과한 사실이 있다면, 긍정 리뷰가 부정적인 이야기를 작성한 후 '그럼에도 불구하고' 좋았다는 의견을 표명할 수 있다는 것이다.

B. Future Works

앞서 한계점에서 언급한 내용을 바탕으로, 필요한 작업은 다음과 같다. 먼저, 요약 시 하나의 텍스트를 요약하는 기존 방식 대신, 1024와 같은 특정 글자 수를 기준으로 처리하는 방식으로 수정해야 한다. 구체적으로, 해당 기준 글자 수를 초과하는 리뷰는 그대로 요약하고, 기준을 넘지 않는 리뷰는 다음 리뷰와 합쳐 최소 기준 글자 수를 만족하도록 만든 후 반복 요약하는 과정으로 개선할 필요가 있다. 또한, 파인 튜닝 없이 한국어 리뷰를 요약했을 때 가장 성능이 좋았던 pororo 라이브러리 대신, 더 많은 데이터를 확보하여 다른 모델을 사용해보는 것도 고려할 수 있다.

다음으로, 긍정적인 리뷰에 부정적인 내용이 포함될 수 있는 문제를 해결하기 위해, 감성 분석 모델을 활용하여 긍정 및 부정 라벨링을 새로 수행하고 요약하는 방법이 필요하다. 이를 통해 리뷰를 보다 중립적으로 구분하여 요약의 품질을 향상시킬 수 있을 것이다. 추가적으로, 리뷰에 대한 감성 분석 연구를 추후 진행한다면, 카테고리별로 더 구체적이고 세분화된 요약을 생성할 수 있을 것이라고 판단된다.

마지막으로, 개발한 사이트에 DB를 생성하여 사용자 경험(UX)을 개선할 필요가 있다. 현재 사이트는 게임 키워드 검색 시 해당 키워드와 일치도가 100%인 게임만 검색할 수 있는 제한점과, 리뷰 요약문 반환 시 로딩 시간이 1분 이상 소요되는 문제를 가지고 있다. 이를 해결하기 위해, Airflow와 같은 도구를 활용한 배치 프로세스를 도입하여 리뷰 데이터를 특정 시간마다 로드하고, 전처리, 필터링, 요약 작업을 수행한 결과를 DB에 저장하여 사용자에게 제공한다면 UX가 크게 개선될 것이라고 생각된다.

ACKNOWLEDGMENT

TBD

REFERENCES

[1] Arne Dreißigacker, Philipp Müller, Anna Isenhardt, and Jonas Schemmel. Online hate speech victimization: Consequences for victims' feelings of insecurity. *Crime Science*, 13(4), 2024.

[2] Lukas Eberhard, Philipp Koncar, Patrick Kasper, and Christian Gutl. Investigating helpfulness of video game reviews on the steam platform. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 43–45. IEEE, 2018.

[3] Jalaluddin Al Mursyidy Fadhlurrahman, Neng Ayu Herawati, Hayyu Rachma Widya Aulya, Ira Puspasari, and Nugraha Priya Utama. Sentiment analysis of game reviews on steam using bert, bilstm, and crf. In *2023 International Conference on Electrical Engineering and Informatics (ICEEI)*. IEEE, 2023.

[4] Katherine C. Lafreniere, Sarah G. Moore, and Robert J. Fisher. The power of profanity: The meaning and impact of swear words in word of mouth. *Journal of Marketing Research*, 59(5):908–925, 2022.

[5] Shubhankar Parijat. Steam sees new all-time concurrent users peak at over 38 million, September 2024. Accessed: 2024-09-23.

[6] Ian Michael Urriza and Maria Art Antonette Clariño. Aspect-based sentiment analysis of user created game reviews. In *2021 24th Conference of the Oriental COCOSDA (O-COCOSDA)*. IEEE, 2021.