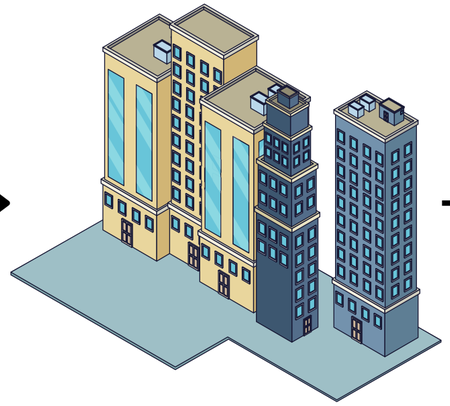
Final Project Text Mining Presentation

Developing an Enhanced Recommendation System for Women's Clothing in E-Commerce: A Comparative Analysis of Traditional Machine Learning and Deep Learning Approaches

Presented by Still Learning Team
Rezky Agung Ardiansyah
Fairuuz Nurdiaz Amanullah

Customer Review



women clothing e-commerce

Understanding customer sentiments

E Commerce Decision Making

10 Features

- Numerical
- Categorical
- Text

How to Handle multiple features with different types of data in sentiment analysis?

Let a look on the of the raw data

	Unnamed: 0	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name
0	0	33	NaN	Absolutely wonderful - silky and sexy and comf...	4	1	0	Initmates	Intimate	Intimates
1	1	34	NaN	Love this dress! it's sooo pretty. i happene...	5	1	4	General	Dresses	Dresses
2	2	60	Some major design flaws	I had such high hopes for this dress and reall...	3	0	0	General	Dresses	Dresses
3	3	50	My favorite buy!	I love, love, love this jumpsuit. it's fun, fl...	5	1	0	General Petite	Bottoms	Pants
4	4	47	Flattering shirt	This shirt is very flattering to all due to th...	5	1	6	General	Tops	Blouses
...	...	...	...	...	...	...	...	...	...	...
23481	23481	34	Great dress for many occasions	I was very happy to snag this dress at such a ...	5	1	0	General Petite	Dresses	Dresses
23482	23482	48	Wish it was made of cotton	It reminds me of maternity clothes. soft, stre...	3	1	0	General Petite	Tops	Knits
23483	23483	31	Cute, but see through	This fit well, but the top was very see throug...	3	0	1	General Petite	Dresses	Dresses
23484	23484	28	Very cute dress, perfect for summer parties an...	I bought this dress for a wedding i have this ...	3	1	2	General	Dresses	Dresses
23485	23485	52	Please make more like this one!	This dress in a lovely platinum is feminine an...	5	1	22	General Petite	Dresses	Dresses

23486 rows × 10 columns

Handling the missing value

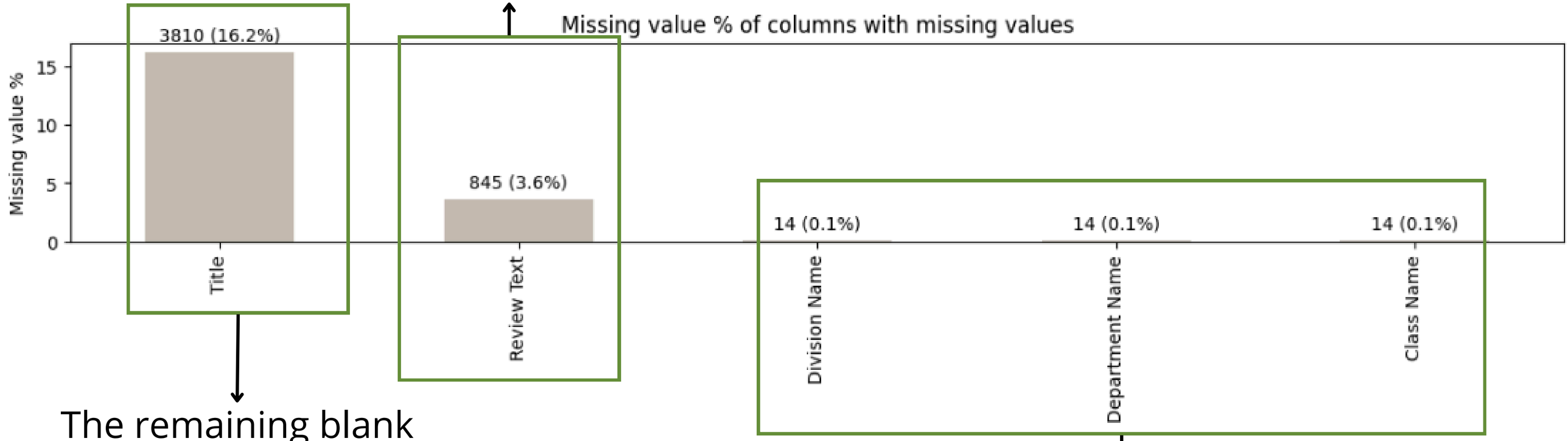
```
df_full.isna().sum()
```

```
Unnamed: 0                    0
Age                           0
Title                      3810
Review Text                 845
Rating                        0
Recommended IND               0
Positive Feedback Count       0
Division Name                14
Department Name              14
Class Name                   14
words count                   0
dtype: int64
```

To handle the missing value, below are the list that we do

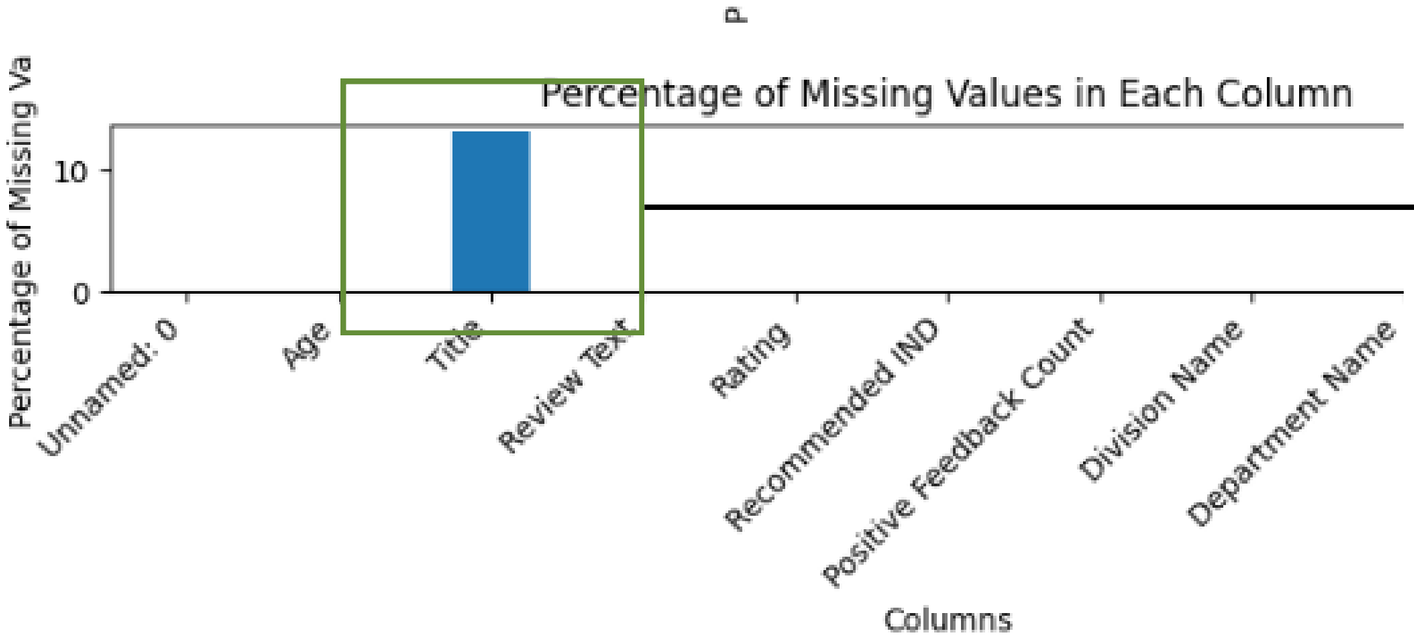
- before handling missing value we perform punctuation removal, and lowercasing to the each data
- For blank Review text we just drop it, then we found a fact that if the reviewe text blank the title also blank we just drop it because our main feature is text features
- The other missing value titles fill by extracted the only adjective of review text
- The information of product Division, Department and Class Name just dropped out

From here we tried to remove blank review text because it’s our main topic is about text



