머신러닝 프로젝트 (주말오전-2조)

감성분석을 통한 다양한 품목 후기의 유사도 비교

이진주 / 윤의정 / 채유진 / 최선아 / 최승혁

품목에 따라 후기 간에 차이점이 있을까?

- 후기는 소비자의 의견을 직관적으로 알 수 있는 데이터
- 판매하는 품목이 달라지면 소비자의 긍/부정 후기의 양상이 달라질까?
- 분석 알고리즘이 달라지면 결과가 달라질까?
- 머신러닝(감성분석)을 통해 각 품목별 후기를 비교해보자!

프로젝트 순서

STEP 1

감성분석 모델 작성

- 로지스틱 회귀
- 나이브베이즈 분류

STEP 2

데이터 학습

- 긍/부정 단어 확인
- 정확도 측정

STEP 3

각 품목별 후기 차이점 비교
모델별 차이점 비교

- 긍/부정 단어 비교
- 정확도 비교

사용 데이터

패션

화장품

IT

가전

- 감정분석용 데이터 : AIHub(https://aihub.or.kr) 이용
- 쇼핑몰 및 SNS 후기에서 추출한 20만개의 데이터
- 5~4점 : 긍정 / 1~2점 : 부정으로 데이터 분석 (긍정 : 부정 = 9 : 1)

알고리즘

로지스틱 회귀분석 (Logistic Regression)	나이브 베이즈 분류 (Naive Bayes Classification)
• 데이터가 특정 카테고리에 속할지를 0과 1사이의 연속적인 확률로 예측하는 회귀 알고리즘 • 간단하고 효율적이며 해석이 용이함 • 이상치에 민감하며, 큰 표본크기가 필요함	• 데이터의 특징이 모두 상호 독립적이라는 가정하에 각 클래스(레이블)에 속할확률을 계산하는 분류 방법 • 성능이 빠르며 작은 표본으로도 수행이 용이함 • 변수가 독립적이지 않을 경우 오류가 발생하며, 추정된 확률은 예측된 범주보다 신뢰가 떨어짐

분석 결과 긍정 단어 Top 10

	패션	IT	가전	화장품	전체
로지스틱 회귀분석	만족합니다 좋네요 편하고 이쁘네요 좋아요 편해요 좋고 감사합니다 최고 예뻐요	최고 좋아요 좋습니다 만족해요 좋네요 만족 만족합니다 만족스러워요 감사해요 감사합니다	만족합니다 만족해요 좋네요 만족스러워요 좋아요 좋습니다 최고 편해요 만족 감사합니다	좋습니다 좋고 좋아요 촉촉하고 좋네요 선물 만족합니다 항상 최고 요즘	최고 감사합니다 만족합니다 만족해요 만족스러워요 촉촉하고 좋네요 좋습니다 만족스럽습니다 강추
나이브 베이즈 분류	좋아요 너무 편하고 같아요 맘에 좋고 가볍고 가격대비 조금 좋네요	좋아요 너무 좋네요 정말 맘에 들어요 만족해요 마음에 좋습니다 아주	좋아요 너무 좋네요 제품 마음에 아주 좋습니다 있어서 만족합니다 만족해요	좋습니다 좋아요 너무 좋고 향이 같습니다 같아요 좋은 많이 향도	좋아요 너무 좋습니다 좋네요 마음에 좋고 제품 정말 아주 같아요

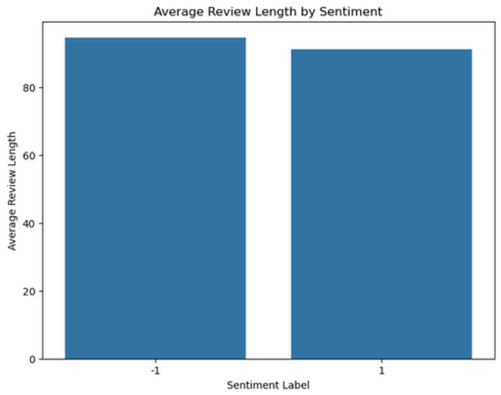
분석 결과 부정 단어 Top 10

	패션	IT	가전	화장품	전체
로지스틱 회귀분석	뒤꿈치 환불 싸구려 빠짐 아프네요 실망 뻣뻣하고 반품 엉망 이유	화나네요 안되서 별로 불만족 비추 반품 짜증나네요 안되고 실망 최악	떨어지네요 무겁네요 무겁고 광고 실망 불만족 별루 비추 안되고 반품	잘못 싸구려 비추 전혀 정품 최악 방송 그냥 실망 반품	불만족 무크 뻣뻣하고 엉망 짜증나네요 비추 아깝네요 반품 최악 별루
나이브 베이즈 분류	너무 그냥 사이즈가 많이 품질이 조금 같아요 생각보다 좋아요 사이즈	너무 좋아요 제품 마음에 ㅠㅠ 있어서 생각보다 별로네요 좋네요 가격이	너무 제품 그냥 좋아요 생각보다 ㅠㅠ 소음이 소리가 많이 그리고	너무 많이 같습니다 향이 같아요 다른 제형이 그냥 않고 효과가	너무 그냥 많이 좋아요 같아요 생각보다 사이즈가 품질이 제품 ㅠㅠ

