Text Analysis with
Product Reviews

Brightics Ambassador
Hasong Cho
2022.01.14

Agenda

01/ Project Overview

02/ Data Summary & Process

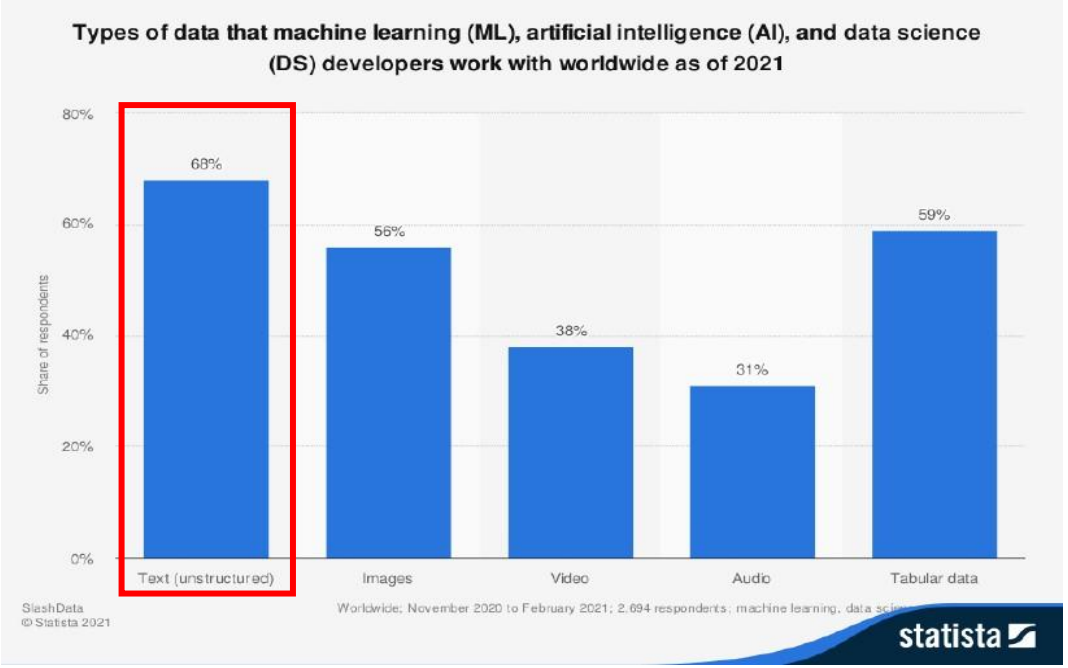
03/ Exploratory Analysis & Modeling

04/ Result & Evaluations

1. Project Overview

Project Background

68% of ML, AI, and DS developers use text data
(As of 2021)



Project Topic

Create a Sentiment Classification Model with Text Data

Text Analysis	Sentiment Analysis
Analyzing and processing process to obtain meaningful information from text for the purpose	Analyze elements such as thoughts, attitudes, and emotions that appear in the text

1. Project Overview

Project Purpose

Predict customers' positive/negative reactions to products by using product reviews and rating data from e-commerce website

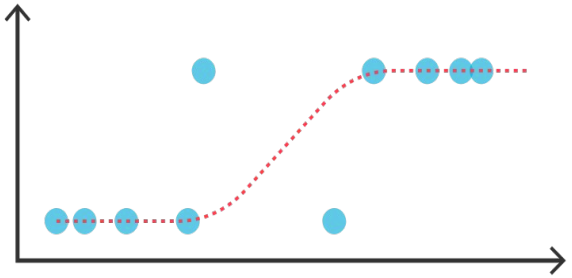
Analysis method

1) Text Analysis

Improve analysis performance through text pre-processing with product review data

2) Logistic Regression

Using product rating data as a label, predict positive/negative reactions of customers to products.