The remaining blank Title obtain by Extract Adjective from Review Text

Small blank data, removing this data might not affect our Data: Remove

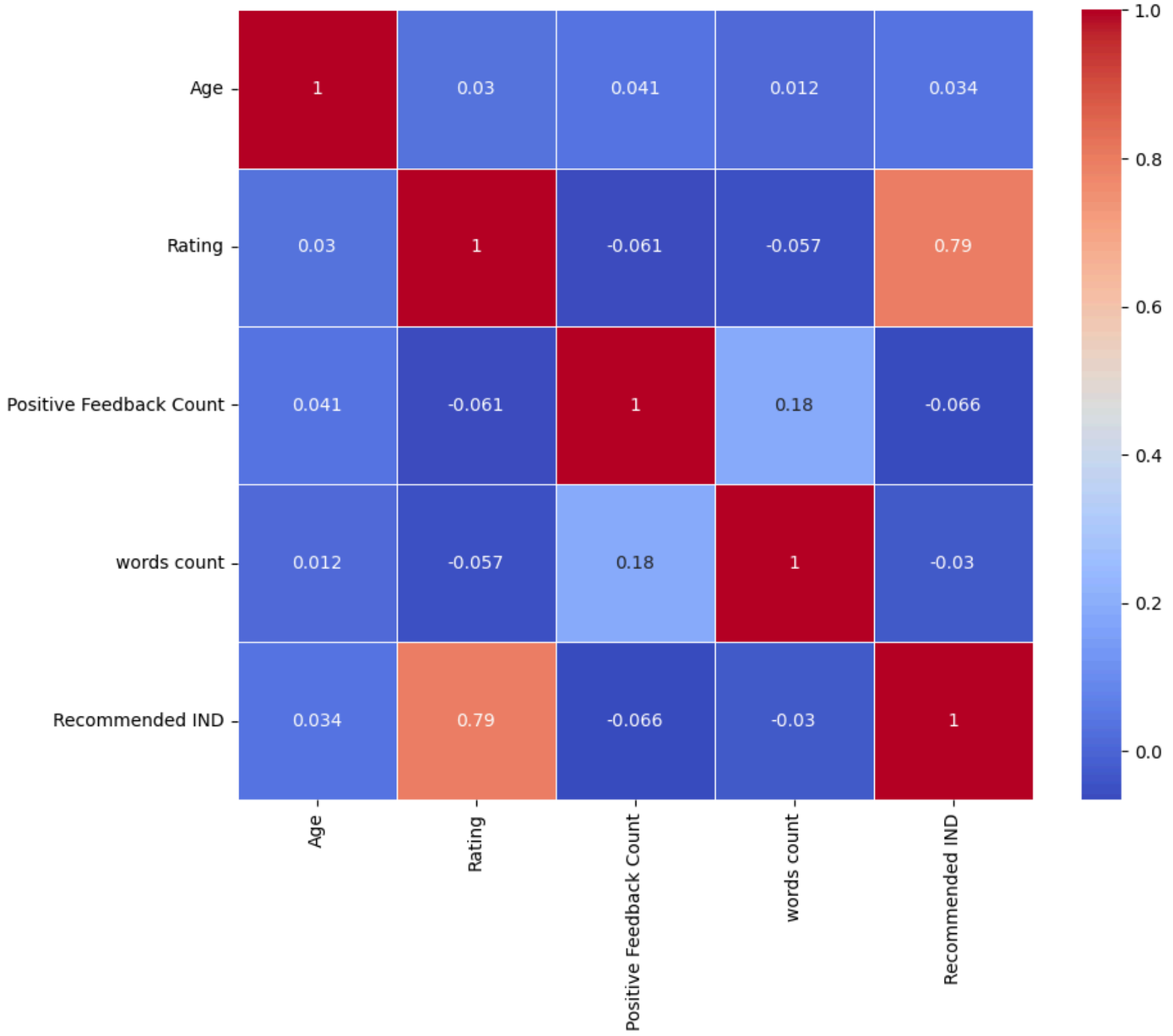


The remaining blank Title obtain by Extract Adjective from Review Text

	Unnamed: 0	Age	Title	Review Text	F
0	0	33	NaN	Absolutely wonderful - silky and sexy and comf...	

Review Text	**Absolutely wonderful – silky and sexy and comfortable**
Title (After Extrax Adjextive)	**wonderful sexy comfortable**

Selection Features

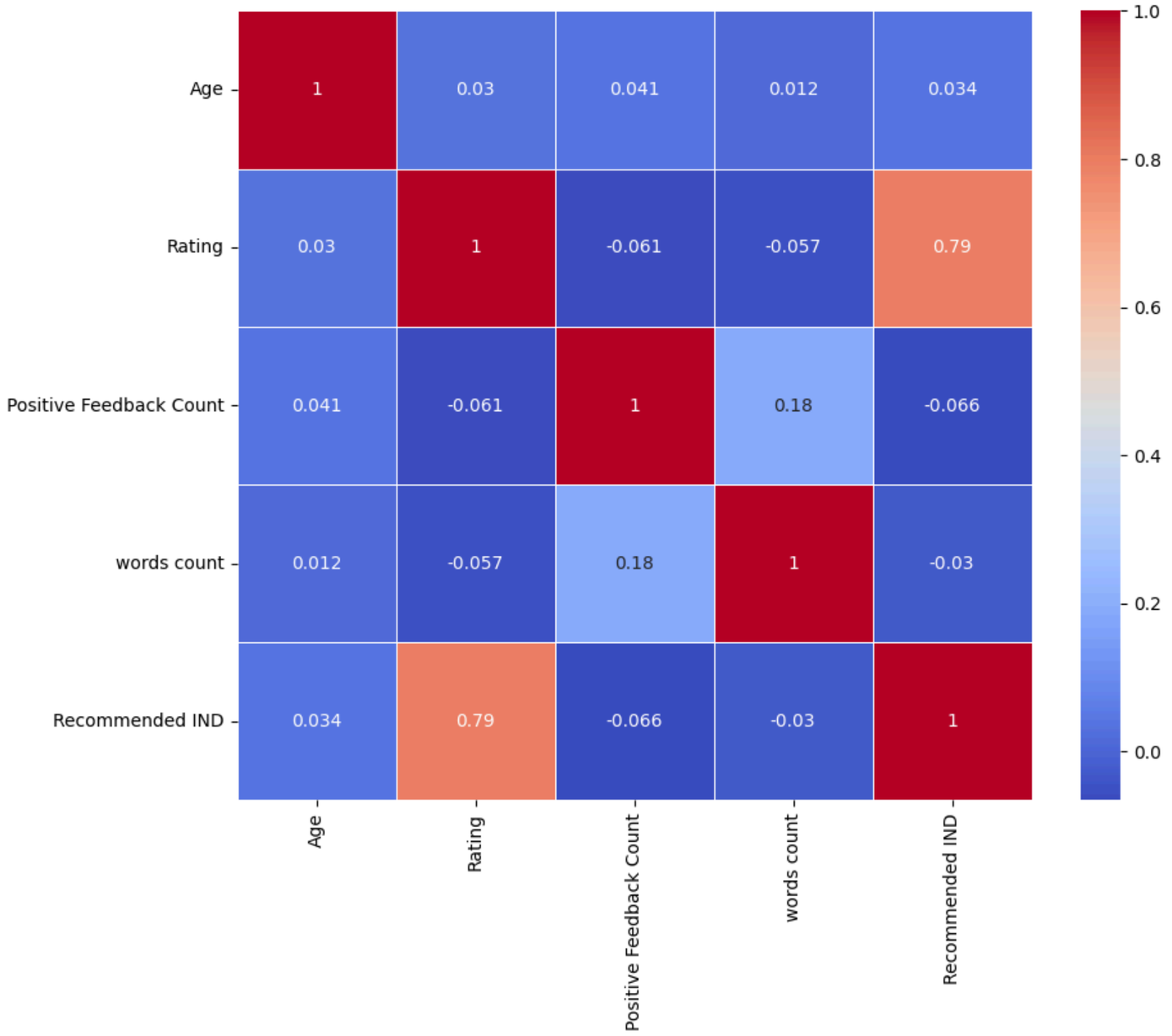


Of the several features above, Recommended IND has a better correlation than other features

Apart from that, to make it easier to make decisions, Recommended IND is better as a label than the others

we take Recomended IND as labels

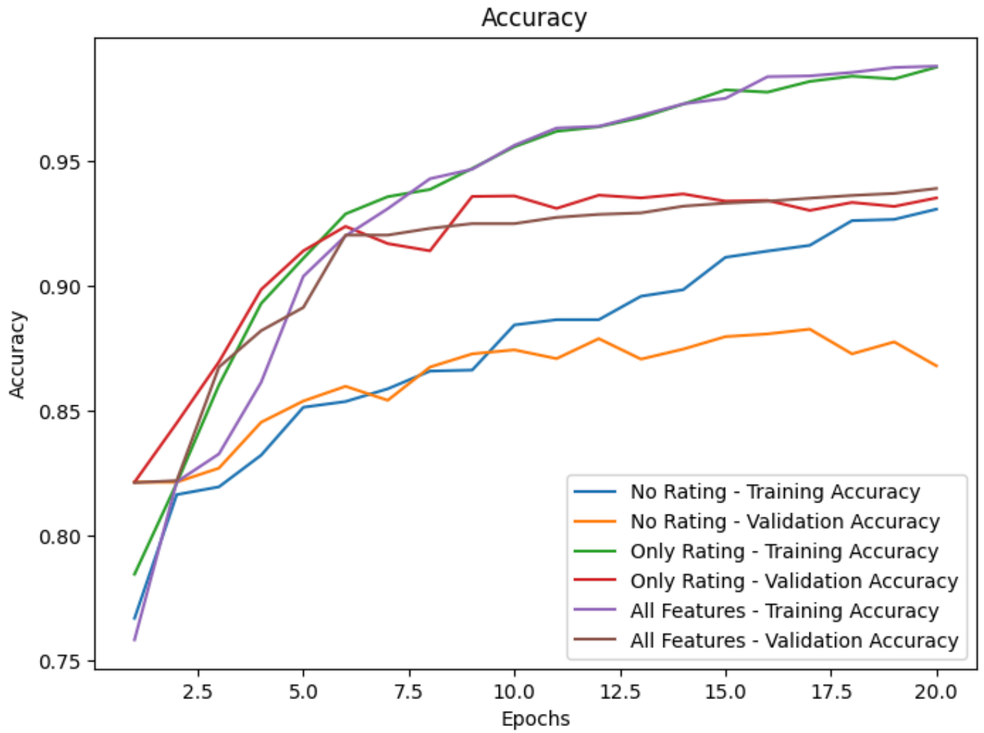
Selection Features



Rating and Recommended IND have a high correlation score. However, the other features show low positive or negative correlations.

