텍스트 마이닝을 이용한 소비자 분석 프로젝트

휴먼

휴먼지능정보공학과 류지혜 이지민

목차

프로젝트 요약

✓ 연구 주제

➡ 비싼 전자제품을 소유하고 있음에 만족? vs 전자제품의 발전된 성능에 만족?

프로젝트 요약

✓ 결론 도출 과정

Review의 특성

: 이미 구입, 사용하고 나서 쓰는 것 -> 사용자가 선택한 이유가 아니라

제품에 만족한 이유가 드러남.

전자제품의 특성

: 가격대가 비싸나, 수요는 많음

프로젝트 요약

✓ 결론 도출 과정

- Text data : 네이버 쇼핑의 에어팟 리뷰
- 에어팟 1세대와 2세대의 리뷰 텍스트 마이닝으로 비교 분석

➡ 소비자가 실제 변화된 성능을 체감하는가?

➡ 소비자의 만족도에 반영된 심리는 무엇인가?(보유효과 등)

프로젝트 요약

✓ 연구 과정

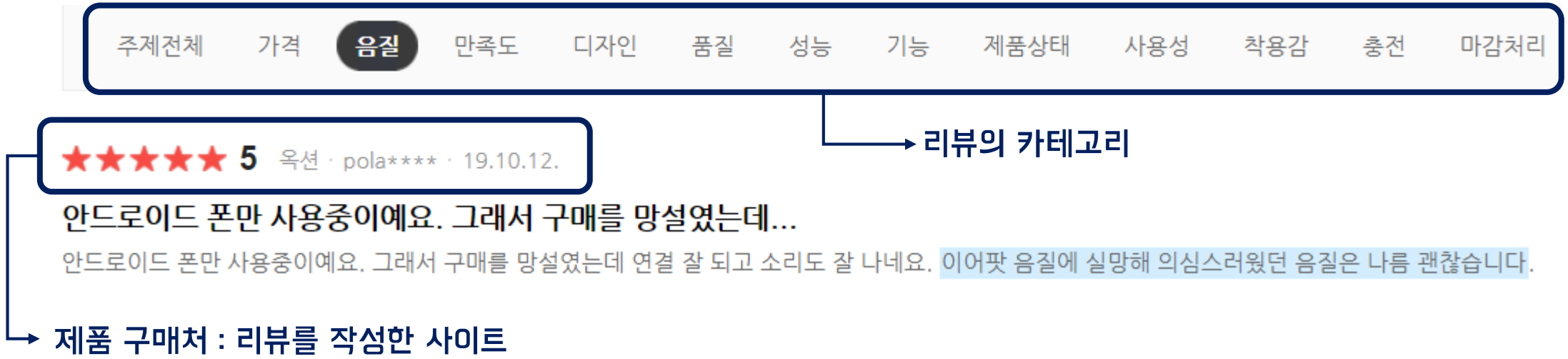
1. **데이터 수집** : 네이버 쇼핑 > 에어팟 1세대, 에어팟 2세대 (유/무선) 리뷰
2. **데이터 전처리** :

   불용어 제거 > 형태소 분석 > 빈도수 분석으로 긍/부정 사전 생성
3. **분석** : 생성된 긍정/부정어 사전으로 감성분석 진행 후 긍정도 변화 비교

데이터 수집

✓ 네이버 쇼핑 리뷰 크롤링

- 네이버 **쇼핑은 '에어팟'을 판매하는 모든 사이트의 리뷰를 볼 수 있음**
- **리뷰의 주제 별로 카테고리화** 되어있음



데이터 수집

✓ 에어팟 1세대 vs 2세대 차이점

W1칩 ➡ **H1칩**

차이 있음	차이 없음
- 페어링 속도, 통화 품질 , 반응속도 개선 - 배터리 개선 - 음성으로 시리 호출 **품질, 성능, 배터리수명, 기능, 조작성**	- 디자인 - 음질 **디자인, 착용감, 휴대성, 음질**

데이터 수집

✓ 데이터 수집 결과

	A	B	C	D
1	type	category	review	star
2	에어팟 1세대	음질	2세대로 배송 받았고 철가루랑 ㅎ	5
3	에어팟 1세대	음질	배송은 4시 이전 주문 당일 발송ㅇ	5
4	에어팟 1세대	음질	아이폰과 아이패드를 둘 다 사용ㅎ	5
5	에어팟 1세대	음질	다른 저렴한 곳도 있었습니다.그ㄷ	4
6	에어팟 1세대	음질	택배도 하루일찍오고 하자 하나도	5

