개인 프로젝트

뉴스 댓글 감정분석

이준수 | **Phone:** 010-7129-8901 | **E-mail:** rjaekawpxm@naver.com

목차

01 프로젝트 개요

02 문제 정의 & 목표

03 데이터 수집 및 전처리

04 분석 방법 및 지표 설계

05 분석 결과 및 검증

06 결론 및 개선 방향

01 프로젝트 개요

프로젝트 주제

뉴스 댓글 **감정분석** 기반
기사 단위 여론 위험도 분석 (v2)

v2 업그레이드 방향

기사 단위 여론 위험도 **수치화**

수치 기반 검증(O/X 테스트) 수행

프로젝트 배경

뉴스 댓글은 대중의 **감정**과 **반응**이 직접적으로 드러나는 데이터

기존 감정분석은 기사 간 **비교 및 위험도** 판단에 한계 존재

02 문제 정의 & 목표

문제 정의 및 인식

- 기존 뉴스 댓글 감정분석은
  전체 댓글의 긍·부정 비율 제시에 그쳐
  기사 간 여론 차이를 비교하기 어려움

- 댓글 수가 기사마다 다름에도
  동일 기준으로 해석되어 **왜곡 가능성** 존재

- 분석 결과가 신뢰 가능한지에 대한
  정량적 검증 과정이 부족함

프로젝트 목표 (v2)

- 뉴스 댓글 감정분석 결과를
  기사 단위로 **요약·비교 가능**하도록 설계

- 여론의 강도와 위험도를
  수치(risk_score)로 표현

- 정규화·민감도 테스트를 통해
  분석 지표의 신뢰성 검증

03 데이터 수집 및 전처리

데이터 수집

네이버 뉴스 기사 URL 수집

- 사회 / 스포츠 섹션 중심

Selenium 기반 댓글 수집

- 기사별 댓글 텍스트
- 작성일
- 공감/비공감 수집

기사 단위 분석을 위해
댓글–기사 URL 매핑 구조

데이터 전처리

중복 댓글 제거

내용이 없는 댓글(NaN) 제외

기사별 댓글 수 집계 기준 정의

분석에 필요한 컬럼만 선별하여 사용

데이터 수집 과정의 문제와 해결

네이버 뉴스 **DOM 구조 변경**으로 크롤링 불안정 발생

상대경로 사용으로 인한 **파일 저장 오류**

→ Pathlib 기반 **프로젝트 루트 기준 경로**로 통일하여 해결

댓글이 없는 기사에 대한 **예외 처리 로직 추가**

뉴스 기사 →  **댓글 수집** →  **전처리** →  **기사 단위 데이터**

04 분석 방법 및 지표 설계

분석 방법

- 뉴스 댓글 텍스트를 대상으로 **감성사전 기반 감정 분석** 수행
- 각 댓글을 **긍정 / 중립 / 부정**으로 분류
- 복잡한 딥러닝 모델 대신 **해석 가능성이 높은 방식**을 선택

분석 방식 선택 이유

- 결과 해석이 직관적이며 **기사 간 비교**에 용이
- 대규모 학습 데이터 없이도 **안정적인 분류** 가능
- **지표 설계 및 수치 검증**에 적합한 구조

기사 단위 요약 지표

n_comments : 기사별 댓글 수

pos_ratio / neu_ratio / neg_ratio : 감정 비율

sentiment_score : 긍·부정 균형 지표

여론 위험도 지표 (risk_score)

