리뷰 데이터로 진행하는
텍스트 분석

Brightics 서포터즈 2기
조하송

Agenda

01/ 프로젝트 개요

02/ 데이터 소개 및 분석 절차

03/ 탐색적 데이터 및 모델링

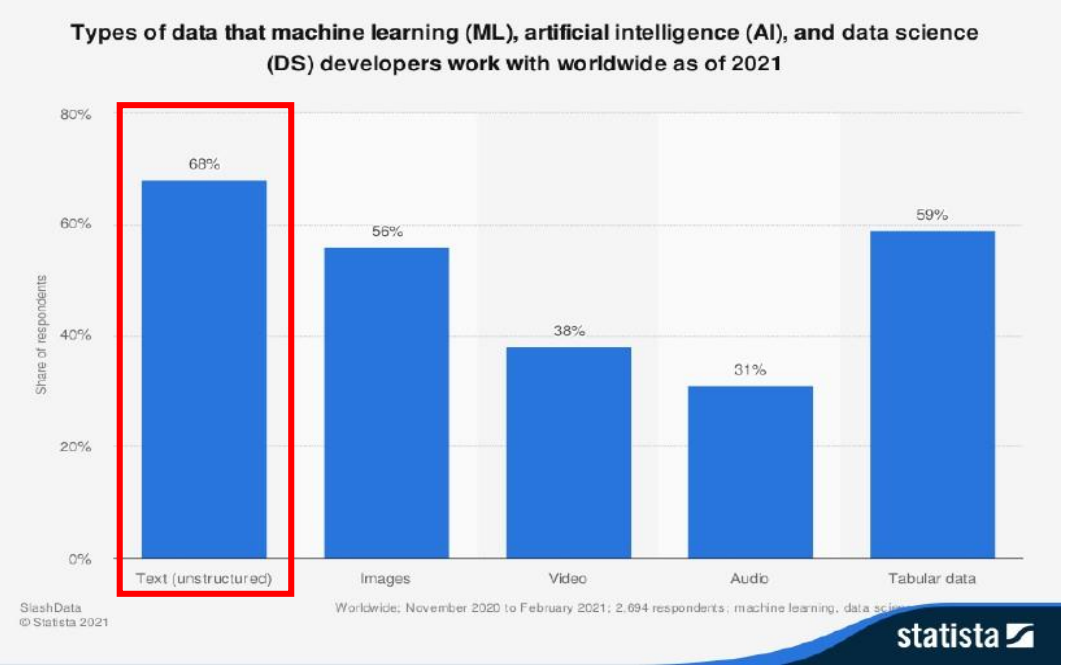
04/ 분석 결과 및 해석

05/ 서포터즈 활동

1. 프로젝트 개요

분석 배경

68%의 ML, AI, DS 개발자는 텍스트 데이터를 사용
(2021년 기준)



분석 주제

텍스트 데이터로 감성 분류 모델 만들기

텍스트 분석	감성 분석
목적에 맞게 텍스트에서 유의미한 정보를 얻어내는 분석 및 처리 과정	텍스트에 나타나는 생각이나 태도, 감성 등의 요소를 분석

1. 프로젝트 개요

분석 목적

의류 쇼핑몰의 상품 후기와 평점 데이터를 활용해 상품에 대한 고객의 긍/부정적인 반응을 예측

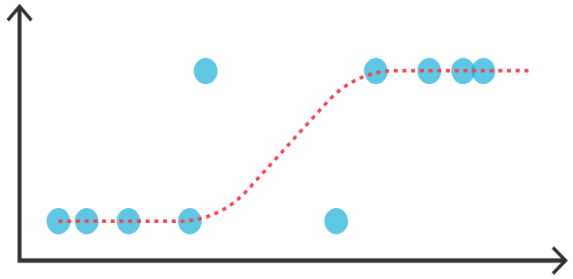
분석 방법

1) 텍스트 분석

상품 후기 데이터로 텍스트 전처리 과정을 통해 분석 성능을 높인다.

2) 로지스틱 회귀

상품 평점 데이터를 레이블로 활용해 고객의 상품에 대한 긍/부정적인 반응을 예측한다.



