

쇼핑 리뷰 감성분석 발표

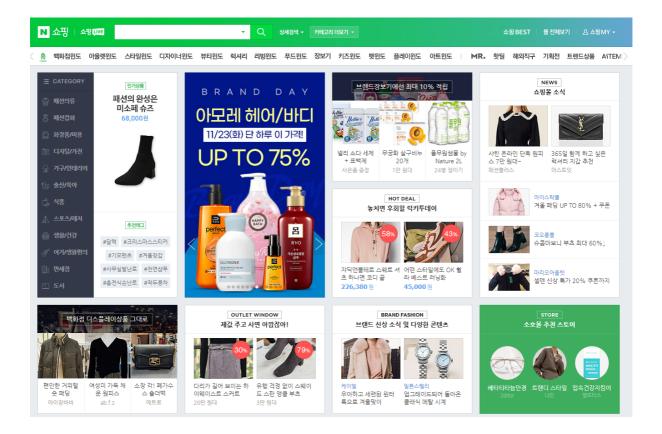
| Project Member | 김의겸 이서영 이세영 조수연 홍지원

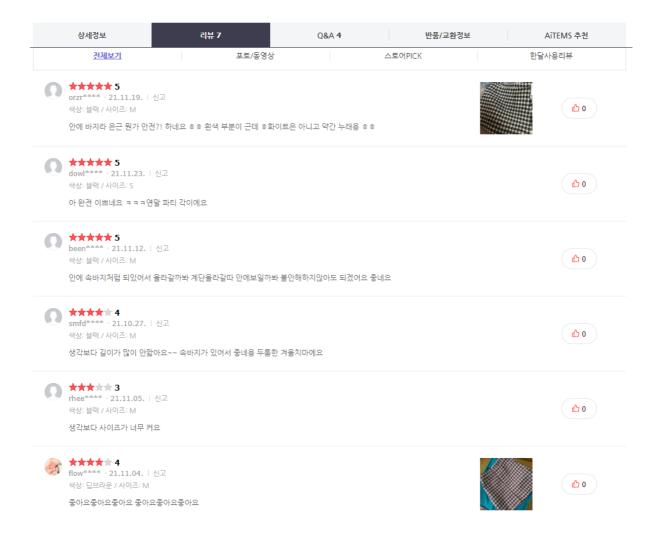
🚺 주제: 네이버 **쇼핑 리뷰 데이터**로 구축한 **감성분석 모델**의 **범용적 가치 및 성능** 탐색



주제 선정의 배경 및 분석 목표:

- (1) 각종 리뷰에 **범용적으로 사용 가능한 좋은 성능의 감성분석 모델의 가능성** 연구
- (2) 별점 리뷰의 한계에 따른 자연어처리(NLP) 기술의 활용 필요성
- 🙎 사용한 데이터
- (1) 네이버 쇼핑(https://shopping.naver.com/)
- ⇒ 사용 데이터는 특정 기간(2020.06~2020.07)동안 작성된 20만 건의 쇼핑 리뷰 데이터

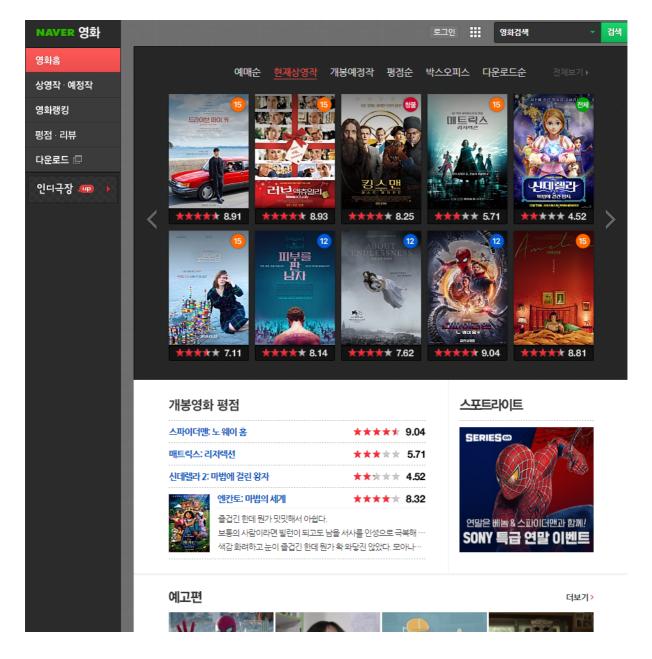






→ 긍정(4~5점)과 부정(1~2점)의 리뷰 데이터 20만 건 수집 (중립적인 3점의 데이터는 제외)