부정 댓글 비율과

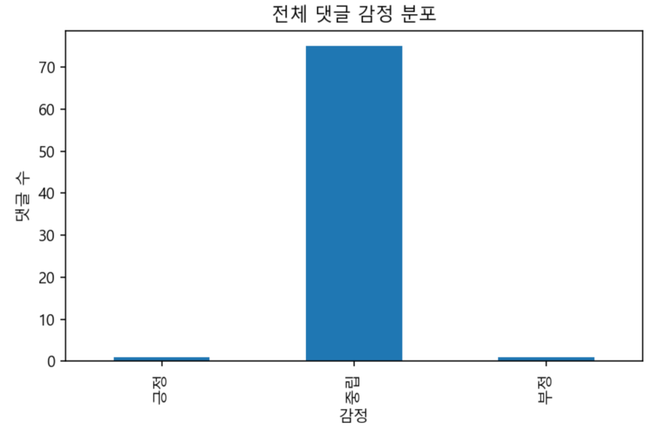
여론 반응 강도를 종합하여 위험도 점수 산출

기사 간 여론 강도 비교를 위한 핵심 지표

“단순 감정 분류를 넘어,

기사 단위 여론 위험도를 수치로 정의하였다.”

05 감정분석 결과 요약



전체 감정 분포

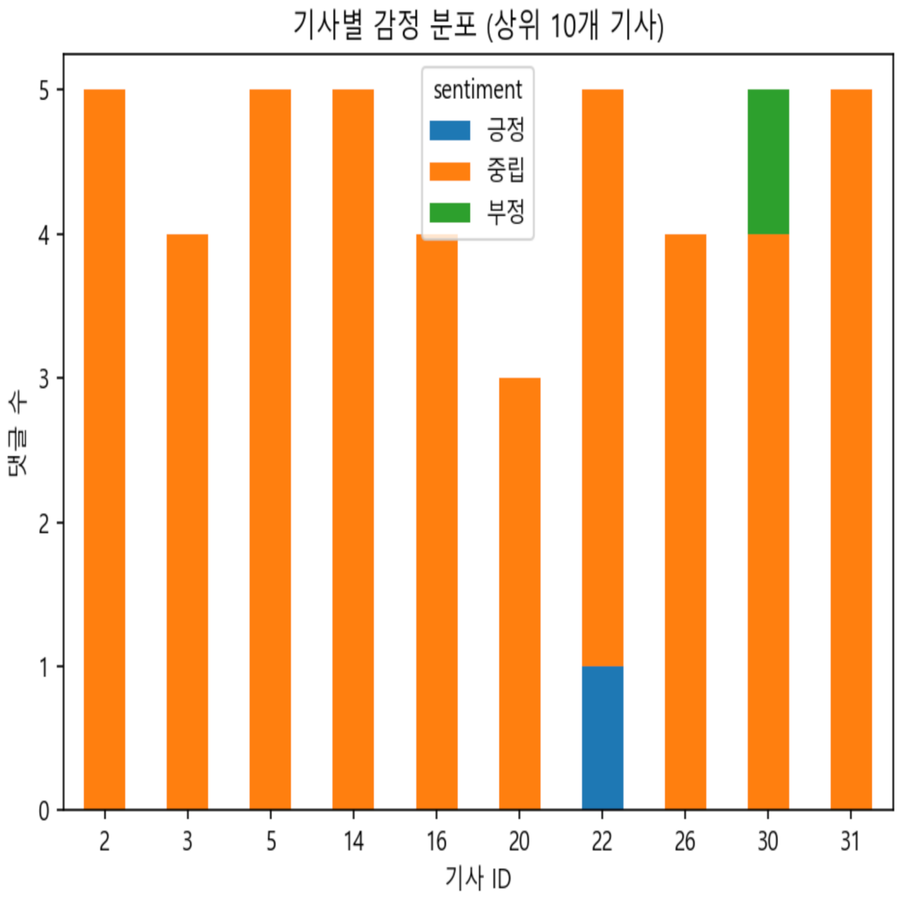
- **중립 감정**의 비중이 가장 **높게** 나타남
- 사실 전달·의견 표현 형태의 댓글이 중립으로 분류되는 구조적 한계에 따른 결과임

기사 단위 감정 분포 차이

- 기사별 댓글 수는 3~5개 수준으로 크지 않음
- 중립 댓글이 주를 이루는 공통된 패턴이 관찰됨

결과 해석

- 전체 감정 비율만으로는 기사 간 여론 반응의 차이를 판단하기 어려움
- 이후 단계에서 기사 단위 **위험도 지표를 설계**할 필요성으로 이어짐



06 기사 단위 여론 위험도 분석

> “여론 위험도는 댓글의 수가 아니라, 부정 감정의 집중도에 의해 결정된다.”