Therefore, we only need to use **Rating** and **Recommended IND** as features. Additionally, age will not be considered as a feature because whether a product is recommended or not does not depend on age.

Selection Features

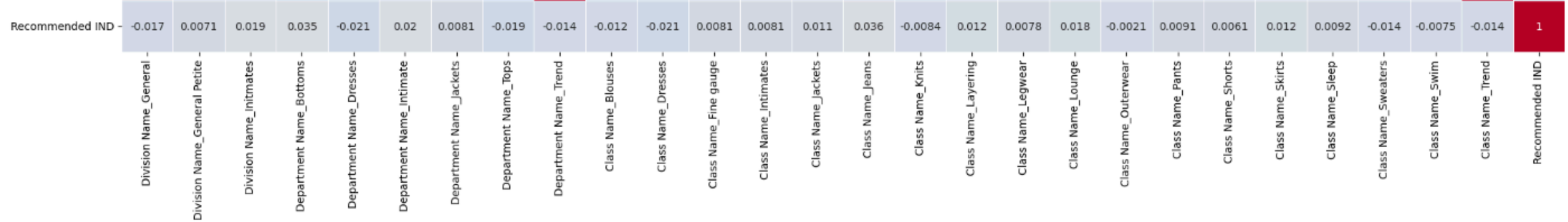


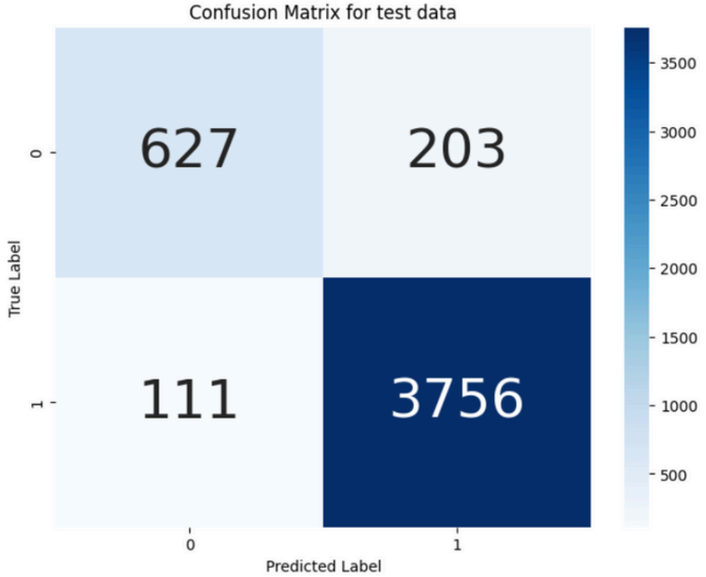
Features	**Loss**	**F1-Score**
Without Rating	0.4366	0.8637
Only Rating	0.2173	0.9367
All Features	0.2052	0.9369

Review	Rating
I wanted to love this. however, the fit was funky and the colors were muted. this is definitely something you must try on in the store. too risky to buy online due to weird fit	2
	R_IND 0

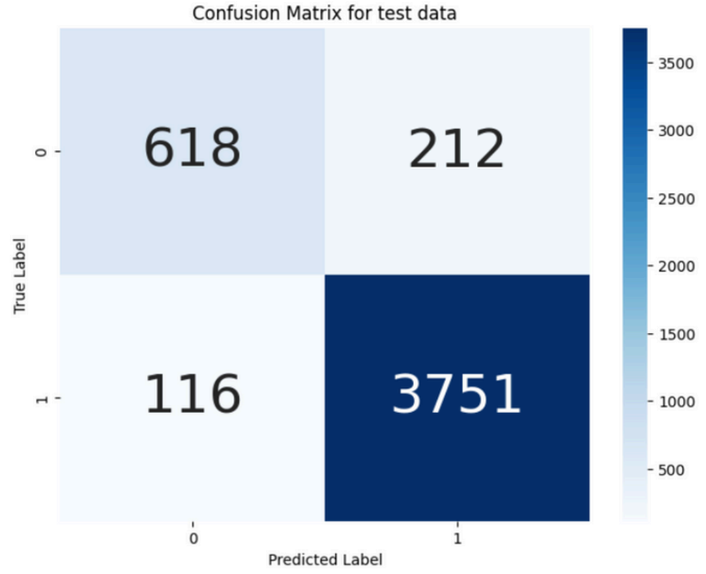
Features	R_IND Pred
Without Rating	1
Only Rating	0
All Features	0

Features Ratings are very important to use in the following sentiment analysis, because positive and negative reviews are somewhat confusing if they are not given a rating





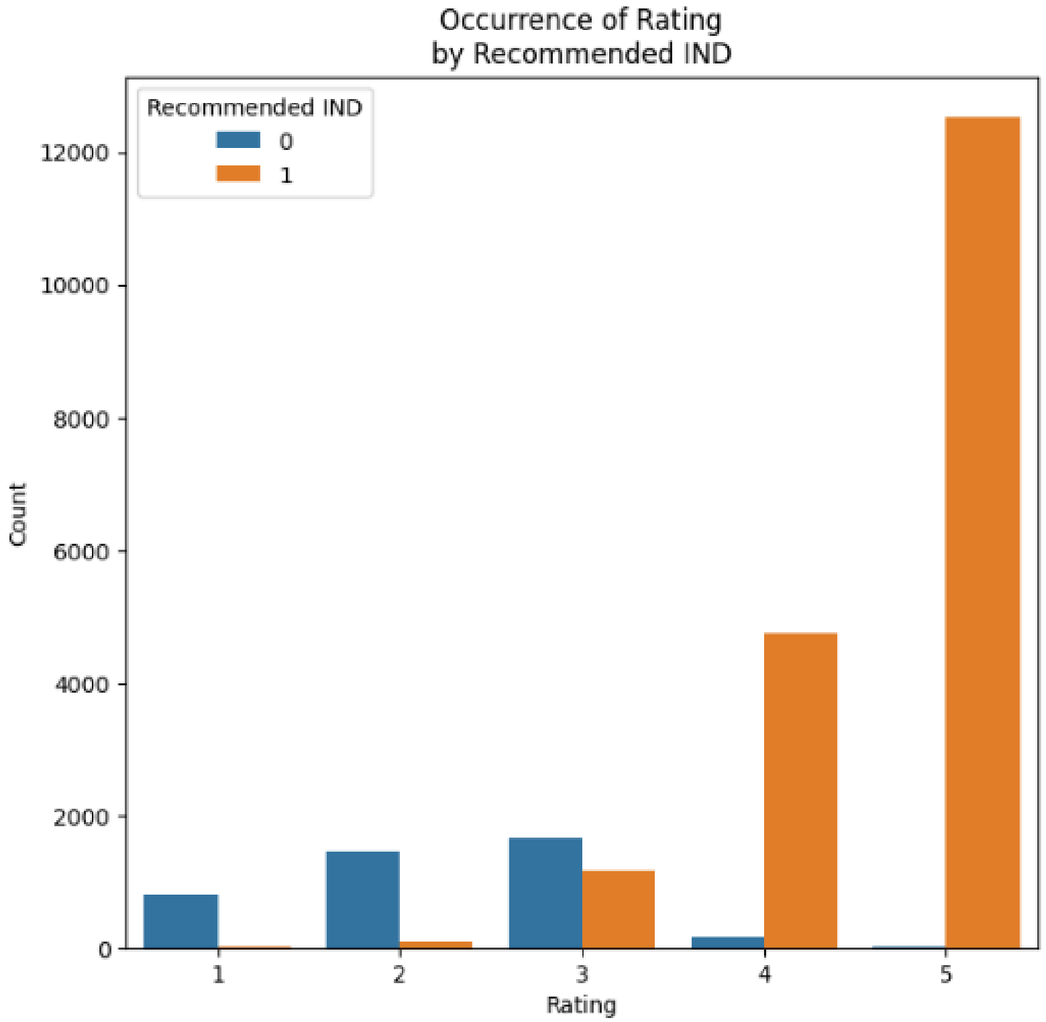
Without Division Name, Departmen Name, and Class Name