SAMSUNG SDS

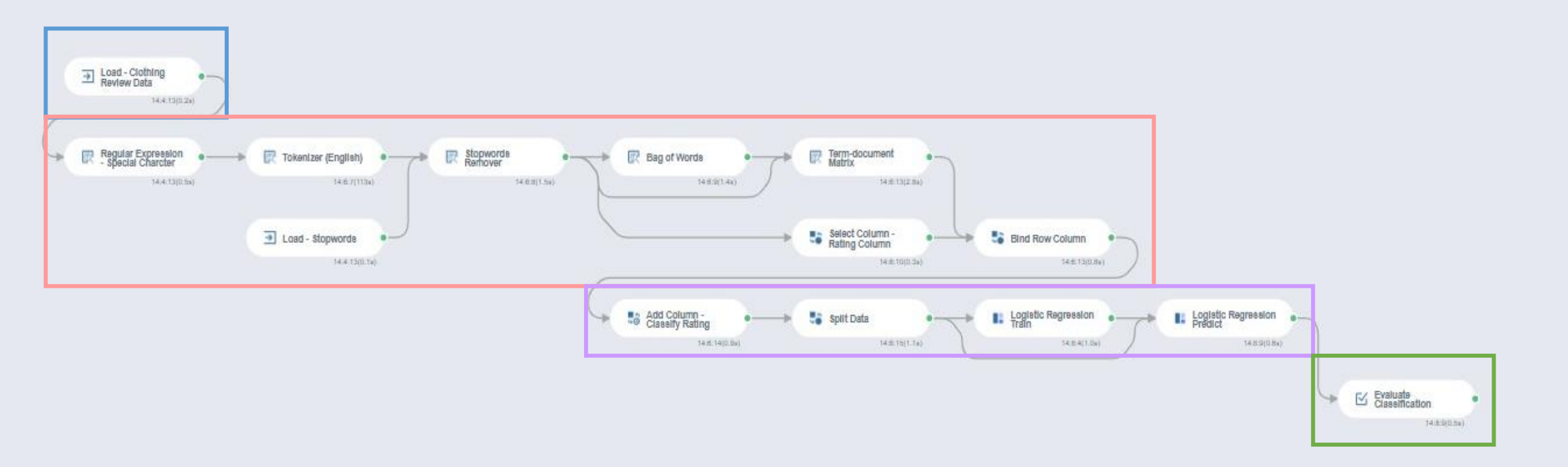
2.1 데이터 소개

	Clothing_ID	Product_Type	Product_Size	Age	Review	Rating
1	767	Intimates	Initmates	33	Absolutely wonderful - silky …	4
2	1080	Dresses	General	34	Love this dress! it's sooo pr…	5
3	1077	Dresses	General	60	Some major design flaws I h…	3
4	1049	Pants	General Petite	50	My favorite buy! I love love l…	5
5	847	Blouses	General	47	Flattering shirt This shirt is v…	5
6	1080	Dresses	General	49	Not for the very petite I love …	2
7	858	Knits	General Petite	39	Cagrcoal shimmer fun I ade…	5
8	858	Knits	General Petite	39	Shimmer surprisingly goes …	4
9	1077	Dresses	General	24	Flattering I love this dress. i …	5
10	1077	Dresses	General	34	Such a fun dress! I'm 5"5' a…	5
11	1077	Dresses	General	53	Dress looks like it's made of…	3
12	1095	Dresses	General Petite	39	This dress is perfection! so …	5
13	1095	Dresses	General Petite	53	Perfect!!! More and more i fi…	5
14	767	Intimates	Initmates	44	Runs big Bought the black x…	5
15	1077	Dresses	General	50	Pretty party dress with som…	3
16	1065	Pants	General	47	Nice but not for my body I to…	4
17	1065	Pants	General	34	You need to be at least aver…	3
18	853	Blouses	General	41	Looks great with white pant…	5
19	1120	Outerwear	General	32	Super cute and cozy A flatte…	5
20	1077	Dresses	General	47	Stylish and comfortable I lov…	5