•
•
•

3679	에어팟 2서	착용감	착용감 음질 모든게 너무 다 좋ㅇ	5
3680	에어팟 2서	착용감	카드사할인까지 받아서 싸게 샀ㄷ	5
3681	에어팟 2서	착용감	생각 보다 착용감이 좋습니다. 즐	5
3682	에어팟 2서	착용감	우선 이전에 에어팟을 써본적은	5
3683	에어팟 2서	휴대성	휴대하기 편하고, 애플 기기 사ㅇ	4
3684	에어팟 2서	휴대성	휴대하기도 좋고 음질도 좋은데	4
3685	에어팟 2서	휴대성	디자인도 깔끔하니 세련되고 휴ㄷ	5

type	token_review
에어팟 1세대	1153
에어팟 2세대	2531

전체 리뷰 수 : 3684

데이터 전처리

✓ 전처리 과정 : 불용어 제거

- 특수기호,

  자/모음 제거

```python
#불용어 제거
import re
def clean_str(text):
    pattern = '([a-zA-Z0-9_.+-]+@[a-zA-Z0-9-]+₩.[a-zA-Z0-9-.]+)' # E-mail제거
    text = re.sub(pattern=pattern, repl='', string=text)
    pattern = '(http|ftp|https)://(?:[-₩w.]|(?:%[₩da-fA-F]{2}))+' # URL제거
    text = re.sub(pattern=pattern, repl='', string=text)
    pattern = '([ㄱ-ㅎㅏ-ㅣ]+)'  # 한글 자음, 모음 제거
    text = re.sub(pattern=pattern, repl='', string=text)
    pattern = '<[^>]*>'         # HTML 태그 제거
    text = re.sub(pattern=pattern, repl='', string=text)
    pattern = '[^₩w₩s]'         # 특수기호제거
    text = re.sub(pattern=pattern, repl='', string=text)
    return text

review_=[]
for i in review:  #문자가 들어있을때는 인덱스 사용하면 안됨!!!!
    a=clean_str(i)
    review_.append(a)  #불용어제거한 review 저장
```

```python
review[1] #불용어 제거 전
```

'배송은 4시 이전 주문 당일 발송으로 다음날 바로 받았구요, 바로 정품 등록했는데 이상 없이 잘 됐어요!~연결도 잘되고, 음악 잠깐 들었는데 음질도 좋고 정말 편해요^^미리 구입해둔 케이스도 장착해서 바로 사용했어요~가격도 L.POINT랑 청구할인 받아서 직구보다 저렴하게 잘 구입했습니다.좋은 상품 감사합니다:)'

```python
review_[1]  #불용어 제거 성공(이모티콘 제거됨)
```

'배송은 4시 이전 주문 당일 발송으로 다음날 바로 받았구요 바로 정품 등록했는데 이상 없이 잘 됐어요연결도 잘되고 음악 잠깐 들었는데 음질도 좋고 정말 편해요미리 구입해둔 케이스도 장착해서 바로 사용했어요가격도 LPOINT랑 청구할인 받아서 직구보다 저렴하게 잘 구입했습니다좋은 상품 감사합니다'

데이터 전처리

✓ 전처리 과정 : 형태소 분석

- 리뷰를 단어별로 토큰화 하여

  긍정/부정 단어 매칭에 활용

형태소 분석

불용어 제거된 리뷰(reivew_)를 형태소분석

```
import nltk
from konlpy.tag import Okt; t=Okt()  #오픈 소스 한국어 분석기

#리뷰하나씩 형태소 추출
a= len(review_) #불용어 제거한 리뷰
token_review_list=[]
for i in range(0,a): # 리뷰하나당 처리하기위해 for문
    token_review=t.morphs(review_[i])  #t=형태소분석기, morphs=형태소 추출,review_=전처리한 리뷰,
    #token_review_str=(' '.join(token_review))  #,로 나누어져있는 형태소를 하나의 str로 묶기(df에 넣기위해서)
    #df['token_reivew'] = token_review_str  #이렇게하면 열전체값이 통일됨
    token_review_list.append(token_review) #하나의 리스트를 만들어서 df에 추가해야함.
df['token_review']=token_review_list #형태소단위로 나누어진 리뷰저장
```