With Division Name, Departmen Name, and Class Name

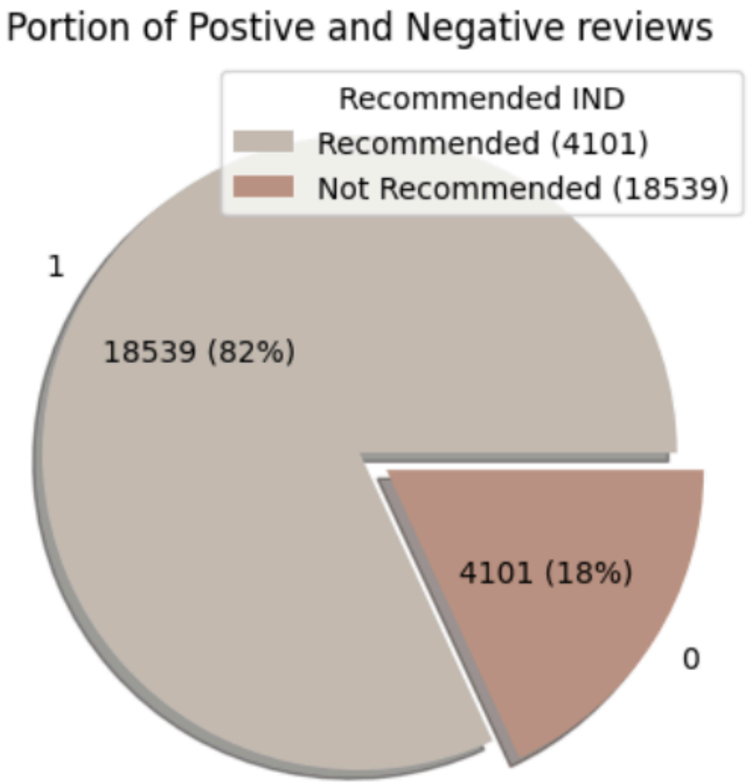
Features Division Name, Department Name, and Class Name show no significant effect on the predicted recommendation and sentiment analysis results

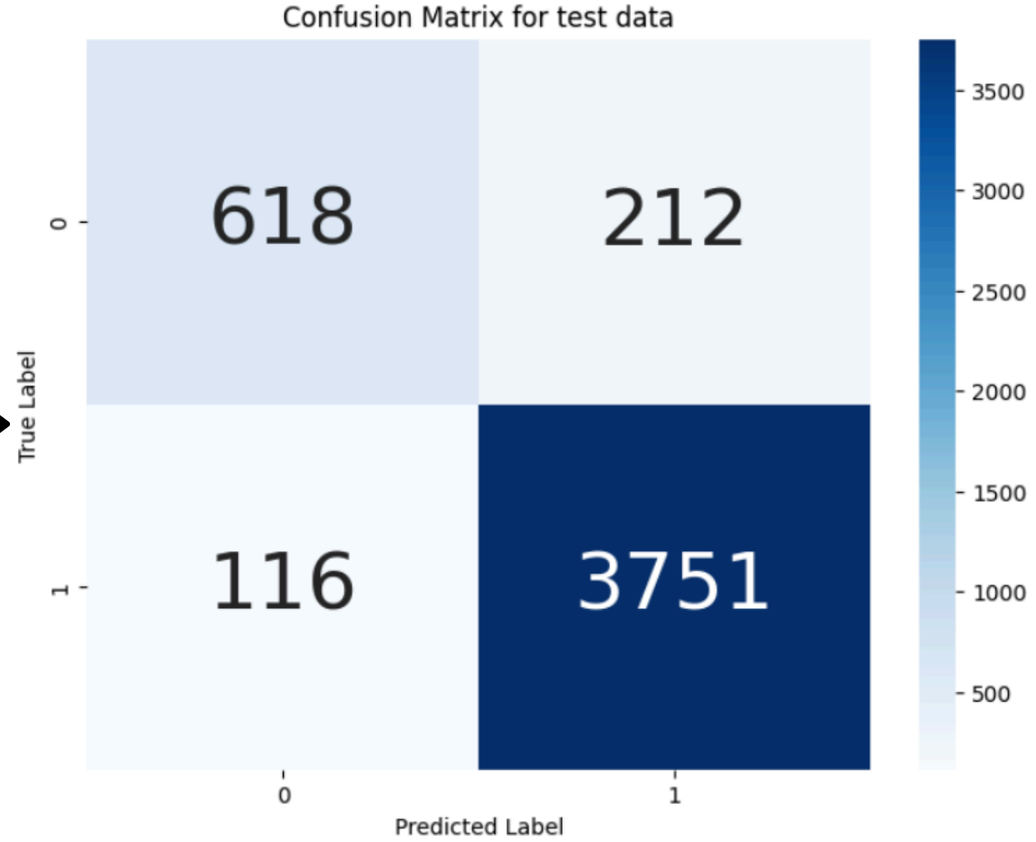
Corelation Rating with Label



Challenge:

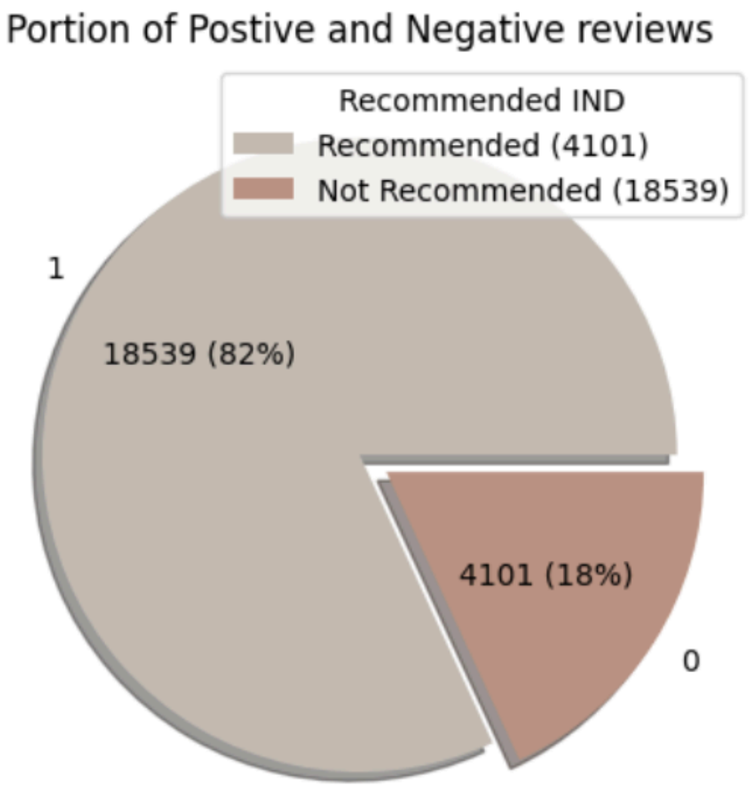
- Customers who give high ratings but do not recommend the product.
- Customers who give low ratings but still recommend the product.
- Products with a rating of 3 result in more non-recommended products than recommended ones.





NOT BALANCED !

The prediction of the model still prefer 1 labels because is contains rich information for 1 labels so imbalanced data is not good to fed into our model



NOT BALANCED !