kaggle

여성 의류 쇼핑몰의 20대 ~ 80 대 사이의 고객이 구매한 제품의 정보, 제품 후기와 평점이 담긴 데이터

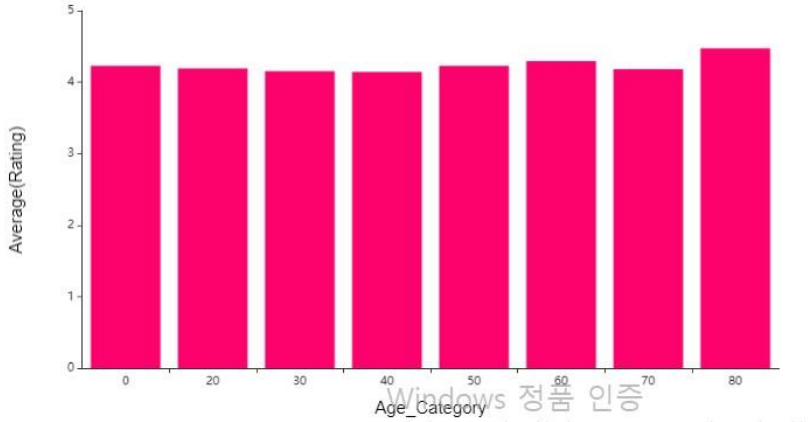
2.2 데이터 분석 절차



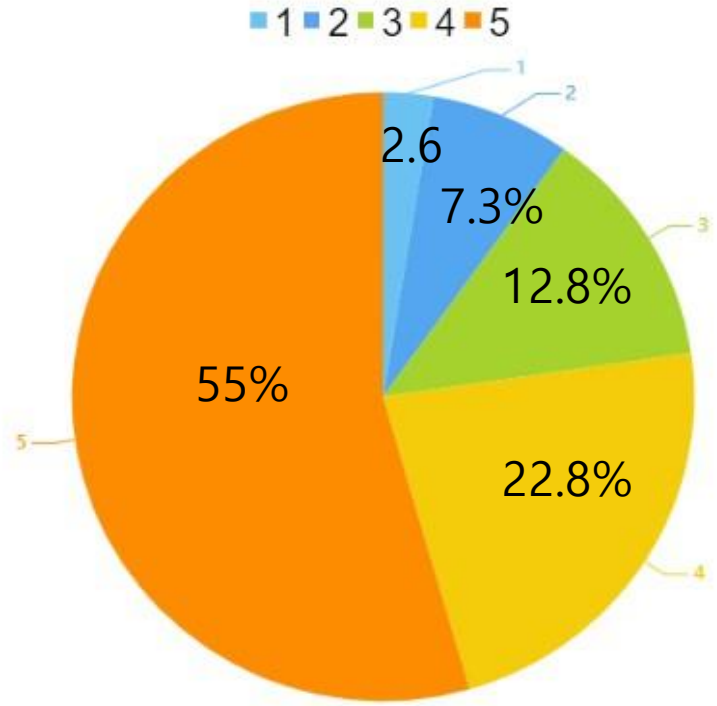
1) 데이터 탐색 | 2) 텍스트 분석 | 3) 분류 알고리즘 | 4) 분류 모델 평가

