야구 댓글 데이터 기반 감정 분석 및 요약 모델 개발

Sentiment Analysis and Summarization of Baseball Game Comments

김민하[1], 문가원[2], 박은비[2], 장민주[3], 김동하[2]

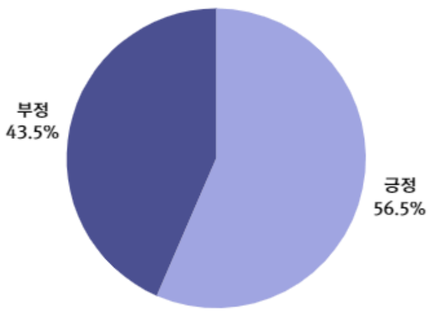
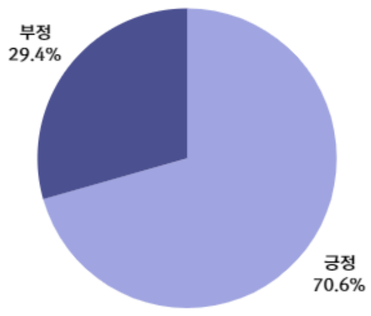
[1] 성신여자대학교 AI융합학부 [2] 성신여자대학교 수리통계데이터사이언스학부 [3] 성신여자대학교 법학부

◆ 연구배경 및 목표

- 최근 스포츠 산업은 단순한 경기 결과 중심에서 벗어나, 팬들과의 감정적 소통을 중시하는 경험 중심 산업으로 변화 중임. 팬들은 유튜브 등 SNS 댓글을 통해 실시간으로 감정과 인식을 표현하며, 이는 선수 및 팀의 이미지 형성에 직·간접적 영향을 미침
- 2025년 KBO 리그는 경기당 평균 관중 15,100명을 기록하며 팬덤 지속 확장 중, KBO 유튜브 채널은 누적 2.65억 조회수와 33만 구독자를 확보
- 그럼에도 불구하고, 이러한 비정형 감성 데이터를 정량적으로 분석·활용하려는 체계적인 연구는 부족한 실정

본 연구는 한국어 기반 감정 분석 요약 기술을 실제 적용하여 팬 의견을 데이터화하고, 실시간 여론 분석 및 전략 수립에 활용 가능한 AI 기반 KBO 리그 팬덤 분석 솔루션을 제시하는 것을 목표로 함.