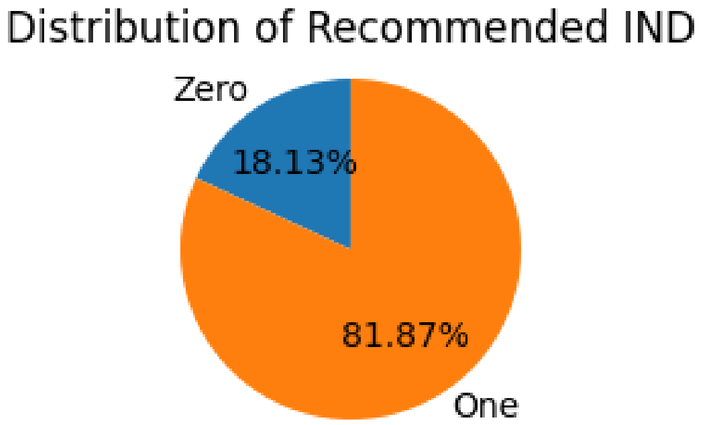
→ Adding Data

SMOTE

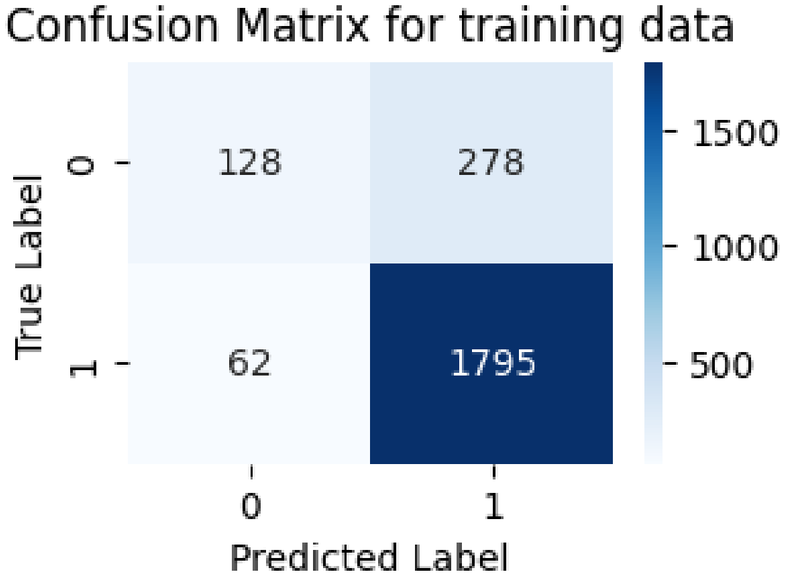
→ Clusterring Data

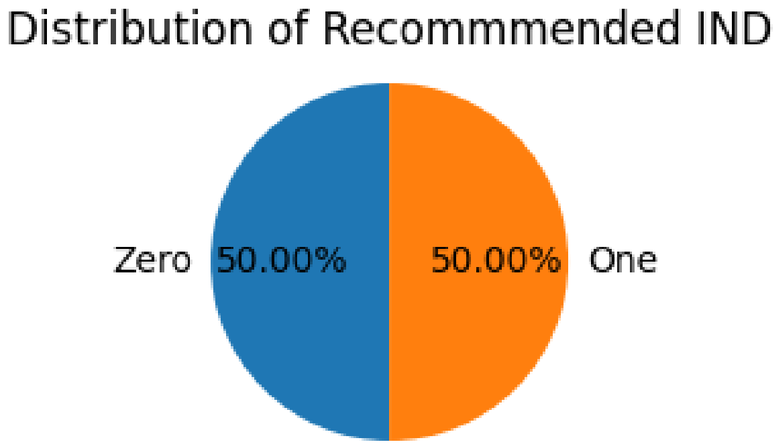
K-Means

Comparation Data and Prediction

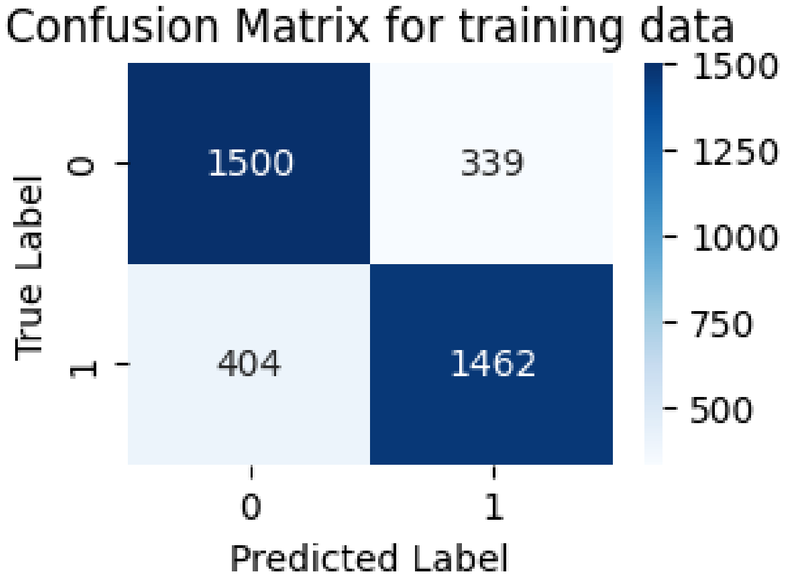


Without SMOTE



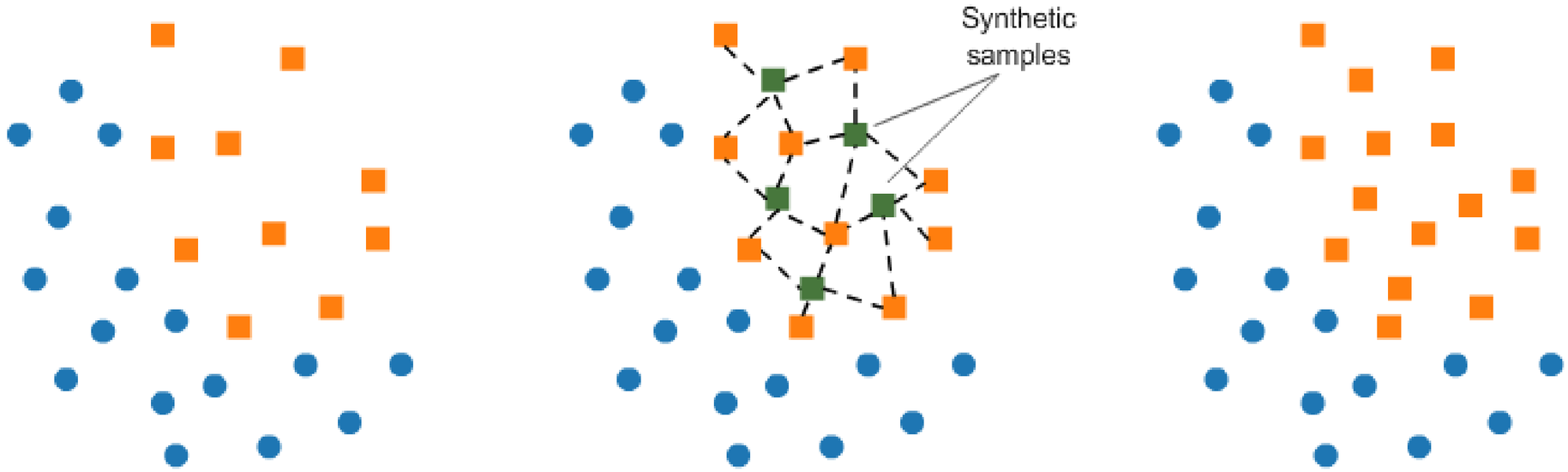


With SMOTE

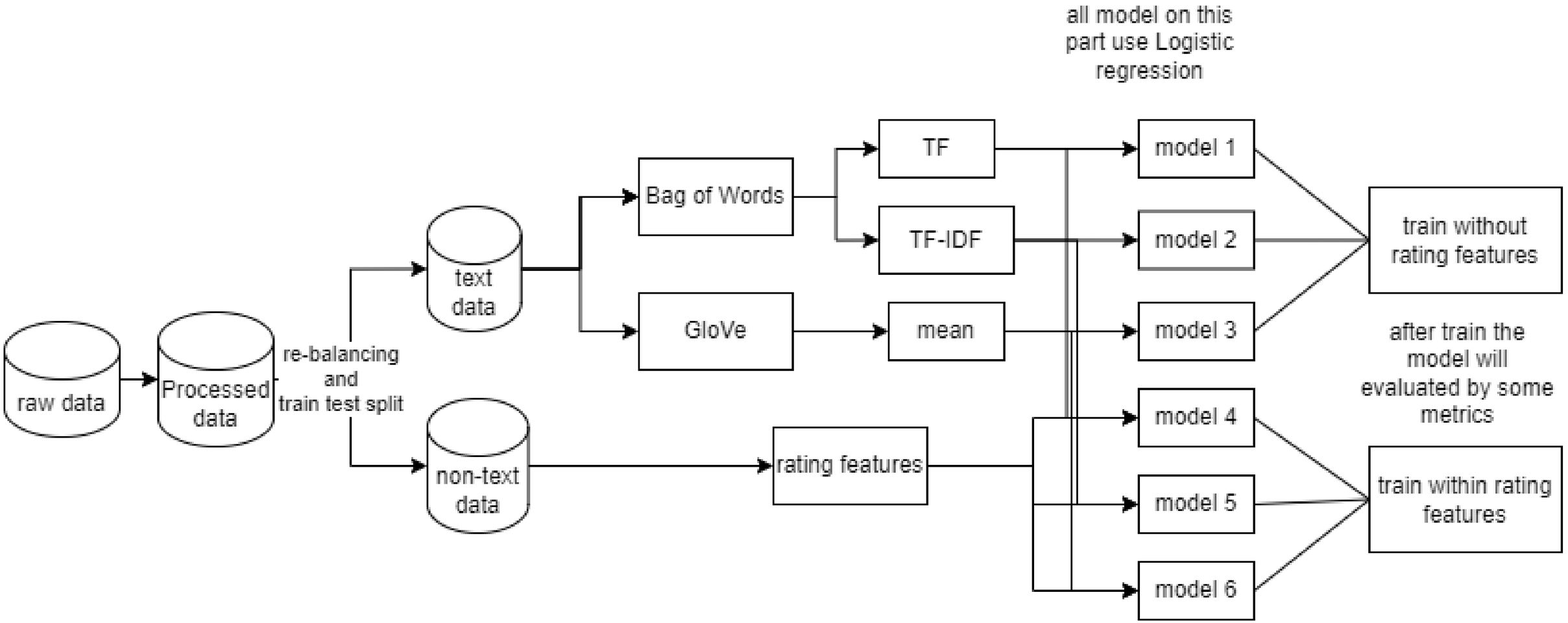


Re-Balancing with SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) is a method used to address class imbalance in machine learning datasets by generating synthetic samples for the minority class. It works by creating new instances of minority class samples by interpolating between existing minority class instances, thus helping to balance the class distribution and improve the performance of classifiers, especially in scenarios where the minority class is underrepresented.



Machine Learning workflow



TF and TFIDF

$$TF(t,d) = \frac{number\ of\ times\ t\ appears\ in\ d}{total\ number\ of\ terms\ in\ d}$$

$$IDF(t) = log\frac{N}{1+df}$$

$$TF-IDF(t,d) = TF(t,d) * IDF(t)$$

Term Frequency (TF): Measures how frequently a term appears in a document.