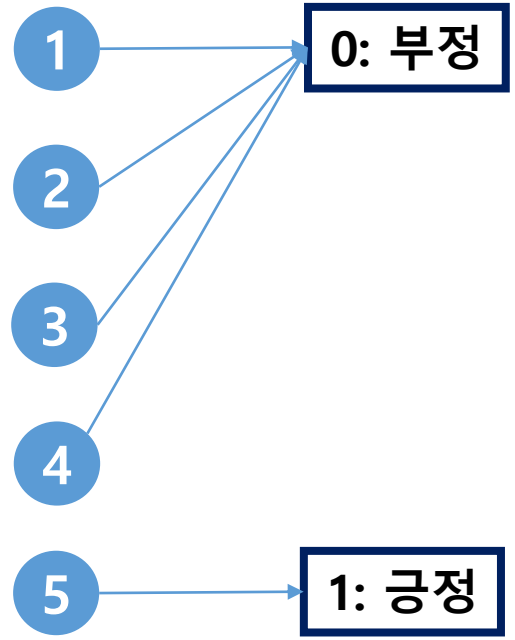
3.1 탐색적 데이터

고객 연령대 vs 평균 평점



상품 평점의 비율





범주형 변수로 변환

3.2 텍스트 분석 절차

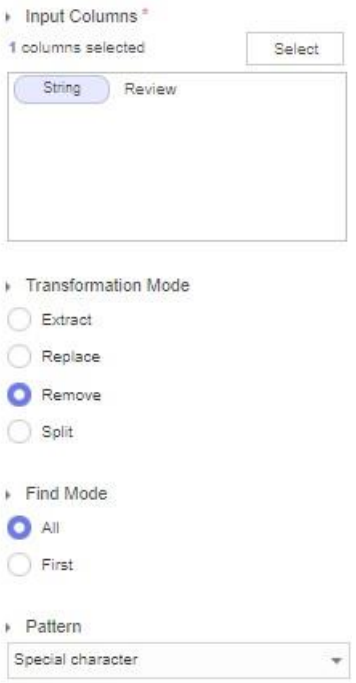
분석 단위로 변환 | 분석 성능 높이는 단계 | 행렬로 표현

문서를 하나의 분석 단위 (Token)로 끊어 내는 작업	텍스트에서 패턴을 찾고 원하는 형태로 문서를 제어	분석 목적에 불필요한 언어를 제거	단어의 순서와 상관없는 빈도수 기반의 단어 표현 방법	문서에 등장하는 단어를 행렬의 형식으로 표현하는 방법

3.3 텍스트 분석 – Regular Expression

Regular Expression: 텍스트에서 패턴을 찾고 원하는 형태로 문서를 제어하는 작업

Before	Regular Expression	After
This dress caught my eye in the store. it's lovely!		This dress caught my eye in the store its lovely
Very cute everyday dress.		Very cute everyday dress
I was looking for something different – it's a bit long!!!		I was looking for something different its a bit long



SAMSUNG SDS

3.4 텍스트 분석 – Tokenization

Tokenization: 문서를 하나의 분석 단위 (Token)로 끊어내는 작업

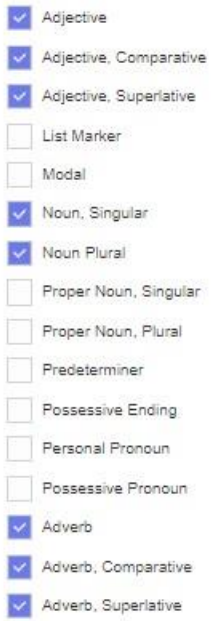
Before

This dress caught my eye in the store its lovely

Very cute everyday dress

I was looking for something different its a bit long

Tokenization



After

dress caught my eye store lovely

very cute everyday dress

i was looking something different bit long

3.5 텍스트 분석 – Stop Words

Stop Words: 분석 목적에 불필요한 언어를 제거하는 작업

Before

dress caught my eye
store lovely

very cute everyday dress

i was looking something
different bit long

Stop Words

	Lists
1	way
2	one
3	bit
4	much
5	also
6	got
7	even
8	tried
9	made
10	get
11	things

After

dress eye
store lovely

cute everyday dress

different long

3.6 텍스트 분석 – Bag of Words

Bag of Words: 단어의 순서와 상관없는 빈도수 기반의 단어 표현 방법

Before

dress eye store lovely cute everyday dress different long

Bag of Words

Minimum Number of Occurrence: 230

	token	document_frequency
1	love	776
2	size	657
3	fit	636
4	dress	622
5	great	621
6	top	552
7	wear	545
8	fabric	417
9	color	408
10	perfect	382
11	small	377
12	cute	366
13	beautiful	361