(2) 네이버 영화 (https://movie.naver.com/)



⇒ 2만 개 데이터를 추출하여 모델 테스트 진행

Process

- 🚺 네이버 쇼핑 리뷰 데이터 수집
- 2 긍정(4~5점)과 부정(1~2점)의 리뷰 데이터 수집 (애매한 혹은 중립적인 3점의 데이터는 제외)
- 🔋 데이터 전처리(pre-processing) 작업
 - 데이터 라벨링(labeling): 긍정(1), 부정(0)

- 데이터 토큰화(tokenizer): 형태소 분석기 Mecab

(+ 불용어 제거, 정규표현식으로 한글 이외의 언어 제거)



토크나이저 비교:

Mecab 빠른 처리 속도, 어간과 어미 구분 O (e.g. 좋네요 → 좋, 네요) Okt 느린 처리 속도, 어간과 어미 구분 x (e.g. 좋네요 → 좋네요)

- 데이터 인코딩: 단어별 정수 인코딩
- 패딩(padding) max len = 80
- 데이터셋 분리: train data, test data

4 모델 학습

- GRU, KcBERT, LSTM, Bi-LSTM



KcBERT: BERT가 대부분 한국어 위키, 뉴스 기사, 책 등 잘 정제된 데이터를 기반으로 학습한 모델이라면, KcBERT는 네이버 뉴스의 댓글과 대댓글 등을 수집해 오탈자, 신조어, 구어체 등을 반영한 Pre-trained BERT 모델

[5] 네이버 영화 리뷰 데이터(후기 + 별점)로 같은 전처리 작업 진행 후 2만 개 데이터에 대한 테스트

Results

1. LSTM

a. 쇼핑 리뷰 정확도: 0.8846

b. 영화 리뷰 정확도: 0.8122 (loss: 0.4088)

2. Bidirectional LSTM

a. 쇼핑 리뷰 정확도: 0.9232

b. 영화 리뷰 정확도: 0.8148 (loss: 0.4122)

3. GRU

a. 쇼핑 리뷰 정확도: 0.9000

b. 영화 리뷰 정확도: 0.7234 (loss: 0.7022)

4. KcBERT

a. 쇼핑 리뷰 정확도: 0.9352

b. 영화 리뷰 정확도: 0.7997 (loss: 0.4669)

(+ Machine Learning Model: Naive Bayes Classification, Logistic Regression, Random Forest)

1. Naive Bayes Classification: accuracy 0.55

2. Logistic Regression: accuracy **0.51**

3. Random Forest Classification: accuracy 0.75 accuracy 0.60

Conclusion

- 1. 쇼핑 리뷰 데이터 상에서 가장 좋은 성능을 보인 모델: KcBERT, Bidrectional LSTM
- 2. 영화 리뷰 데이터를 통해 범용성 측면에서 가장 좋은 성능을 보인 모델: Bidirectional LSTM

단, KcBERT가 영화 리뷰 데이터에서 잘 작동하지 않았기 때문에 단정 짓기는 어려움.



미국 스탠포드대학 Christopher Manning 교수의 <u>CS224n</u> 강의 中
"The de facto consensus in NLP in 2017 is that **no matter what the task**,
you throw a **BiLSTM** at it, with attention if you need information flow, and
you get great performance!"



BERT 성능의 한계: 여러 단어가 조합됐을 경우, 해당 단어 간 상관 관계를 고려하지 않는 점

Further Questions

- 1. BERT 작동이 잘 안되며 구현이 매우 까다로움.
- 2. Fine-tuning을 통해 훈련 데이터의 맥락 의존성을 해결할 수 있을까?
- 3. 다양한 **토크나이저** 간 성능 비교(e.g. Hannanum, Kkma, Komoran, Mecab, Okt)
- 4. 다양한 <mark>언어</mark>를 포함한 감성분석 모델 (e.g. 영어)