주요 결과 요약

- 1위 기사 Risk Score = 60

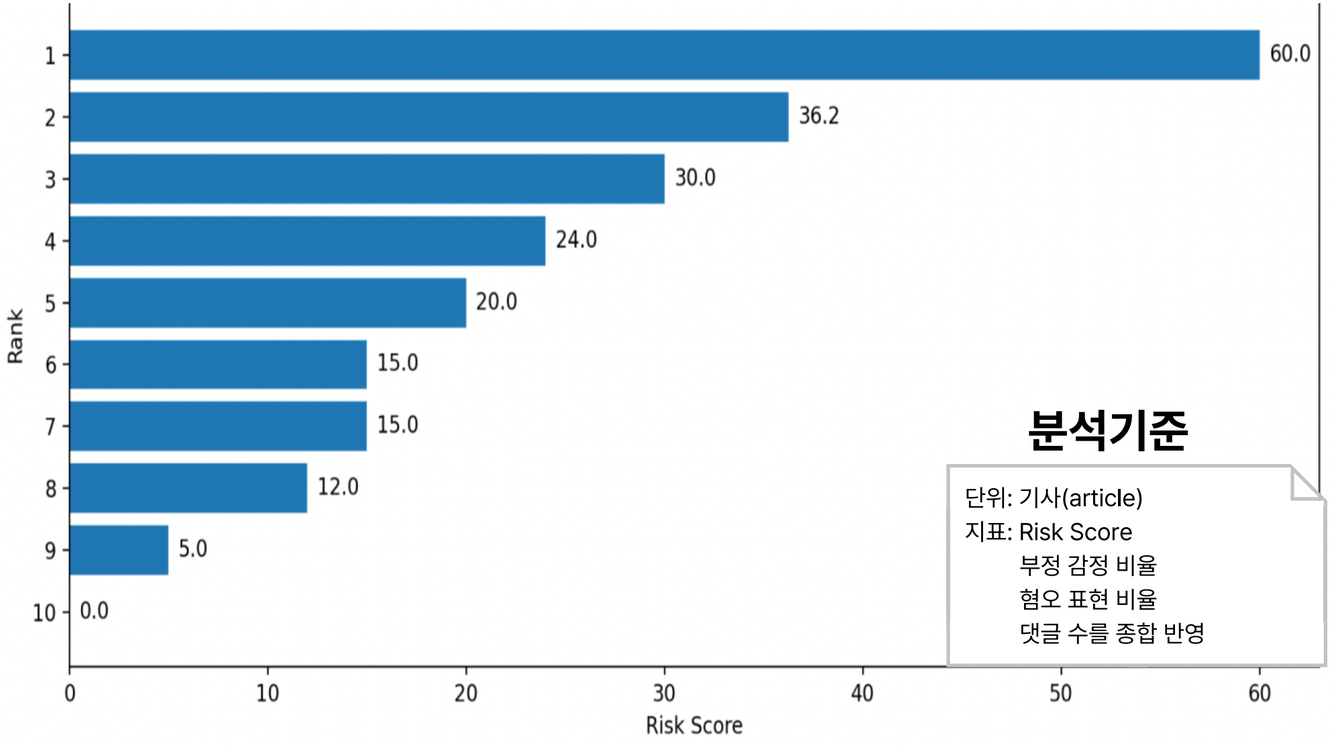
  다른 기사 대비 현저히 높은 위험도
  댓글 수는 적지만, 부정·공격적 표현 집중