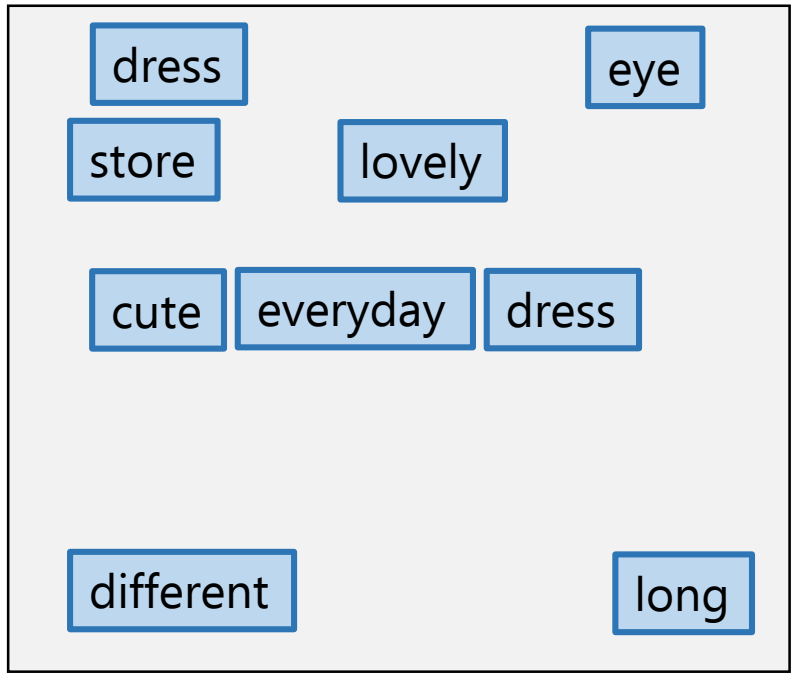
After

dress eye store lovely cute everyday dress different long

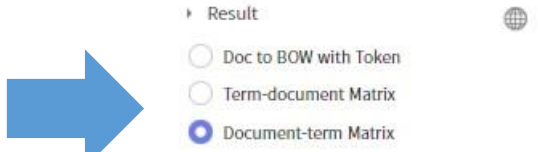
3.7 텍스트 분석 – Document-Term Matrix

Document-Term Matrix: 문서에 등장하는 단어를 행렬의 형식으로 표현하는 방법

Before



Document-Term Matrix



After

Doc#	dress	lovely	cute	…
Doc1	1	1	1	0
Doc2	0	0	1	1
Doc3	0	0	0	1

3.8 머신러닝 – Logistic Regression

로지스틱 회귀로 상품에 대한 고객의 긍/부정적인 반응을 예측

Document – Term Matrix

Doc#	dress	lovely	cute	Rating
Doc1	1	1	1	0
Doc2	0	0	1	1
Doc3	0	0	0	1

Logistic Regression Train

Logistic Regression Predict

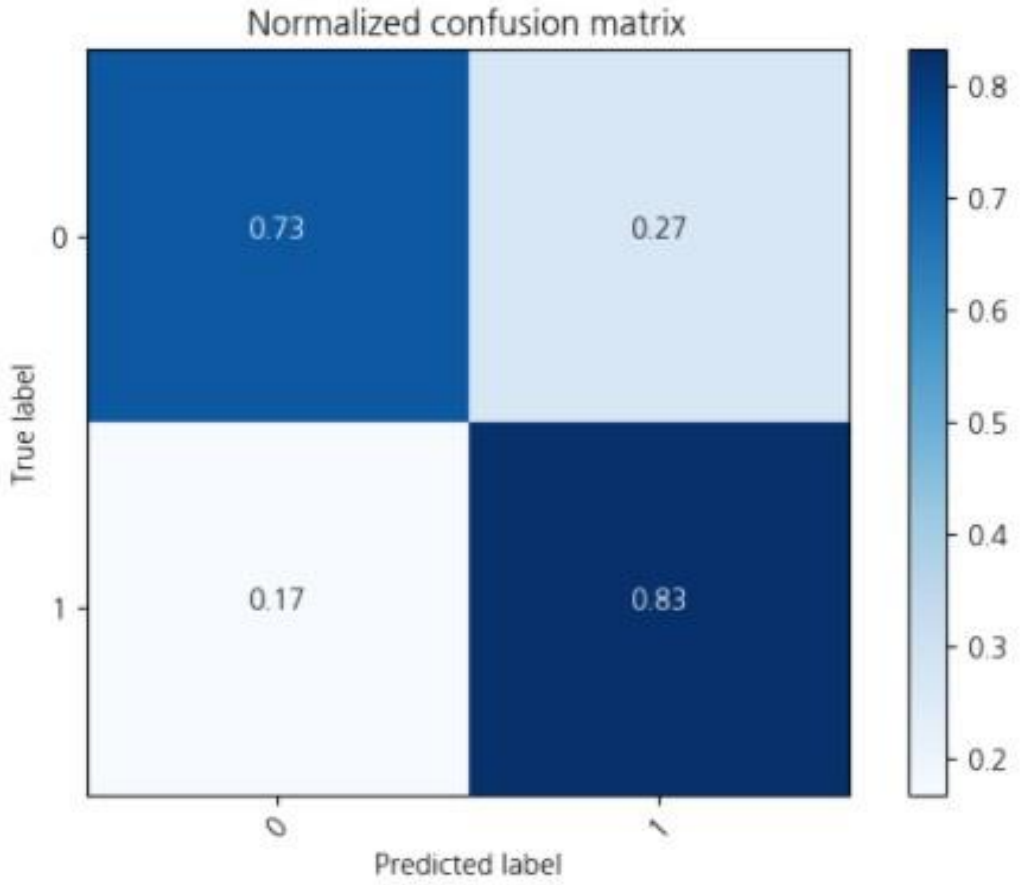
Rating_Pos_Neg	prediction	probability_0	probability_1
1	0	0.606318457955235	0.3936815420447...
1	0	0.507898628533753	0.4921013714662...
0	0	0.707873108421831	0.2921268915781...
1	1	0.4144519889043...	0.5855480110956...
0	0	0.670308041344152	0.3296919586558...
0	0	0.502287962856025	0.4977120371439...
0	1	0.0865165451617...	0.9134834548382...
1	1	0.2481441137735...	0.7518558862264...
0	0	0.9685682099067...	0.0314317900932...
1	1	0.0420869680352...	0.957913031964747
0	0	0.9215722996277...	0.0784277003722...
1	1	0.2898372975519...	0.7101627024480...
1	1	0.2121135594973...	0.7878864405026...
0	0	0.9070732492568...	0.0929267507431...
0	0	0.9372695015410...	0.0627304984589...
1	1	0.1680782227182...	0.8319217772817...
0	1	0.427291703150259	0.572708296849741
1	1	0.2075543664098...	0.7924456335901...
1	1	0.3523020219035...	0.6476979780964...
0	0	0.8864160063951...	0.1135839936048...

