



쇼핑 리뷰 감성분석 발표

| Project Member | 김의겸 이서영 이세영 조수연 홍지원

1 주제: 네이버 쇼핑 리뷰 데이터로 구축한 감성분석 모델의 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



2 사용한 데이터

(1) 네이버 쇼핑(<https://shopping.naver.com/>)

⇒ 사용 데이터는 특정 기간(2020.06~2020.07)동안 작성된 20만 건의 쇼핑 리뷰 데이터

The screenshot shows the Naver Shopping homepage with a green header. The main content area is divided into several sections:

- Left Sidebar:** A vertical menu with categories like '패션의류' (Fashion), '패션잡화' (Fashion Accessories), '화장품/미용' (Cosmetics), '디지털/가전' (Digital/Electronics), '가구/인테리어' (Furniture/Interior), '출산/육아' (Pregnancy/Childcare), '식품' (Food), '스포츠/레저' (Sports/Leisure), '생물/건강' (Biological/Health), '여가/생활편의' (Hobby/Life Convenience), '민세점' (Department Store), and '도서' (Books).
- Top Banner:** A large blue banner for 'BRAND DAY' featuring hair and body products, with a discount of 'UP TO 75%'.
- Hot Deal Section:** A section titled 'HOT DEAL' with a sub-header '농치면 후회할 럭키투데이' (Don't miss the lucky Tuesday). It features two items: a jacket for 58% off (226,380 원) and a watch for 43% off (45,000 원).
- Outlet Window Section:** A section titled 'OUTLET WINDOW' with a sub-header '계값 주고 사면 아깝잖아!' (It's a pity to buy without a discount!). It features two items: a skirt for 30% off (20만 원대) and a sweater for 79% off (3만 원대).
- Brand Fashion Section:** A section titled 'BRAND FASHION' with a sub-header '브랜드 신상 소식 및 다양한 콘텐츠' (Brand new arrivals and various content). It features two items: a jacket for 58% off (226,380 원) and a watch for 43% off (45,000 원).
- Store Section:** A section titled 'STORE' with a sub-header '소호운 추천 스토어' (Sohoon's recommended store). It features three items: a watch for 58% off (226,380 원), a jacket for 43% off (45,000 원), and a watch for 43% off (45,000 원).

상세정보	리뷰 7	Q&A 4	반품/교환정보	AITEMS 추천
전체보기	포토/동영상	스투어PICK	한달사용리뷰	
<p>★★★★★ 5</p> <p>orzi**** · 21.11.19. 신고</p> <p>색상: 블랙 / 사이즈: M</p> <p>안에 바지라 은근 뭔가 안전?! 하네요 ㅎㅎ 흰색 부분이 근데 화이트는 아니고 약간 누레음 ㅎㅎ</p>		 <p>0</p>		
<p>★★★★★ 5</p> <p>dowl**** · 21.11.23. 신고</p> <p>색상: 블랙 / 사이즈: S</p> <p>아 완전 이쁘네요 ㅋㅋㅋ연말 파티 각이예요</p>		<p>0</p>		
<p>★★★★★ 5</p> <p>been**** · 21.11.12. 신고</p> <p>색상: 블랙 / 사이즈: M</p> <p>안에 속바지처럼 되있어서 올라갈까봐 계단을라갈때 안에보일까봐 불안해하지않아도 되겠어요 좋네요</p>		<p>0</p>		
<p>★★★★☆ 4</p> <p>smfd**** · 21.10.27. 신고</p> <p>색상: 블랙 / 사이즈: M</p> <p>생각보다 길이가 많이 안짧아요~~ 속바지가 있어서 종네용 두툼한 겨울치마예요</p>		<p>0</p>		
<p>★★★★☆ 3</p> <p>rhee**** · 21.11.05. 신고</p> <p>색상: 블랙 / 사이즈: M</p> <p>생각보다 사이즈가 너무 커요</p>		<p>0</p>		
<p>★★★★☆ 4</p> <p>flow**** · 21.11.04. 신고</p> <p>색상: 딥브라운 / 사이즈: M</p> <p>좋아요좋아요좋아요 좋아요좋아요좋아요</p>		 <p>0</p>		



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

(2) 네이버 영화 (<https://movie.naver.com/>)

NAVER 영화

영화홈

상영작 · 예정작

영화랭킹

평점 · 리뷰

다운로드

인디극장

로그인

영화검색

검색

예매순

현재상영작

개봉예정작

평점순

박스오피스

다운로드순

전체보기

드라이브 마이 카

8.91

리브 애츨리

8.93

킹스맨

8.25

매트릭스 리저렉션

5.71

신데렐라

4.52

노랑

7.11

피부를 판 남자

8.14

ABOUT ENDLESSNESS

7.62

스파이더맨: 노 웨이 홈

9.04

안칸토: 마법의 세계

8.32

개봉영화 평점

스파이더맨: 노 웨이 홈

★★★★★ 9.04

매트릭스: 리저렉션

★★★★★ 5.71

신데렐라 2: 마법에 걸린 왕자

★★★★★ 4.52

안칸토: 마법의 세계

★★★★★ 8.32

즐겁긴 한데 뭔가 밋밋해서 아쉽다.

보통의 사람이라면 빌런이 되고도 남을 서사를 인성으로 극복해 ...

색감 화려하고 눈이 즐겁긴 한데 뭔가 확 와닿진 않았다. 모어나...

스포트라이트

SERIES

연말은 베놈 & 스파이더맨과 함께!

SONY 특급 연말 이벤트

예고편

1

2

3

4

더보기

⇒ 2만 개 데이터를 추출하여 모델 테스트 진행

Process

- 1 네이버 쇼핑 리뷰 데이터 수집
- 2 긍정(4~5점)과 부정(1~2점)의 리뷰 데이터 수집 (애매한 혹은 중립적인 3점의 데이터는 제외)
- 3 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 오타자, 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

- 쇼핑 리뷰 정확도: **0.8846**
- 영화 리뷰 정확도: **0.8122** (loss: 0.4088)

2. Bidirectional LSTM

- 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, Bidirectional 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 CS224n 강의 중

"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 **토큰라이저** 간 성능 비교(e.g. Hannanum, Kkma, Komoran, Mecab, Okt)
4. 다양한 **언어**를 포함한 감성분석 모델 (e.g. 영어)