- 2~5위 기사

  Risk Score 20~36 구간에 집중
  댓글 수와 무관하게 부정 감정 밀도가 높을수록 상위권 진입

- 하위 기사

  댓글 수가 존재하더라도
  감정 강도가 낮으면 위험도는 낮게 유지



분석기준

단위: 기사(article)
지표: Risk Score
부정 감정 비율
혐오 표현 비율
댓글 수를 종합 반영

07 위험도 지표 정규화 및 신뢰성 검증

> **Risk Score는 댓글 수 편차에도 안정적으로 작동하는 신뢰 가능한 지표이다.**

주요 결과 요약

- 댓글 수 편차로 인한 위험도 왜곡 여부 검증
- Risk Score를 **댓글 수로 정규화**하여 비교 분석 수행

검증 결과

- Risk Score ↔ 정규화 지표 상관계수 = 0.87
- 정규화 전·후 Top5 기사 100% 동일 (5/5)

```
== risk_score ↔ risk_norm 상관 ==
pearson corr = 0.8723

== 요약 ==
Top5 겹침 개수 : 5 / 5
```

- 정규화 이후에도 위험 기사 순위가 유지됨
- 위험도 지표는 댓글 수에 과도하게 의존하지 않음

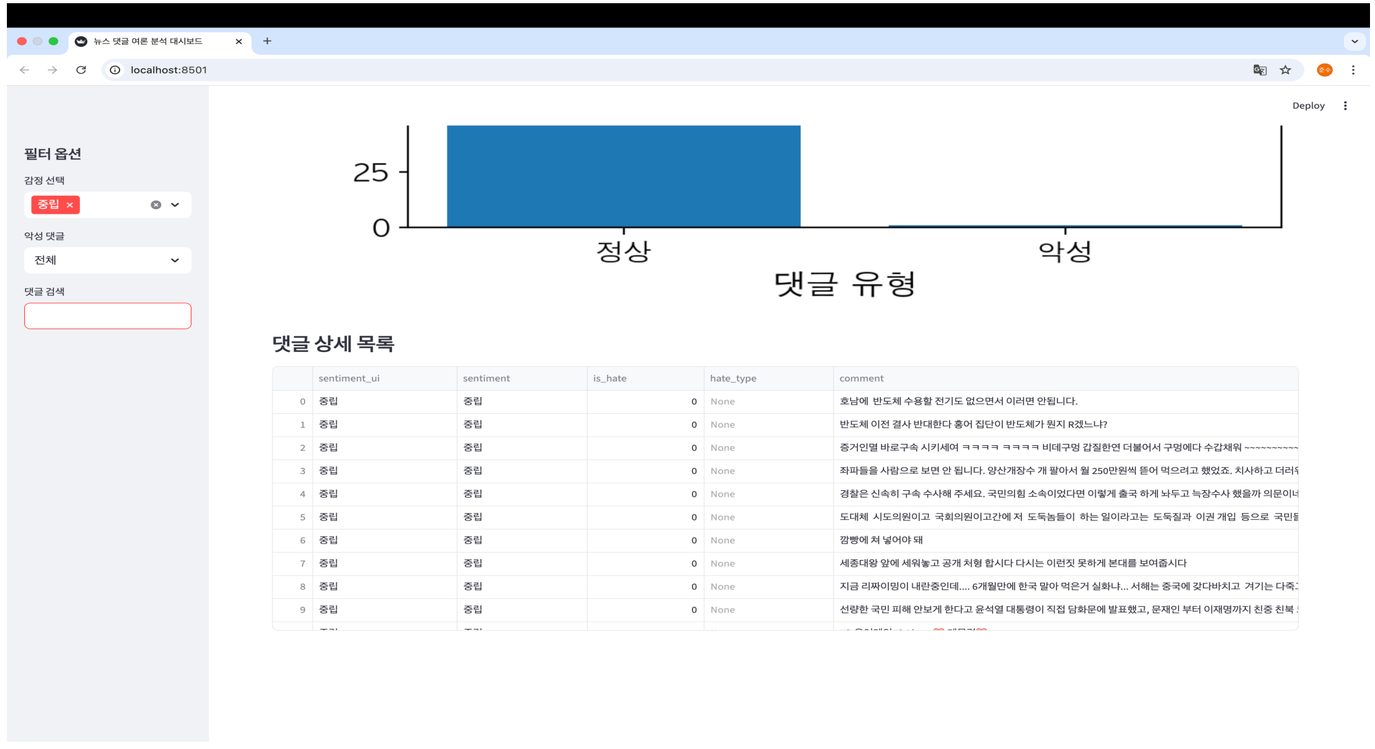
정규화 지표(risk_norm)는 risk_score / n_comments로 계산함