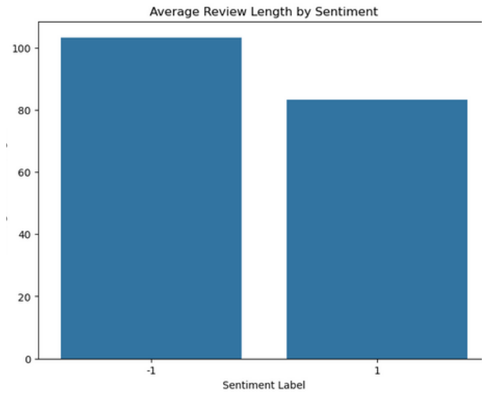
분석 및 결론 품목 후기에 따른 분석

- 긍정 단어의 경우 패션/IT/화장품/가전이 대부분 비슷한 단어의 조합을 보임
  - “좋아요”, “만족”, “감사”, “최고”
- 부정 단어의 경우, IT와 가전이 유사한 모습을 보임
  - “반품”, “안되고”
  - 패션, 화장품에 비해 고가인 제품군이 많아 그만큼 반품 관련 리뷰, 문의가 많은 것으로 예상됨
- 가전, 화장품의 경우 부정 단어 내에서 고객 유입 경로를 언급하기도 함
  - 가전의 경우 광고, 화장품의 경우 방송 채널으로 구매 후 부정 후기 남긴 경우가 많은 것으로 추측됨
  - 광고 및 방송을 통한 구매 후 고객 이탈 방지 전략을 세우는 것이 바람직하다고 예측할 수 있음

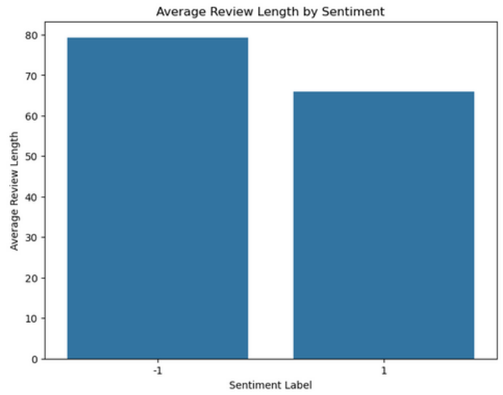
분석 및 결론 후기 데이터 길이 분석



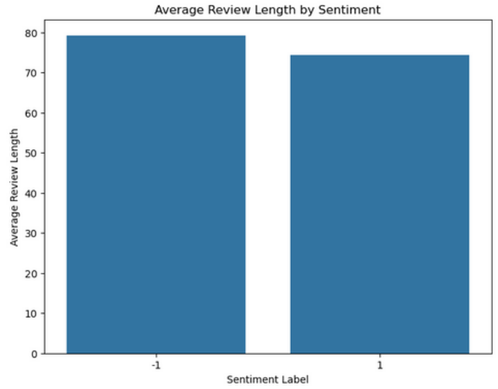
화장품



가전



패션



IT

- 모든 리뷰 데이터에서 부정 후기가 더 긴 양상을 보임 (왼쪽-부정 / 오른쪽-긍정)
- 부정 후기의 길이가 더 길지만, 긍정 후기 대비 압도적으로 숫자가 적음
- 고객이 부정 후기를 잘 남기진 않지만, 작성하게 되면 길고 자세히 작성한다는 것을 추측할 수 있음
  따라서 긍정후기보다 부정후기에 보다 신경을 쓰고 대안방안을 마련해야 함

분석 및 결론 알고리즘 모델에 따른 분석

- 로지스틱 회귀와 나이브베이즈 각 모델에서 추출된 긍/부정 단어의 특성이 다른 양상을 보임.
  - 로지스틱 회귀 : 상관관계에 따라 추출 > 긍/부정 단어가 비교적 뚜렷하게 구분
  - 나이브베이즈 : 클래스에 속할 확률에 따라 추출 > 긍/부정보다는 섹터가 구분되는 단어가 추출됨 (사이즈, 소음, 향 등)
- “좋아요”, “너무” 의 경우 긍/부정 단어 모두에 나타남

- 각 품목을 학습시킨 알고리즘에 다른 품목을 대입한 경우 알고리즘 차이와 상관없이 0.8~0.95의 높은 정확도를 보임

∴ 모델의 특성에 따른 차이는 있으나 섹터에 따른 두드러지는 차이는 없다고 볼 수 있다!

감사합니다!

질문이 있으신가요?