SAMSUNG SDS

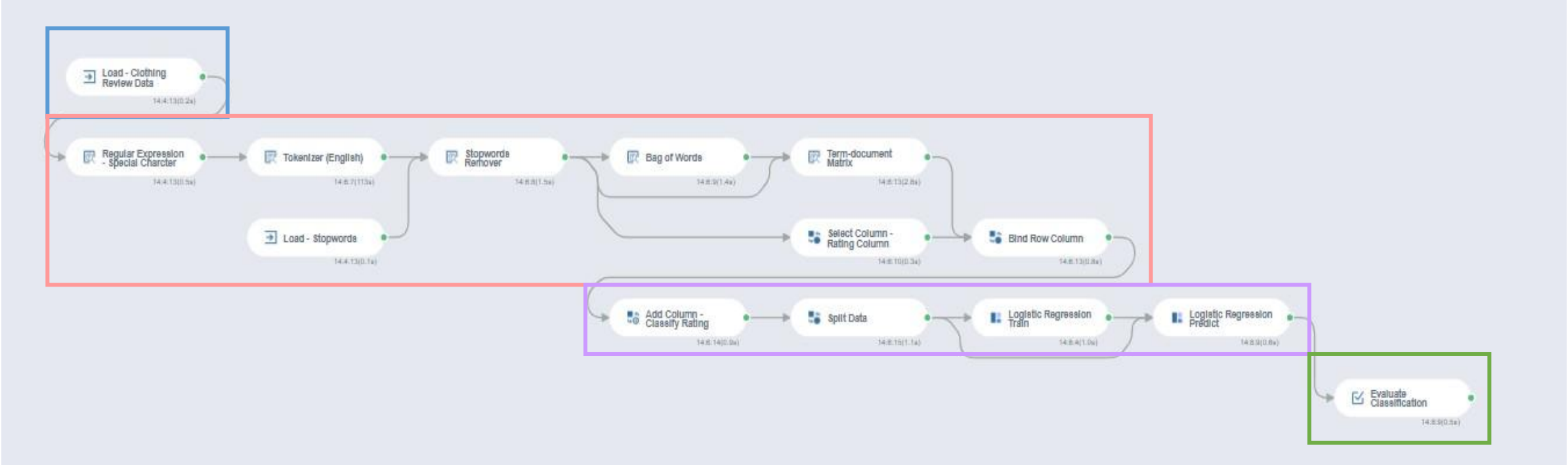
2.1 Data Overview

	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

Data contains product information, product reviews, and ratings purchased by customers in their 20s to 80s at an online women's clothing store

2.2 Analysis Process

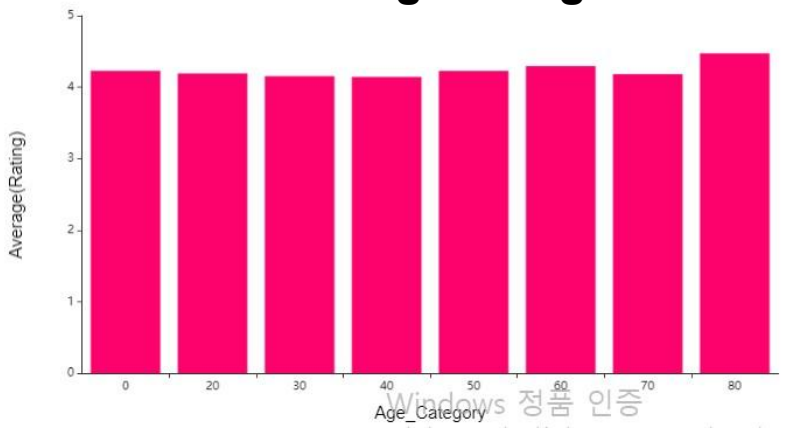


1) Exploratory Analysis → 2) Text Analysis → 3) Classification Model → 4) Evaluation

