

뉴스 댓글 기반 기사 위험도 분석 및 개선 프로젝트

(개인프로젝트)

1. 프로젝트 개요

본 프로젝트는 뉴스 기사 댓글 데이터를 활용하여 대중의 감정 반응을 분석하고, 기사 단위 여론 위험도를 정량화하는 것을 목표로 수행되었습니다.

초기 버전(v1)에서는 댓글 감정 분포 분석에 집중했으나, 분석 과정에서 기사 간 비교가 어렵고 결과 해석에 한계가 존재함을 확인했습니다.

이에 따라 v2에서는 댓글 감정 분석 결과를 기반으로 기사 단위 여론 위험도를 수치화하고 비교 가능하도록 분석 구조를 개선했습니다.

2. 문제 정의 (v1 분석의 한계)

기존 분석 방식에서는 다음과 같은 문제가 존재했습니다.

- 댓글 감정 분포만 제공되어 기사 간 비교가 어려움
- 댓글 수 차이에 따라 분석 결과 해석 왜곡 가능성 존재
- 분석 결과의 신뢰성을 정량적으로 검증하기 어려움

즉,

댓글 감정 분석 결과를 기사 위험도로 해석하기 어려운 구조적 문제가 존재했습니다.

3. 해결 접근 방식 (v2 개선)

기존 댓글 감정 분류 분석을 유지하면서,

다음과 같은 개선 방향을 적용했습니다.

① 기사 단위 위험도 지표 설계

댓글 감정 비율을 기반으로 기사별 여론 위험도를 계산하는

Risk Score 지표를 새롭게 설계했습니다.

이를 통해 기사 간 위험도 비교가 가능해졌습니다.

② 데이터 수집 및 전처리 개선

분석 과정에서 다음 문제를 해결했습니다.

- 기사별 댓글 구조 정리
- 댓글이 없는 기사 예외 처리
- 데이터 누락 및 구조 불일치 문제 해결

이를 통해 기사 단위 분석 구조를 안정적으로 구성했습니다.

③ 지표 신뢰성 검증

위험도 지표가 데이터 규모에 따라 왜곡되지 않는지 확인하기 위해

- 댓글 수 정규화 적용
- 정규화 전/후 위험도 비교 수행
- 상관계수 및 순위 비교를 통한 안정성 검증

을 수행하여 분석 지표의 일관성을 확인했습니다.

④ 분석 결과 시각화 및 서비스화

분석 결과를 사용자가 직접 탐색할 수 있도록

- Streamlit 기반 대시보드를 구현
- 기사별 감정 분포 및 위험도 시각화
- 실제 댓글 데이터 조회 기능 제공

을 통해 분석 결과 활용성을 높였습니다.

4. 핵심 분석 결과

분석 결과 다음을 확인할 수 있었습니다.

- 댓글 수가 많지 않아도 부정 감정이 집중될 경우 위험도가 높아짐
- 기사 단위 비교를 통해 여론 위험도가 높은 기사 식별 가능
- 지표 정규화 이후에도 위험 기사 순위가 유지되어 신뢰성 확보

즉,

단순 감정 분류를 넘어 기사 단위 위험도 분석이 가능함을 확인했습니다.

5. 문제 해결 과정

문제 1. 댓글 감정 분석 결과의 비교 어려움

기사마다 댓글 수가 달라 직접 비교가 어려웠습니다.

→ 기사 단위 위험도 지표를 설계하여 해결

문제 2. 데이터 구조 불일치 문제

기사별 댓글 데이터 구조가 일정하지 않았습니다.

→ 전처리 및 예외 처리 로직을 추가해 해결

문제 3. 지표 신뢰성 검증 필요

위험도 지표가 데이터 규모에 따라 변형될 가능성이 존재했습니다.

→ 정규화 및 상관 분석을 통해 안정성 검증 수행

6. 프로젝트 의의

- 단순 감정 분석을 넘어 기사 위험도 분석으로 확장
- 분석 결과를 서비스 형태로 구현하여 활용성 확보
- 분석 과정에서 발생한 문제를 개선하며 모델 구조를 발전시킨 사례

7. 사용 기술 및 개발 환경

- Python
- Pandas, NumPy
- 감정 분석 기반 댓글 분류
- Streamlit 대시보드 구현
- 기사 및 댓글 데이터 크롤링 및 전처리

GitHub Repository

(v1)

https://github.com/rjaekawpxm1-netizen/portfolio/tree/main/Projects/naver_news_sentiment

(v2)

https://github.com/rjaekawpxm1-netizen/portfolio/tree/main/Projects/naver_news_sentiment%202