08 분석 결과 시각화 : Streamlit 대시보드 구현

> “분석 결과를 사용자가 직접 탐색할 수 있도록 대시보드 형태로 구현하였다”

대시보드

- 감정 분석·악성 댓글 분석 결과를 대시보드 형태로 시각화
- 필터 기능을 통해 감정/악성 여부별 댓글 탐색 가능
- 실제 댓글 데이터 제공으로 분석 결과의 해석 가능성 강화
- 분석 → 결과 → 서비스화까지 이어지는 전체 파이프라인 완성



09 프로젝트 한계점

한계점

“ 본 프로젝트는 제한된 데이터 환경에서 분석을 수행하여 결과 해석에 주의가 필요하다. ”

데이터 규모의 제한

- 기사당 댓글 수가 많지 않음
- 일부는 1~5개의 댓글만으로 분석
- 감정 분포가 중립에 편중되는 현상

→ 소규모 데이터에서는
감정 분포 왜곡 가능성 존재

감정 분석 모델의 단순성

- 사전 기반 감정 분류 방식 사용
- **맥락·풍자·은어 표현**에 대한 해석 한계
- 정치·사회 이슈 댓글에서
감정 강도 과소/과대 평가 가능

기사 주제·섹션 정보 미반영

- 기사 내용의 민감도
(정치·사회·연예 등) 미구분
- 동일 지표를 모든 기사에 동일 적용
- 주제별 감정 반응 차이 반영 부족

10 개선 방향 및 확장 가능성

개선 방향

“ 데이터 확장과 모델 고도화를 통해 여론 위험도 분석의 신뢰성을 높일 수 있다. ”

데이터 규모 확장

- 기사당 댓글 수 확대
- 기간 단위 수집(일/주/월) 적용
- 기사 수 및 댓글 수 누적 기반 분석

감정 분석 모델 고도화

- KoBERT, KLUE-BERT 등 딥러닝 기반 모델 적용
- 혐오·공격성 전용 분류기 추가
- 감정 강도(score) 예측 방식 도입

위험도 지표 세분화

- 기사 섹션별 위험도 기준 차등 적용
- 댓글 수 가중치 동적 조정

  시간 흐름에 따른 위험도 변화 추적

11 결론 및 의의

결론

- 네이버 뉴스 댓글 데이터를 기반으로
  감정(긍·중·부정) 및 악성 댓글 분석을 수행하였다.

- 전체 댓글 분포에서는 중립 의견이 다수를 차지하였으나,
  여론이 전반적으로 **안정적임**을 보여주는 동시에
  부정·악성 반응이 집중되는 기사를
  상대적으로 명확히 식별할 수 있는 조건으로 작용하였다.

- 기사 단위 위험도 분석 및 정규화 테스트 결과,
  댓글 수 편차에도 불구하고 고위험 기사 식별 결과는
  일관되게 유지되었다.

의의

- 중립 의견이 많은 환경 속에서도,
  소수의 강한 부정·악성 반응을 분리하여
  해석할 수 있는 분석 구조를 제시하였다.

- 사전 기반 감정 분석의 한계를 인지하고,
  절대적인 감정 판단이 아닌 상대 비교와
  위험 신호 탐지 목적의 지표로 활용하였다는 점에서 의미가 있다.

- 감정·악성 분석 결과를 Streamlit 대시보드로 시각화하여
  분석 결과의 **해석 가능성과 실질적 활용성**을 함께 확보하였다.