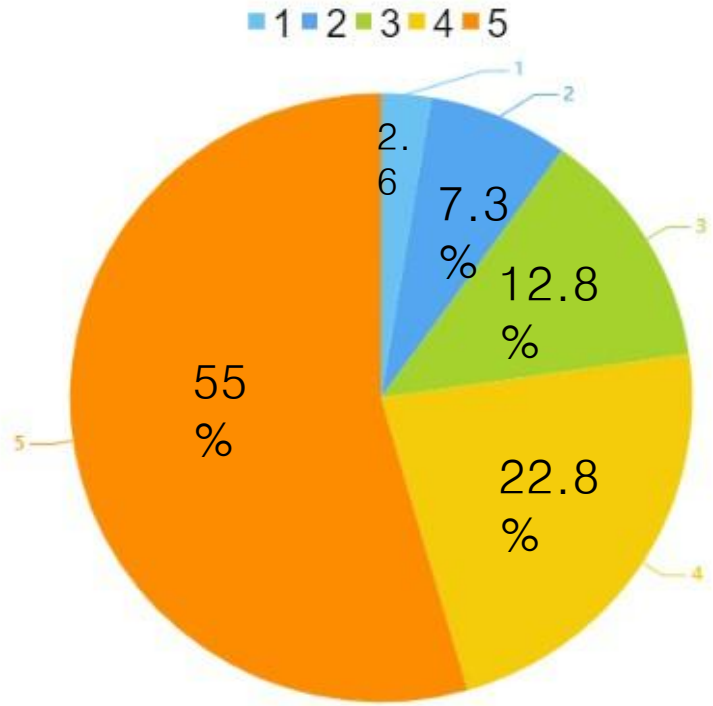
SAMSUNG SDS

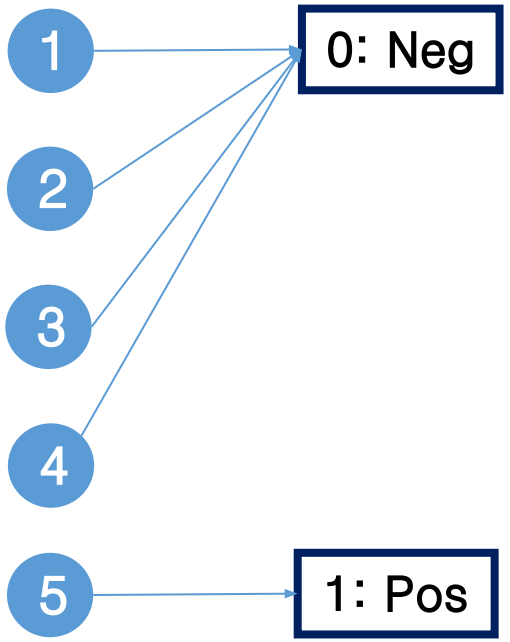
3.1 Exploratory Analysis

Customer age group vs average rating



Percentage of Product Ratings





Convert to categorical variable

3.2 Text Analysis Process

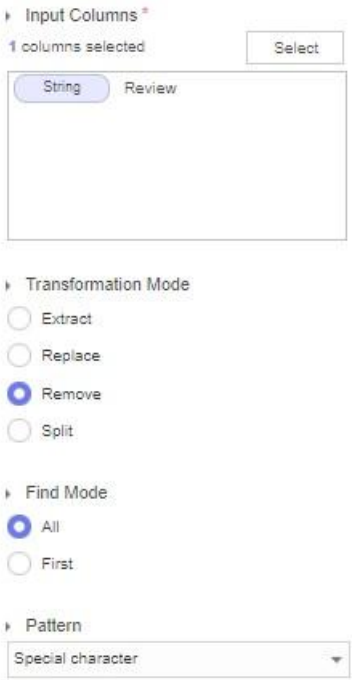
Convert to unit of analysis → **Steps to improve analytical performance** → **express as a matrix**