TF-IDF: Weighs a term's frequency in a document against its frequency across all documents, highlighting important terms unique to each document.

Example

for the constraint on the vectorizer, we added the minimal occurence of a word is 10 to prevent curse of dimensionality

'absolutely wonderful silky and sexy and comfortable wonderful sexy comfortable'

	absolutely	wonderful	silky	and	comfortable	wonderful	sexy	comfortable	huge
0	1	2	1	2	2	2	2	2	0

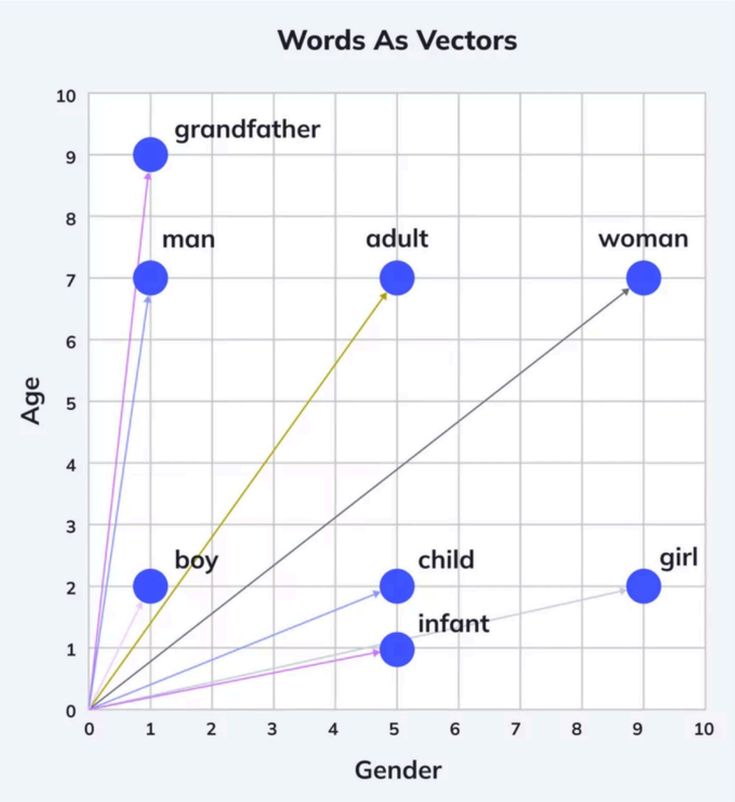
← only TF

	absolutely	wonderful	silky	and	comfortable	wonderful	sexy	comfortable	huge
0	0.238986	0.574449	0.340758	0.126147	0.325865	0.574449	0.612109	0.325865	0.0

← with TF-IDF

GloVe

GloVe (Global Vectors for Word Representation) is an unsupervised learning algorithm for generating word embeddings by aggregating global word-word co-occurrence statistics from a corpus.



In this section, we will use the 6B pretrained GloVe embeddings with dimensions of 50, 100, and 200 for comparison.

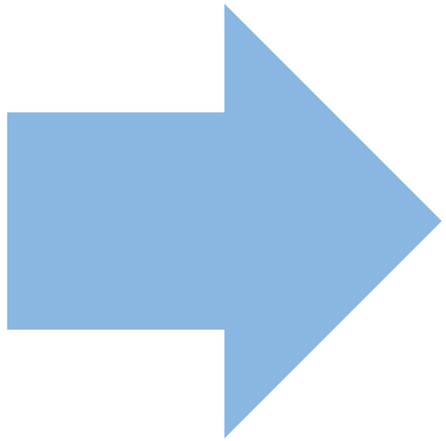
Our challenge is to handle the multiple vectors generated for each piece of text, as logistic regression requires a single vector per data point.

To address this, we will apply mean pooling to combine the word vectors into a single vector for each text.

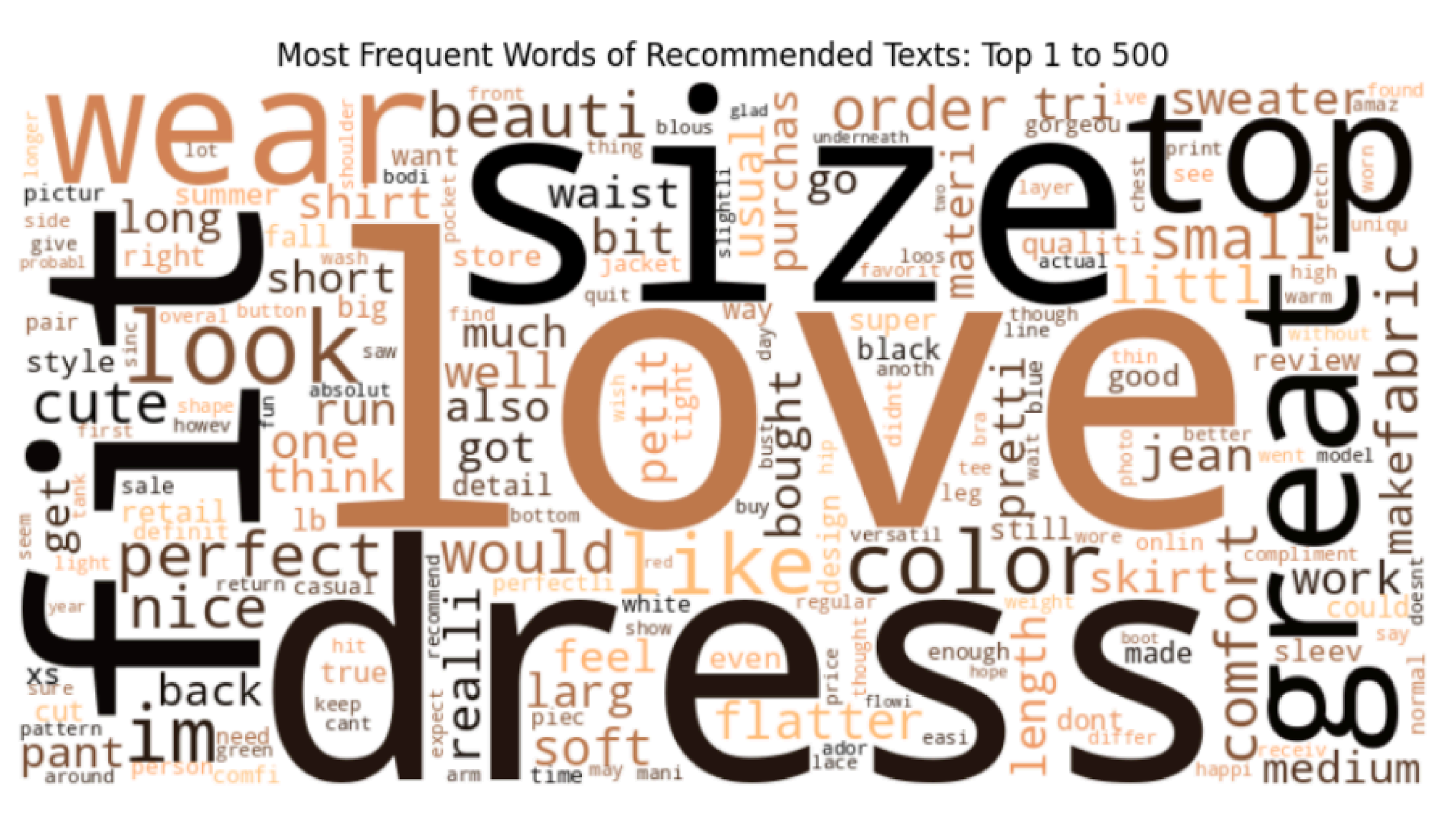
Mean Pooling Vectors

This approach is widely used due to its simplicity and effectiveness in capturing the overall meaning of a text by averaging the word vectors.

	this	product	is	amazing
0	0.53074	0.15882	0.61850	-0.013117
1	0.40117	-0.27394	0.64254	0.421970
2	-0.40785	0.25375	-0.46552	-0.263170
3	0.15444	0.76122	0.37570	0.342180
4	0.47782	0.30715	0.74838	0.999220
...	...	...	...	...
45	0.28164	0.21341	0.19403	0.257900
46	0.12819	0.20546	-0.12466	0.500160
47	0.28762	0.76339	-0.27557	-0.122940
48	0.14440	0.38060	0.30899	-0.209300
49	0.23611	0.70857	0.48497	0.669110



	mean
0	0.323736
1	0.297935
2	-0.220697
3	0.408385
4	0.633142
...	...
45	0.236745
46	0.177287
47	0.163125