◆ 분석 방법

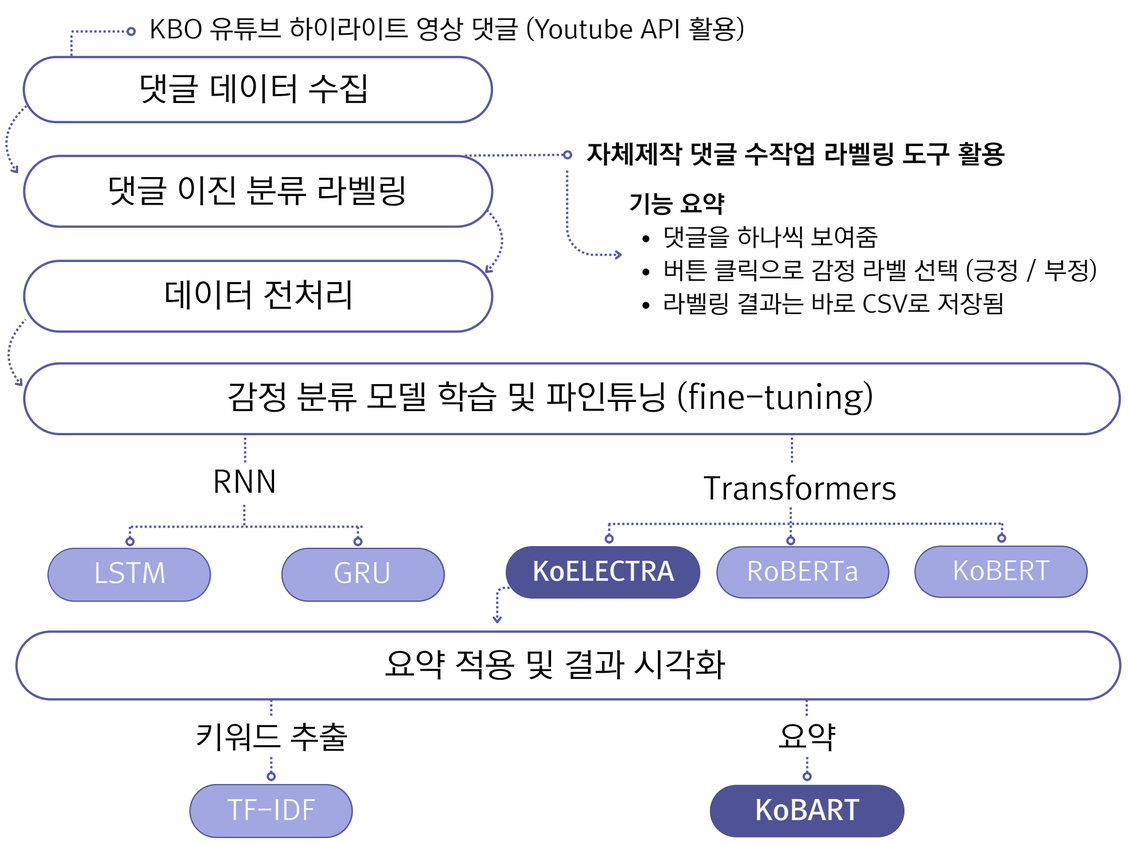
	학습 데이터	테스트 데이터
기간	2025년 3-4월 하루 한 경기 임의 선정	2025년 5월 17일 삼성 vs 롯데 경기
분석 대상	KBO 리그 유튜브 하이라이트 영상 댓글	
수집 방법	YouTube Data API	
데이터 라벨링	Streamlit 기반 웹 인터페이스 자체 제작 및 이진 감정 수작업 라벨링	
데이터 수	4,440개	500개
데이터셋 구성		

수집된 댓글은 중복 문장, 이모지, 특수문자 등을 제거하고, 모델 유형에 따라 전처리 방식을 달리 적용함.

- LSTM 및 GRU 모델에는 댓글에서 의미 있는 형태소만 추출한 후, 이를 토큰 인덱스로 변환하고 고정된 길이로 패딩 처리하여 시퀀스 형태로 학습 데이터를 구성함.
- Transformers 기반 모델에는 각 모델에 특화된 토크나이저(tokenizer)를 활용하여 문장을 토큰화하고, attention mask 등을 함께 생성해 입력함.

이후, 감정 분류를 위한 텍스트 분류 모델을 설계하고 비교·학습하였으며, 분류 결과를 바탕으로 TF-IDF 기반 키워드 추출과 KoBART 기반 문장 요약을 수행하여 주요 팬 의견을 요약함.

- **Data Analysis Pipeline**



◆ 분석 결과

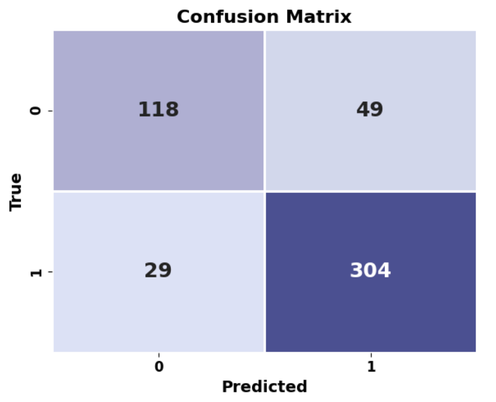
단위 : %

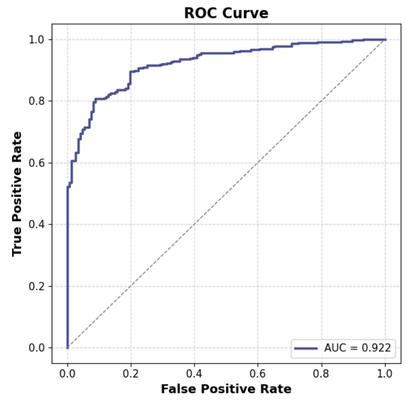
Model	LSTM	GRU	KoBERT	RoBERTa	KoELECTRA
Accuracy	72.0	70.6	78.8	83.4	85.4
F1-score	83.2	82.6	84.0	87.3	89.5