Tokenizer (English)	Regular Expression	Stopwords Remover	Bag of Words	Term-document Matrix
The task of breaking a document into one unit of analysis (Token)	Find patterns in text and control your document the way you want it	Eliminate unnecessary language for analysis purposes	A word representation method based on frequency counts regardless of order	A way to represent words appearing in a document in the form of a matrix

3.3 Text Analysis– Regular Expression

Regular Expression:
The task of finding patterns in text and controlling the document in the desired form

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 Text Analysis – Tokenization

Tokenization:
The task of breaking a document into one unit of analysis (Token)

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

- [x] Adjective
- [x] Adjective, Comparative
- [x] Adjective, Superlative
- [ ] List Marker
- [ ] Modal
- [x] Noun, Singular
- [x] Noun Plural
- [ ] Proper Noun, Singular
- [ ] Proper Noun, Plural
- [ ] Predeterminer
- [ ] Possessive Ending
- [ ] Personal Pronoun
- [ ] Possessive Pronoun
- [x] Adverb
- [x] Adverb, Comparative
- [x] Adverb, Superlative

After

dress caught my eye store lovely

very cute everyday dress

i was looking something different bit long

3.5 Text Analysis – Stop Words

Stop Words:
The task of removing language unnecessary for analysis purposes.

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 Text Analysis – Bag of Words

Bag of Words:
A word representation method based on frequency counts regardless of word order

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

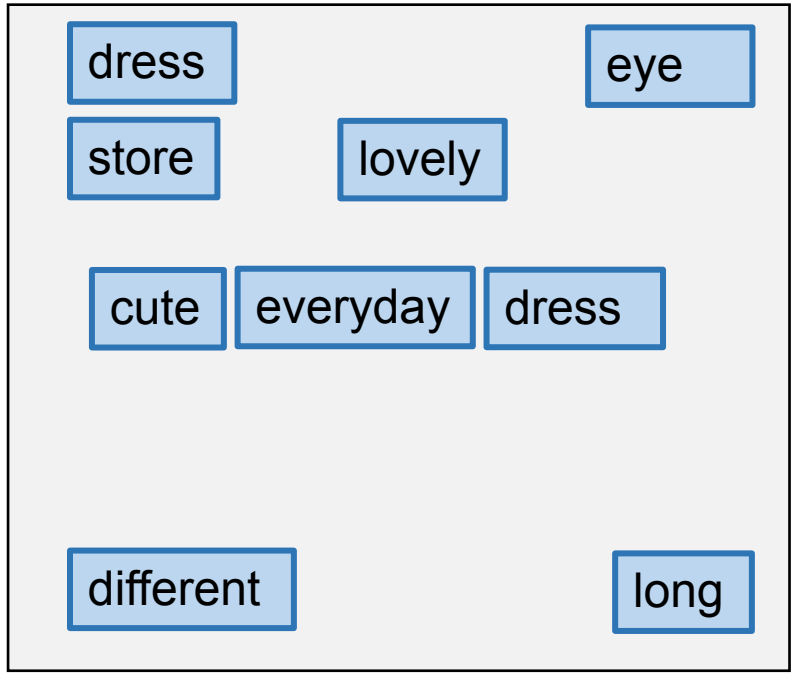
After

dress
eye
store
lovely
cute
everyday
dress
different
long

3.7 Text Analysis – Document-Term Matrix

Document-Term Matrix:
A method of expressing words appearing in a document in the form of a matrix

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 Machine Learning – Logistic Regression

Use logistic regression to predict customer's positive/negative reaction to a product