p.s) join을 하면 리뷰가 형태소 단위로 찢어진 list에서 다시 string이 되기 때문에 하면안됨

```
df
```

	type	category	review	star	token_review
0	에어팟 1세대	음질	2세대로 배송 받았고 철가루랑 하얀색 실리콘케이스도 잘 받았습니다 역시 사용해보니 ...	5	[2, 세대, 로, 배송, 받았고, 철, 가루, 랑, 하얀색, 실리콘, 케이스, 도...
1	에어팟 1세대	음질	배송은 4시 이전 주문 당일 발송으로 다음날 바로 받았구요, 바로 정품 등록했는데 ...	5	[배송, 은, 4시, 이전, 주문, 당일, 발송, 으로, 다음, 날, 바로, 받았구...
2	에어팟 1세대	음질	아이폰과 아이패드를 둘 다 사용하고 있는데 아이폰에서 음악을 듣거나 영상을 보다 아...	5	[아이폰, 과, 아이패드, 를, 둘, 다, 사용, 하고, 있는데, 아이폰, 에서, ...
3	에어팟 1세대	음질	다른 저렴한 곳도 있었습니다.\n\n그래도 제조일자 최신에 페이백으로 다시 받는 금...	4	[다른, 저렴한, 곳도, 있었습니다, \n\n, 그래도, 제조, 일자, 최신, 에,...
4	에어팟 1세대	음질	택배도 하루일찍오고 하자 하나도 없고 깔끔했어요 뽁뽁이로 포장도 해주셔서 너무 감사...	5	[택배, 도, 하루, 일찍, 오고, 하자, 하나, 도, 없고, 깔끔했어요, 뽁뽁이,...
...	...	...	...	...	...
3781	에어팟 2세대 (무선)	착용감	생각 보다 착용감이 좋습니다. 즐거운 사운드를 느꼈습니다	5	[생각, 보다, 착용, 감, 이, 좋습니다, 즐거운, 사운드, 를, 느꼈습니다]

데이터 전처리

✓ 전처리 과정 : 빈도 분석

- 빈도가 높은 긍정어/부정어를 감성분석 사전에 추가

In [33]:
```
sort=sorted(frequency.items(), key=lambda x: x[1], reverse=True)

```

```
('잘', 1008),
('좋아요', 933),
('좋고', 841),
('너무', 835),
('은', 799),
('에어팟', 689),
('가', 624),
('사용', 528),
('을', 507),
('구매', 452),
('품질', 447),
('좋네요', 439),
('빠르고', 422),
('디자인', 387),
('으로', 378),
('좋습니다', 369),
('가격', 368),
('로', 358),
('보다', 346),
```

데이터 분석

✓ 단어 사전

- 기존 단어 사전에 데이터 빈도분석 결과에 나타난 긍/부정어 추가

긍정어

괜찮구요 돼요 편한 만족합니다 좋습니다 좋네요 좋아요 좋은데 저렴 편리함 재구매 맘에듭니 안심 꼼꼼히 깔끔한 양호 앙증맞고 깨끗한 깔끔해서 편리하네요 감사해요 필수 우수합니다 꼼꼼하게 좋았구요 빵빵 좋을 좋긴 무조건 좋았고 저렴하고 편하구요 짱짱 비싼만큼 세상 당일 싸고 다행히 잘쓰고 놀랐어요 싼 금방 편리해요 대박 감동 예뻐요 깨끗하게 좋고요 되더라구요 빨랐어요 굿굿 깔끔하게 좋아합니다 잘쓸게요 편리 좋으네요 이뻐요 빠름 다행 좋아용 이쁘네요 번창 잘쓰겠습니다 편리합니다 편이 만족스럽네요 사세요 만족스러워요 좋아해요 기대 만족하고 빠르네요 괜찮네요 선명하고 가볍고 좋음 편의 간편하고 안전하게 좋았어요 믿고 편합니다 빨라요 저렴한 굳 깨끗하고 좋구 만족스럽습니다 진작 좋구요 좋아하네요 편하네요 예쁘고 만족해요 편해요 좋 신세계 좋아서 편리하고 깔끔하고 저렴하게 이쁘고 굿 들어요 추천 짱 감사합니다 좋은 편하고 최고 역시