Most Frequent Words of Recommended Texts: Top 1 to 500
wear
beauti
size
top
order
tri
sweater
found
amaz
gorgeou
underneath
blous
glad
thing
lot
want
bodi
pictur
summer
shirt
waist
bit
go
materi
print
see
layer
chest
stretch
uniqu
worn
long
fall
right
store
jacket
slightli
usual
purchas
qualiti
small
actual
loos
favorit
high
warm
short
big
overal
button
wash
quit
find
way
much
super
though
line
without
pair
style
look
love
littl
review
good
thin
absolut
saw
sinc
shape
howev
first
cute
run
well
also
got
petit
tight
wish
day
bust
black
anoth
didnt
blue
pretti
bra
tee
wait
photo
better
went
model
jean
fit
dress
flatter
comfort
bought
make
fabric
one
think
detail
sale
retail
definit
lb
bottom
buy
design
hip
leg
versatil
still
wore
onlin
compliment
work
perfect
would
like
color
skirt
could
nice
return
casual
light
red
recommend
reall
white
show
regular
weight
doesnt
say
sleev
xs
back
hit
true
keep
cant
cut
pattern
pant
im
need
green
person
comfi
around
feel
even
larg
piec
soft
time
may
mani
arm
expect
price
thought
flowi
easi
ador
lace
enough
hope
made
boot
dont
differ
receiv
happi
medium
normal
lengh

Most Frequent Words of Not recommended Texts: Top 1 to 500

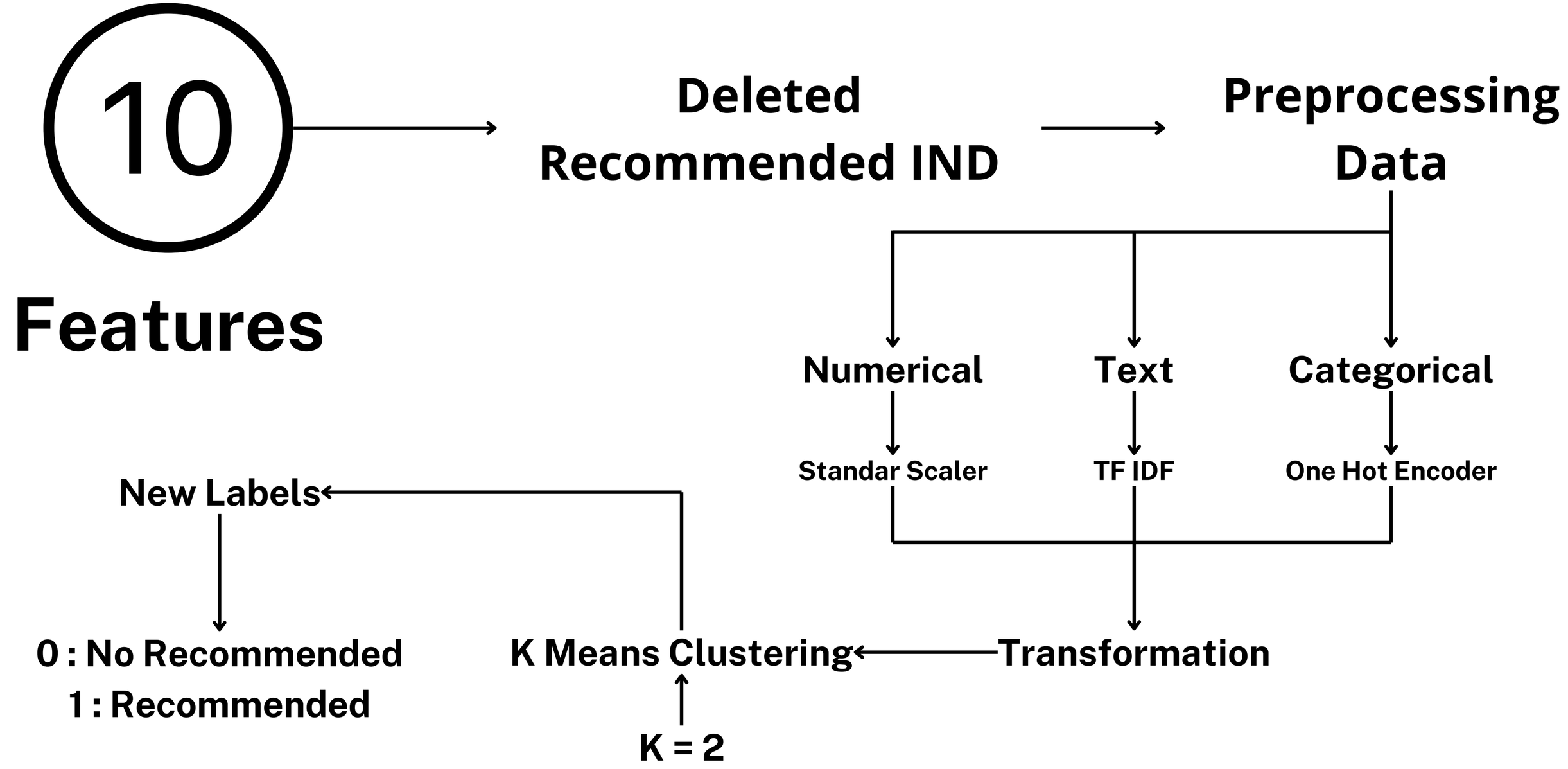
Result Logistic Regression Without Rating

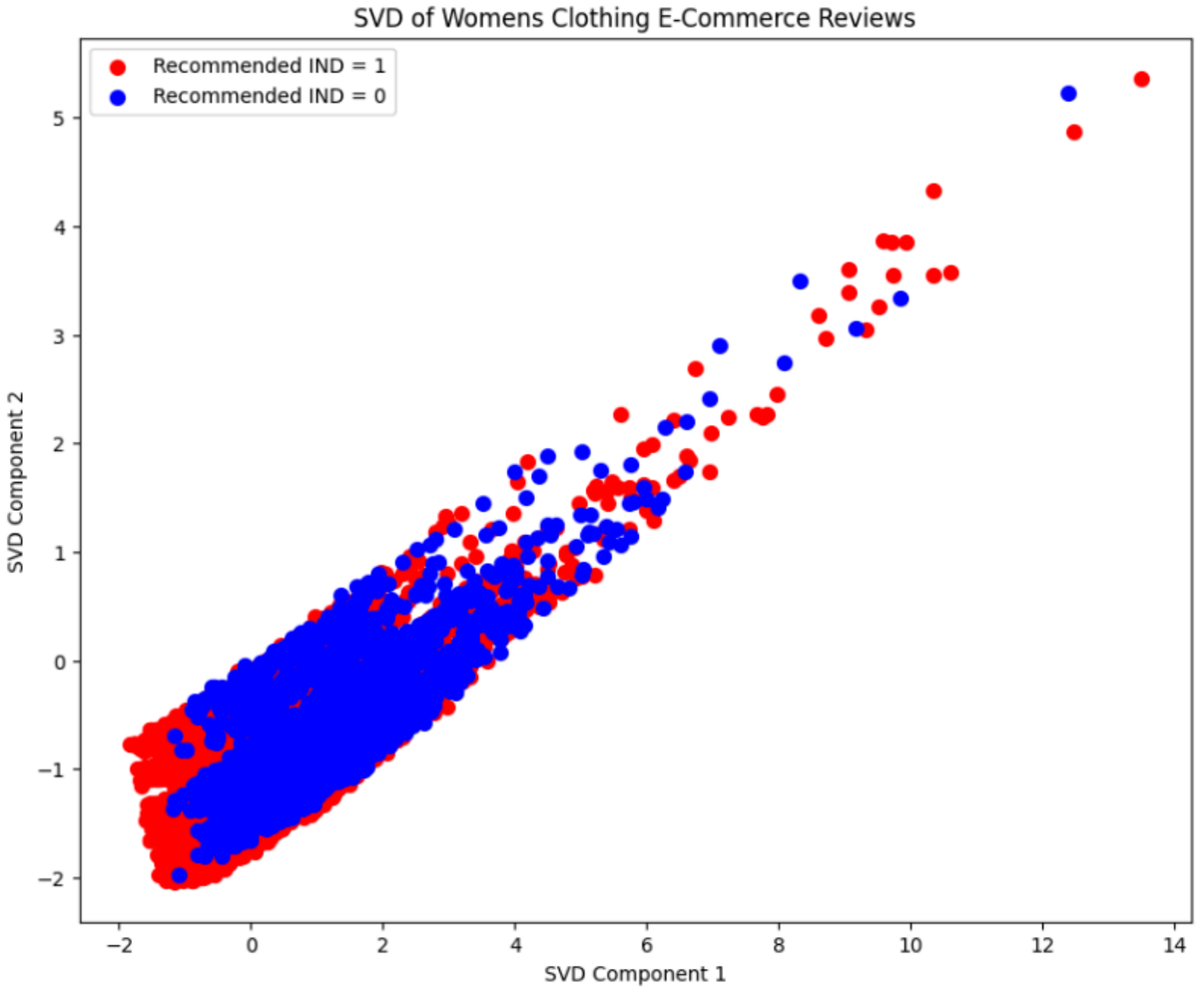
Text Representation	Accuracy	Precision	Recall	F1-Score	Loss
TF	0.928	0.927	0.926	0.926	0.2249
TF-IDF	0.925	0.940	0.903	0.921	0.2338
Glove 50d	0.777	0.785	0.745	0.764	0.4631
Glove 100d	0.821	0.830	0.794	0.812	0.4062
Glove 200d	0.848	0.857	0.825	0.840	0.3668