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

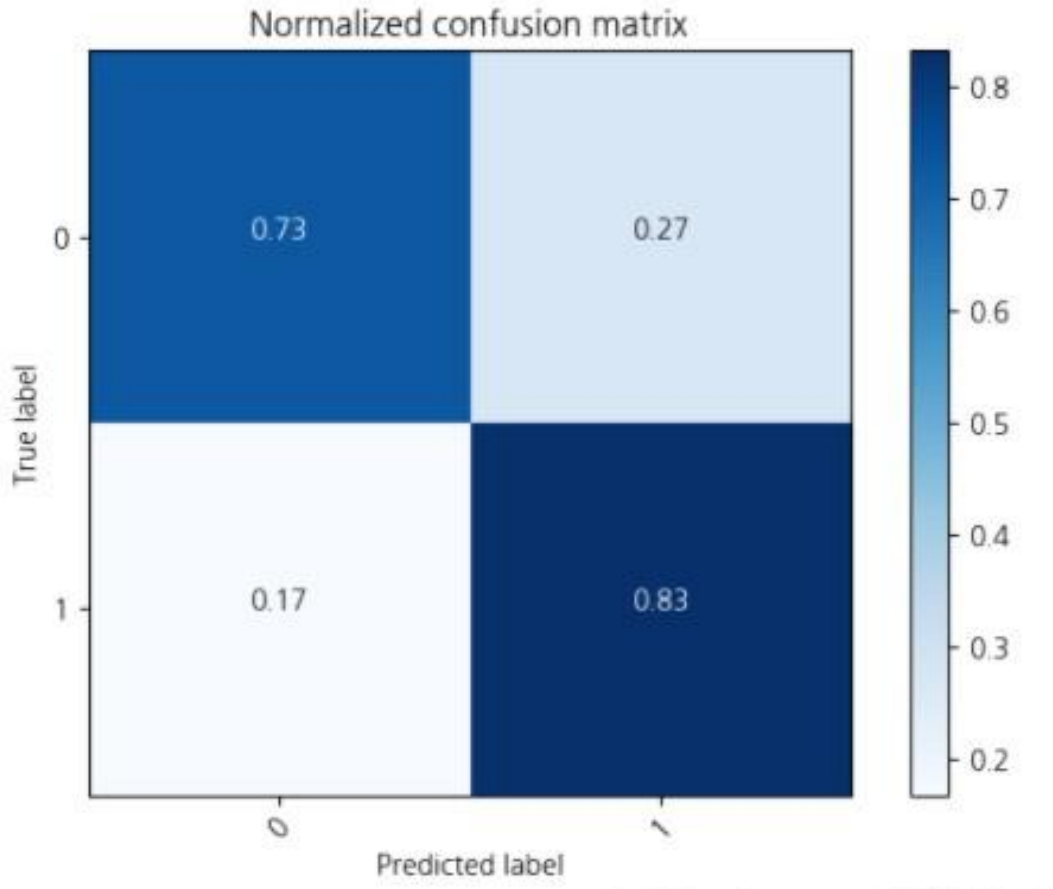
Logistic Regression Predict

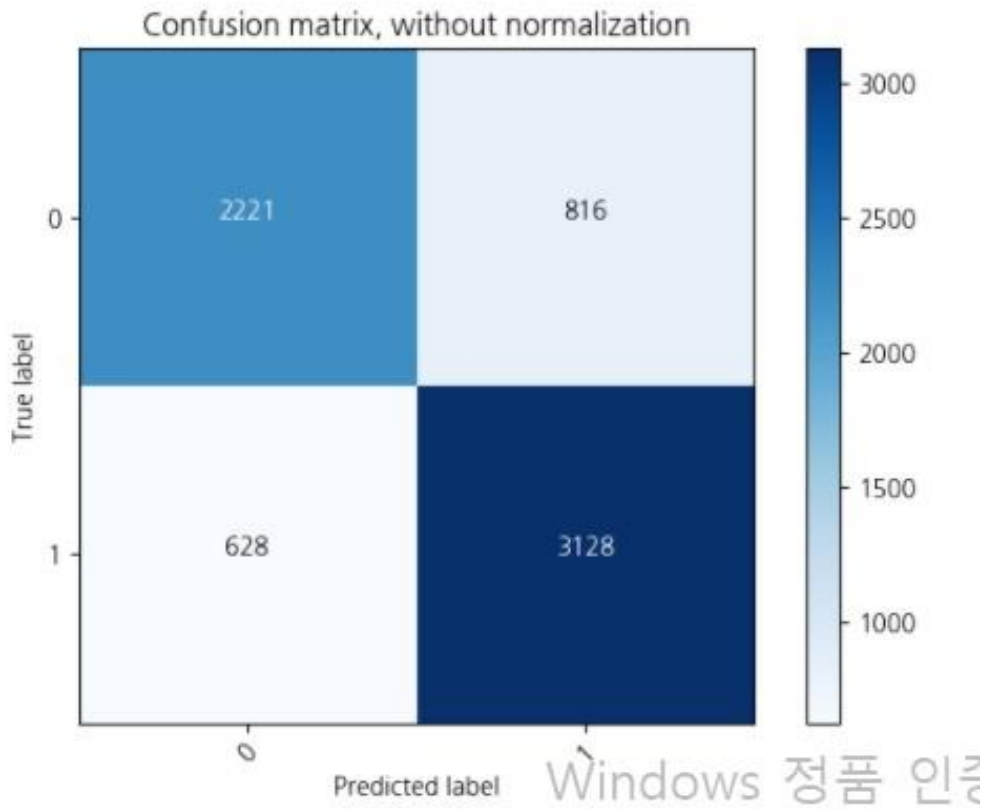
Rating_Pos_Neg	prediction	probability_0	probability_1
1	0	0.606318457955235	0.3936815420447...
1	0	0.507898628533753	0.4921013714662...
0	0	0.707873108421831	0.2921268915781...
1	1	0.4144519889043...	0.5855480110956...
0	0	0.670308041344152	0.3296919586558...
0	0	0.502287962856025	0.4977120371439...
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...
0	0	0.9545023921841...	0.0484976078188...

4.1 Evaluation

Evaluate Classification Result

Accuracy : 0.7874282349477403





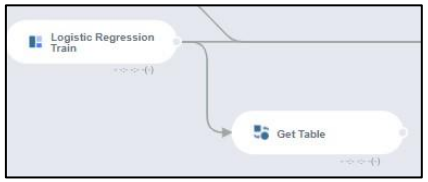
4.2 Result Interpretation

Actual Pos → Correctly Predicted

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!

Actual Neg → Correctly Predicted

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.