- 테스트 데이터의 감정 라벨 불균형을 고려하여, 정확도(Accuracy)보다 F1-score를 중심으로 모델 성능을 평가함.
- KoELECTRA는 정확도 85.4%, F1-score 89.5%로 가장 우수한 성능을 보여, 본 연구의 최종 분류 모델로 선정됨. (Transformer 기반의 contextual embedding이 문맥을 잘 반영)

혼동행렬 기반 분석

- 긍정 댓글은 정분류율이 높았으나, 부정 댓글은 긍정으로 오분류되는 경향이 나타남.





ROC Curve 기반 분석

- AUC 값 약 0.92로, 임계값 변화에도 안정적인 분류 성능 확인됨.

오분류 사례에 대해 K-means 클러스터링을 적용한 결과, 모델이 부정 감정을 긍정으로 오인하는 주요 원인은 긍정과 부정 표현이 혼재된 문장 구조 때문인 것으로 분석됨. 예를 들어, “오늘은 잘했지만 여전히 내야가 불안하다”와 같은 문장은 이중적인 감정을 포함하고 있어 모델에게 혼란을 주는 대표적인 사례였음.

- **KoBART 요약 결과**

1차전 고승민 때문에 이겼다 2차전 전민재 돌아왔구나 최준용이 너무 기다렸다 너무 기부니 좋다 진짜 김태형 감독님은 찐 명장이었다 코치진 꾸리고 선수들이 성장했다 롯데

뭐 하냐 결승점을 저 따위로 내주냐 기강 좀 잡아라 1305 이건 뭔 송구냐 ㅋㅋ 타격도 못해 수비도 못해 수비도 못해 손호영 왜 계속 쓰는거고 정훈 오늘 좀

- **TF-IDF 키워드 추출 결과**

긍정	부정
전민재, 롯데, 민재, 진짜, 좋다, 최준용, 오늘, 이기다, 김원중, 수비, 돌아오다, 사랑, 선수, 보다, 야구, 우리, 유격수, 복덩이	손호영, 삼성, 군, 좀, 진짜, 없다, 가다, 경기, 감독, 보다, 수비, 아니다, 같다, 왜, 박진만, 있다, 송구

TF-IDF는 어휘 빈도 기반으로 키워드를 빠르고 직관적으로 추출할 수 있으나, 의미 없는 단어가 포함되거나 문맥이 반영되지 않는 한계가 있음.
KoBART는 문맥과 의미를 반영하여 경기 내용과의 일치도가 높고, 팬 반응을 간결하고 자연스럽게 요약하는 데 강점을 보임.

◆ 기대효과 및 향후 과제

실용적 기대효과

- 본 연구를 활용하여 팬 댓글을 실시간으로 분석함으로써, 이슈 발생 시 신속한 여론 파악 및 대응이 가능함.
- 플랫폼 간 데이터 통합 분석을 통해 팬 반응 모니터링 체계를 구축하고, 요약 결과를 챗봇 및 자동 응답 시스템과 연계함으로써, 감정 기반의 자동 대응 메시지 제공이 가능함. 이를 통해 커뮤니케이션 효율을 높이고 운영 비용을 절감할 수 있음.
- AI 기반 커뮤니케이션 플랫폼으로 발전시켜, 브랜드 이미지 향상과 팬 참여도 제고에 긍정적 기여가 될 것이라 생각됨.

한계점 및 향후 과제

- 이분법적 감정 분류 방식은 세부 감정 반영의 한계
  → Likert 척도 기반 라벨링 기법 도입으로, 감정 강도를 정량화하여 데이터 정밀도 및 분석 설명력 향상 기대
- 하이퍼파라미터 튜닝이 제한적
  → AutoML 및 GridSearch를 활용한 자동 튜닝을 통해, 요약 및 피드백 기능 정밀도 향상 및 모델 성능 개선 가능성 확보