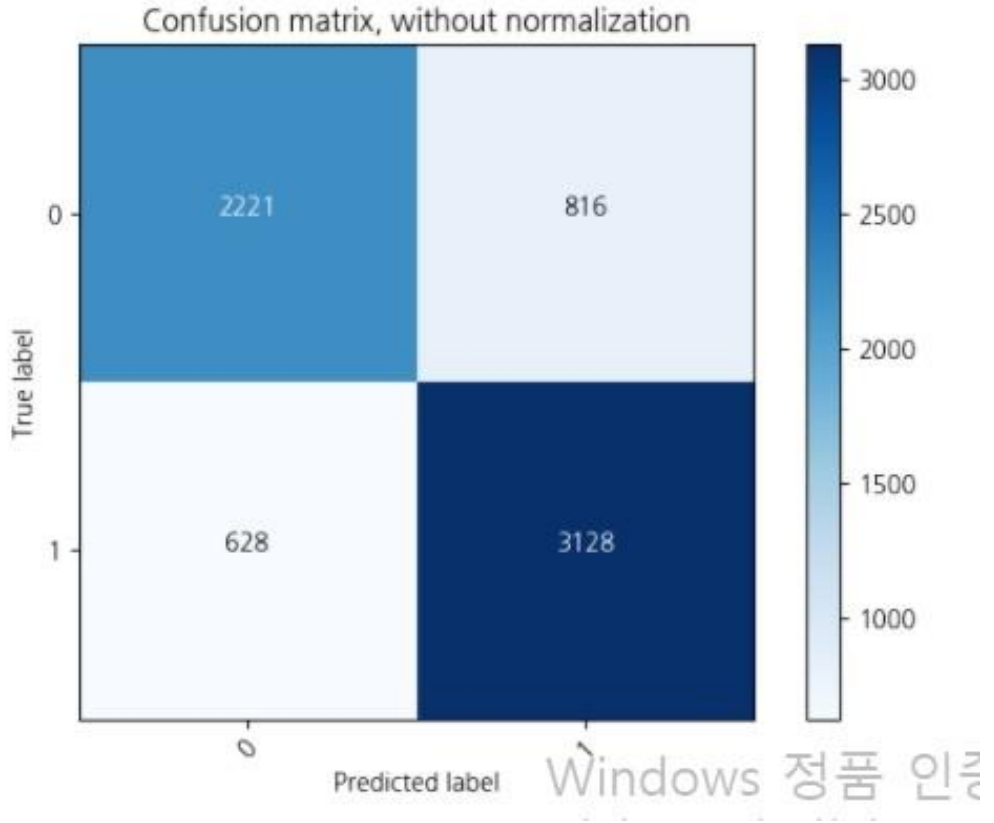
SAMSUNG SDS

4.1 분석 결과

Evaluate Classification Result

Accuracy : 0.7874282349477403





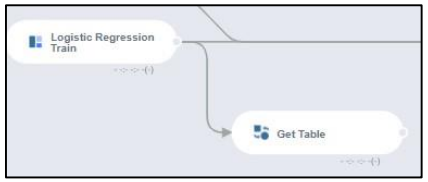
4.2 결과 해석

실제 긍정 → 예측 정확

My favorite buy!
I love love love this jumpsuit.
It's fun flirty and fabulous!
Every time I wear it I get nothing but great compliments!

실제 부정 → 예측 정확

Dress looks like it's made of cheap material.
Dress runs small where the zipper area runs.
I ordered the sp which typically fits me and it was very tight!
The material feels very cheap … disappointed.



*Get Table 함수를 Logistic Regression Train 함수에 연결시키면 모든 Feature의 coefficient를 볼 수 있다.

Word	Coefficient
favorite	1.1
compliments	0.9
love	0.59
fun	0.22

Word	Coefficient
disappointed	-1.41
cheap	-1.08
tight	-0.17
small	-0.07

4.3 분석 개선점

실제 부정 → 예측 오류

The dress is lovely to look at on the hanger and it was the right length for me. Gorgeous on the hanger but not for me. I was so in love with this dress when I saw it in the store but so disappointed when I put it on.

Word	Coefficient
gorgeous	0.66
love	0.59
right	0.15
lovely	0.12
disappointed	-1.41