* By connecting the Get Table function to the Logistic Regression Train function, you can see the coefficients of all features

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 Further Improvements

Actual Neg → Incorrectly Predicted

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

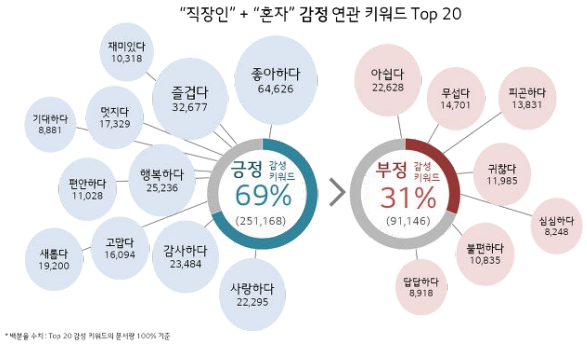
4.4 Implications

Analysis Summary and Effects

- **Summary:** Developed a machine learning model to predict customer ratings from text data
- **Effects:** It is possible to develop a system that can be used for practical purposes through preprocessing, model advancement, etc., and a simple review system can be operated.

Implications

- Customer response after new product launch
- Countermeasures through Negative Feedback Monitoring
- Real-time user sentiment analysis
- Product trend tracking



Various data available

- Data such as movie reviews or hotel reviews can be applied

Dataset

Amazon Fine Food Reviews

Dataset

Disneyland Reviews

Dataset

Hotel Reviews

Thank You!