부정어

작살 사지 기스 우려 굳이 이상 아쉽네요 솔직히 끊기는 비싸서 늦게 불편해서 비싸지만 안되서 보다는 당황 소음 가품 청구 없어요 잡음 비싼 불량 하자 현상 끊김 고장 단점 하지만 문제 후회 환불 먼지 묻지마 유격 결함 너무 불량 반품 끊 교환 대응 어떻게 문제 책임 사과 죄송합니다 비정상 이상해서 안된다고 그딴 팔지 힘들어요 병행 불편함을 찝찝하네요 안좋아서 안좋았어요 답답하네요 뽑기운이라 어렵고 심합니다 증상 수리 머리카락 안되네요 아쉽습니다 비꼬고 조롱 속상하네요

데이터 분석

✓ 기초 통계

statistics1 : 기종 별, 리뷰 카테고리 별 리뷰의 개수 ¶

```
1  #긍정도 평균
2  statistics1=pd.pivot_table(df3.drop(df3.columns[[2,4]], axis=1), index = ['type','category'], aggfunc='count')
3  statistics1
```

	변화 있음					변화 없음			
카테고리	기능	품질	배터리 수명	성능	조작성	착용감	디자인	음질	휴대성
1세대	70	116	10	111	45	49	150	586	16
2세대	129	286	14	225	36	97	351	1373	20
총합	199	402	24	336	81	146	401	1959	36

음질의 리뷰수가 가장 많고, 품질,성능,디자인이 그 다음 으로 많음
=> 반복되어 언급되는 카테고리에 관심이 있다고 볼 수 있음.

데이터 분석

✓ 기초 통계

statistics2 : 기종 별 리뷰의 개수 ¶

```
1 statistics2=statistics1.groupby('type').sum()
2 statistics2
```

type	
에어팟 1세대	1153
에어팟 2세대	306
에어팟 2세대(유선)	2225

전체 리뷰 수 : 3864

statistics3 : 기종 별, 별점 별 리뷰의 개수

```
1 statistics3 = pd.pivot_table(df3.drop(df3.columns[[1, 4]], axis=1), index = ['type','star'], aggfunc='count')
2 statistics3
```

		review
type	star	
에어팟 1세대	1	2
	2	1
	3	26
	4	139
	5	985
에어팟 2세대	1	8
	2	1
	3	52
	4	342
	5	2128

대부분의 리뷰 별점 4~5점대 분포
소비자들이 부정적인 리뷰를 잘 남기지 않음을 알 수 있음.

데이터 분석

✓ 감정분석

	type	category	star	token_review	positive
0	에어팟 1세대	음질	5	['2', '세대', '로', '배송', '받았고', '철', '가루', '랑', ...	0.800000
1	에어팟 1세대	음질	5	['배송', '은', '4시', '이전', '주문', '당일', '발송', '으로'...	0.782609
2	에어팟 1세대	음질	5	['아이폰', '과', '아이패드', '를', '둘', '다', '사용', '하고'...	0.705882
3	에어팟 1세대	음질	4	['다른', '저렴한', '곳도', '있었습니다', '\n\n', '그래도', '제...	0.411765
4	에어팟 1세대	음질	5	['택배', '도', '하루', '일찍', '오고', '하자', '하나', '도',...	0.515152
...	...	...	...	...	...
3679	에어팟 2세대	착용감	5	['생각', '보다', '착용', '감', '이', '좋습니다', '즐거운', '사...	1.000000
3680	에어팟 2세대	착용감	5	['우선', '이전', '에', '에어팟', '을', '써', '본적', '은', ...	0.571429
3681	에어팟 2세대	휴대성	4	['휴대', '하기', '편하고', '애플', '기기', '사이', '의', '호환...	0.454545
3682	에어팟 2세대	휴대성	4	['휴대', '하기도', '좋고', '음질', '도', '좋은데', '떨어져서', ...	0.545455
3683	에어팟 2세대	휴대성	5	['디자인', '도', '깔끔하니', '세련되고', '휴대', '성도', '좋아요']	1.000000