4.4 분석 효과

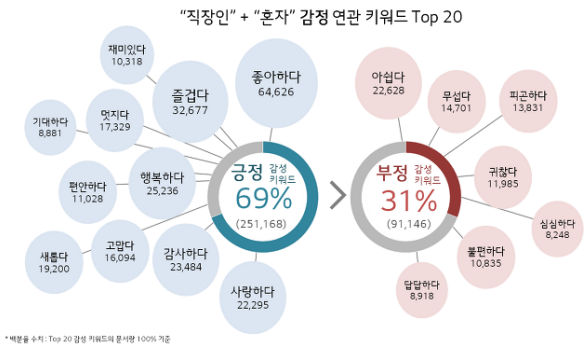
분석 요약 및 효과

텍스트로 고객의 평점을 예측하는 머신러닝 모형 개발

전처리, 모델 고도화 등을 통해 실적용이 가능한 시스템을 개발하고 이를 통해 간편한 형태의 리뷰 시스템을 운용 가능

감성분석의 효과

- 신제품 출시 후 고객반응
- 부정적인 피드백 모니터링을 통한 대응방안
- 실시간 사용자 감성분석
- 상품 트렌드 추적



다양한 데이터 활용 가능

- 영화나 호텔 후기와 같은 데이터 적용 가능

Dataset

Amazon Fine Food Reviews

Dataset

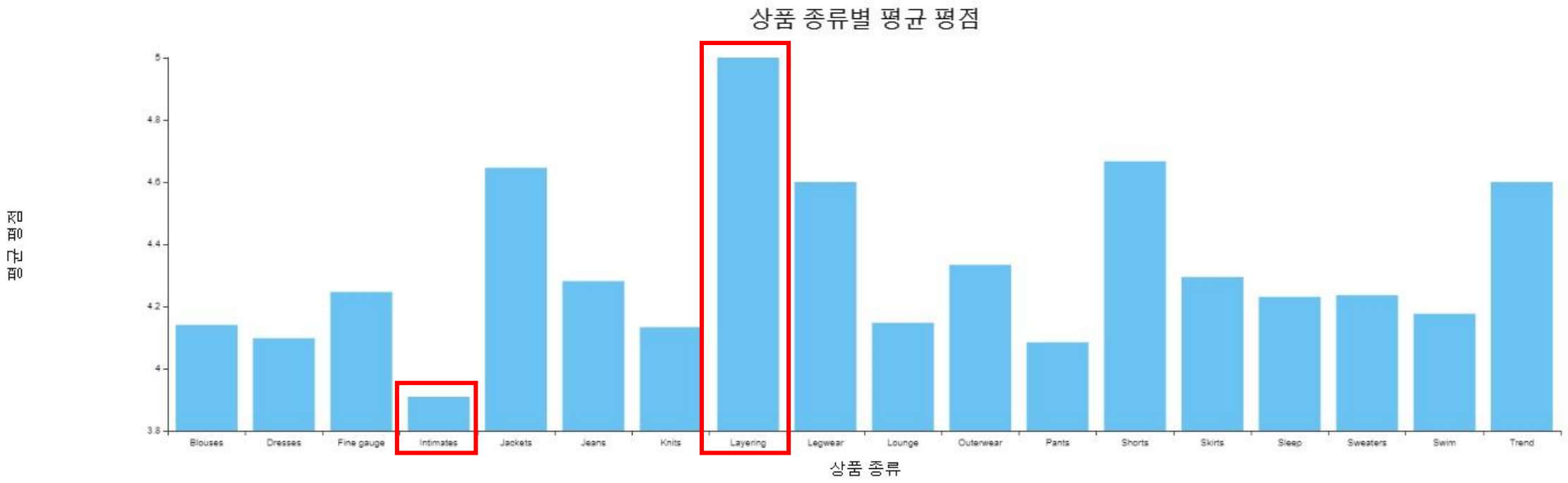
Disneyland Reviews

Dataset

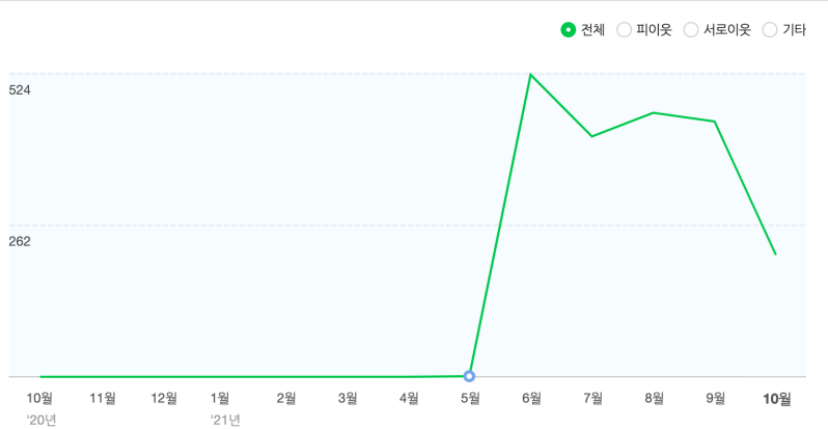
Hotel Reviews

SAMSUNG SDS

4.5 추가 주제


상품 종류별 평균 평점
5
4.8
4.6
4.4
4.2
4
3.8
Blouses
Dresses
Fine gauge
Intimates
Jackets
Jeans
Knits
Layering
Legwear
Lounge
Outerwear
Pants
Shorts
Skirts
Sleep
Sweaters
Swim
Trend
상품 종류

5. 서포터즈 활동 코멘터리



기간	전체	피이웃	서로이웃	기타
2021.10. 월간	211	6	45	160
2021.09. 월간	442	4	55	383
2021.08. 월간	457	4	78	375
2021.07. 월간	416	13	143	260
2021.06. 월간	523	81	42	400

데이터만 있으면 원하는 대로
분석해주는 카페가 있다?...

SAMSUNG SDS

감사합니다