3684 rows × 5 columns

```
#5점, 0.8이상 긍정도
five_star=df3[((df3['star']==5)&(df3['positive']>=0.8))][['type','category','star','token_review','positive']]
five_star
```

	type	category	star	token_review	positive
0	에어팟 1세대	음질	5	['2', '세대', '로', '배송', '받았고', '철', '가루', '랑', ...	0.800000
12	에어팟 1세대	음질	5	['좀', '더', '일찍', '살것을', '전', '인강', '용', '으로', ...	1.000000
26	에어팟 1세대	음질	5	['배송', '도', '12일', '만에', '와', '굉장히', '빨랐고', '제...	0.818182
28	에어팟 1세대	음질	5	['얼머전', '에', '에어팟', '구매', '하고', '신세계', '를', '맛...	0.833333
47	에어팟 1세대	음질	5	['주문', '하고', '이틀', '만에', '왔습니다', '배송', '완료', '...	0.833333
...	...	...	...	...	...
3672	에어팟 2세대	착용감	5	['뱅앤올룹슨', 'E', '8', '을', '사용', '했었습니다', 'E', '...	0.800000
3674	에어팟 2세대	착용감	5	['착용', '감도', '좋고', '정품', '이라', '더', '좋아요']	1.000000
3676	에어팟 2세대	착용감	5	['배송', '엄청', '빨리', '왔네요', '무선', '이라', '엄청', '편...	1.000000
3679	에어팟 2세대	착용감	5	['생각', '보다', '착용', '감', '이', '좋습니다', '즐거운', '사...	1.000000
3683	에어팟 2세대	휴대성	5	['디자인', '도', '깔끔하니', '세련되고', '휴대', '성도', '좋아요']	1.000000

별점과 리뷰의 긍정도가 상이한 리뷰는 제외함

Ex) 너무 좋고 만족합니다 => 별점 1점, 긍정도 1점

데이터 분석

✓ 감정분석

```
statistics2_=statistics1_.groupby('type').mean()
statistics2_
```

type	positive
에어팟 1세대	0.914774
에어팟 2세대	0.933005

에어팟 1세대 긍정도 평균 **0.91**
에어팟 2세대 긍정도 평균 **0.93**

=> 총 긍정도는 증가

데이터 분석

✓ 감정분석

type	star	positive
에어팟 1세대	3	0.485714
	4	0.712733
	5	0.946088
에어팟 2세대	1	0.126705
	3	0.513920
	4	0.708207
	5	0.954231

- 별점과 긍정도가 상이한 리뷰 제외한 결과

- **별점이 낮을수록 긍정도가 낮고,**
  별점이 높을 수록 긍정도가 높아지는 것을 확인

데이터 분석

✓ 감정분석 결과

type	category	positive
에어팟 1세대	기능	0.893521
	디자인	0.923919
	배터리 수명	0.878788
	성능	0.905500
	음질	0.912772
	조작성	0.906815
	착용감	0.916037
	품질	0.944222
	휴대성	0.951389
에어팟 2세대	기능	0.933138
	디자인	0.941952
	배터리 수명	0.968750
	성능	0.934699
	음질	0.921734
	조작성	0.894987
	착용감	0.930584
	품질	0.936678
	휴대성	0.934524

변화 있음						변화 없음			
카테고리	기능	품질	배터리 수명	성능	조작성	착용감	디자인	음질	휴대성
1세대	0.89	0.94	0.88	0.90	0.90	0.91	0.92	0.91	0.95
2세대	0.93	0.93	0.97	0.93	0.89	0.93	0.94	0.92	0.93
긍정도	+0.4	-0.1	+0.9	+0.3	-0.1	+0.2	+0.2	+0.1	-0.2

결론

✓ 소비자 심리 분석

변화 있음						변화 없음			
카테고리	기능	품질	배터리 수명	성능	조작성	착용감	디자인	음질	휴대성
긍정도	+0.4	-0.1	+0.9	+0.3	-0.1	+0.2	+0.2	+0.1	-0.2

긍정도 평균 변화량 **+0.28** **+0.08**

1. 전체적으로 에어팟 1세대보다 2세대의 긍정도가 높다.
2. 제품 사양의 개선이 있는 카테고리의 긍정도 평균 변화량이 변화 없는 카테고리보다 3.5배 높았다.

소비자들은 에어팟2세대의 개선된 성능 자체에 만족하여 리뷰의 긍정도가 28% 증가하였다.
하지만 성능이 전혀 변화가 없는 카테고리의 긍정도가 소폭(8%) 증가한 것으로 보아, 어느 정도 보유 효과와 심적 회계원리가 작용하였다고 볼 수 있다.