Result Logistic Regression With Rating

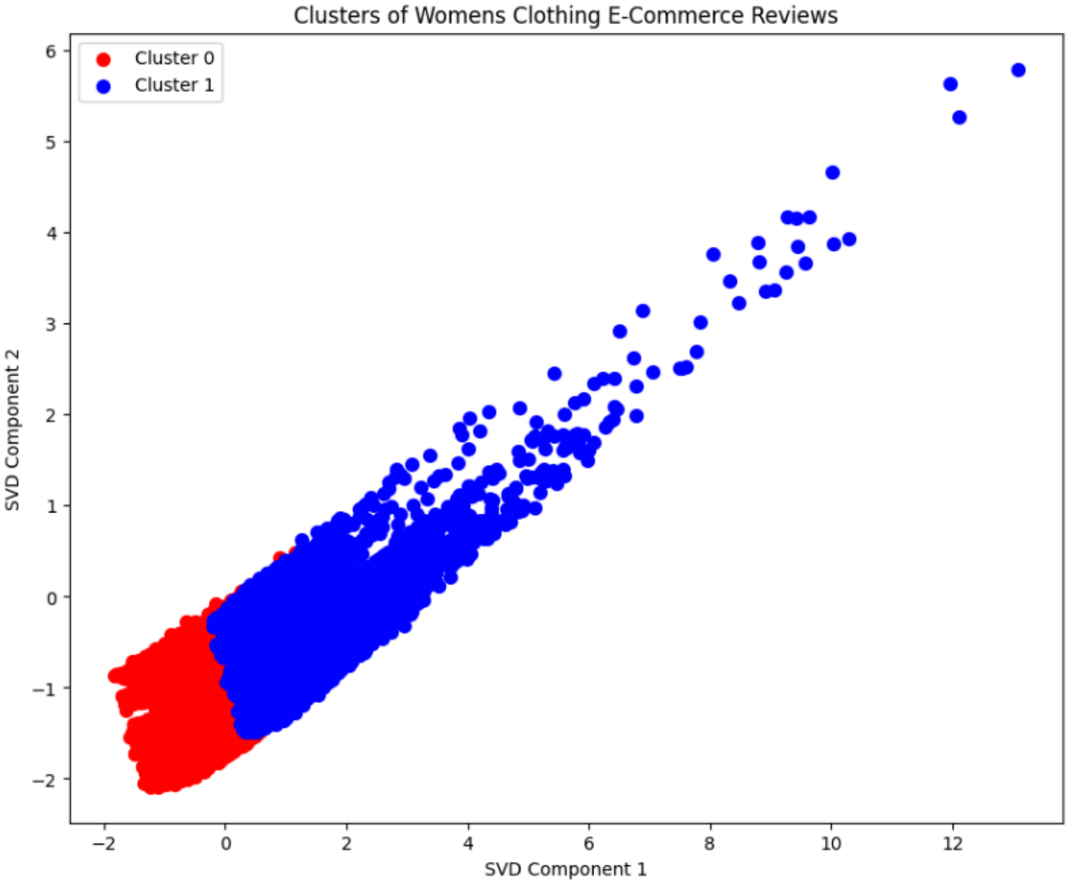
Text Representation	Accuracy	Precision	Recall	F1-Score	Loss
TF	0.958	0.952	0.962	0.957	0.1272
TF-IDF	0.958	0.968	0.944	0.956	0.1297
Glove 50d	0.961	0.979	0.939	0.959	0.1398
Glove 100d	0.958	0.975	0.938	0.956	0.1397
Glove 200d	0.958	0.972	0.940	0.956	0.1375

Clustering Works



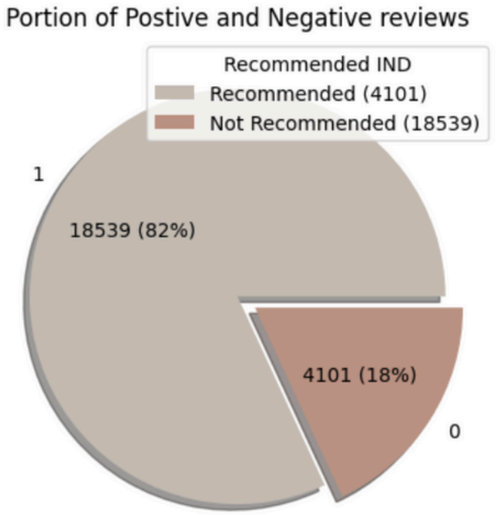


Data Distribution Before Clustering

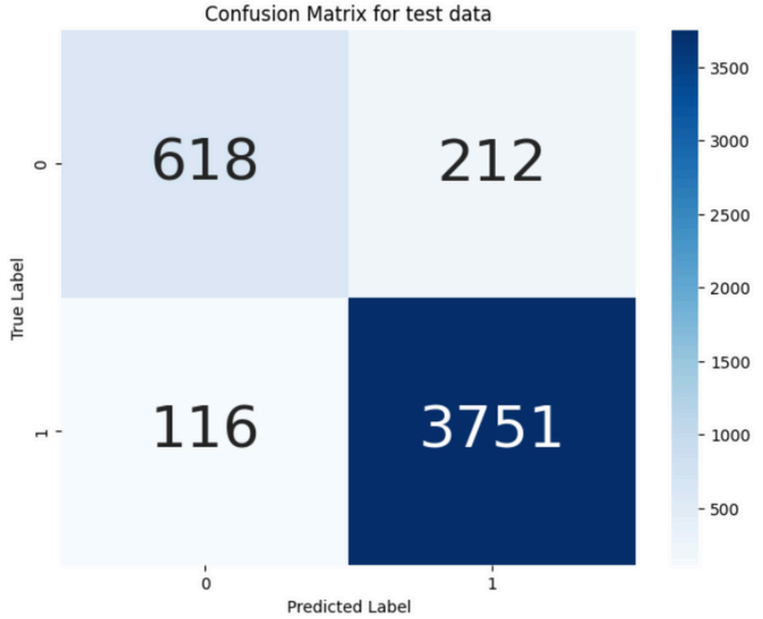


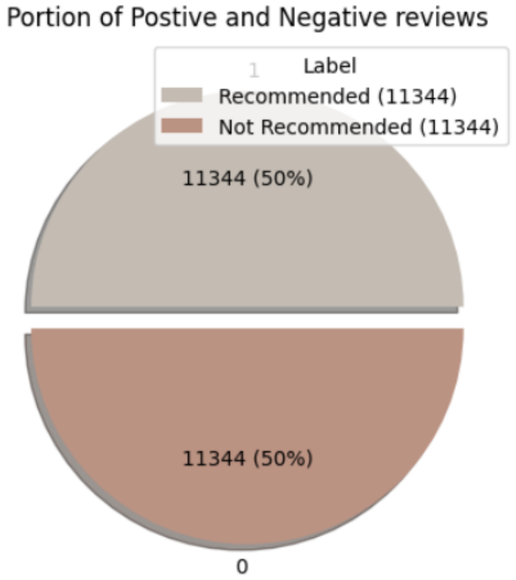
Data Distribution After Clustering

Comparation Data and Prediction

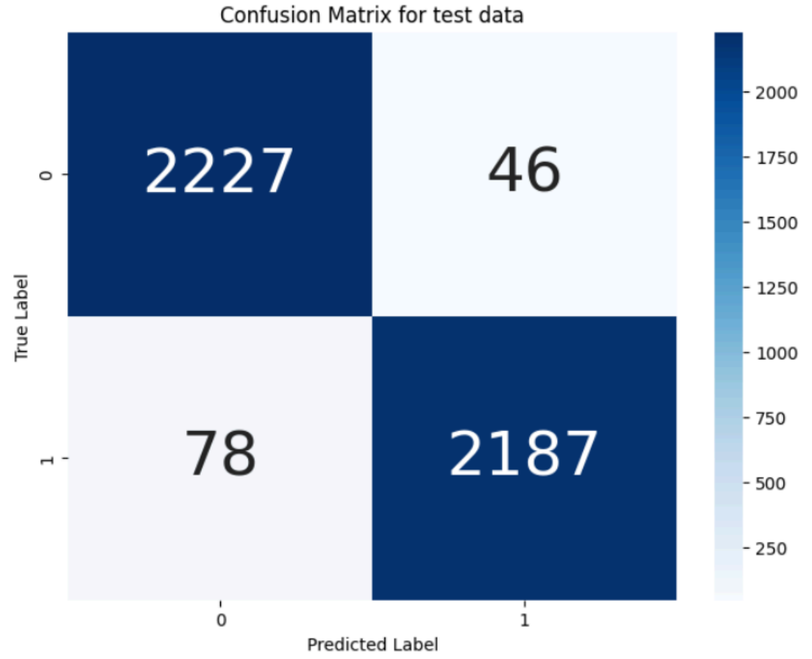


Without K-Means Clustering



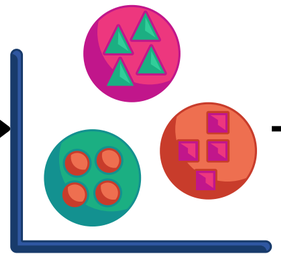


With K-Means Clustering



CNN-LSTM Workflow

Raw Data



K-Means

Numerical Features

Categorical Features

Text Features

GloVe Embeddings

Concatenated

CNN Layer 1

Max Polling

CNN Layer 2

Max Polling

LSTM

Fully Conected 1

Fully Conected 2

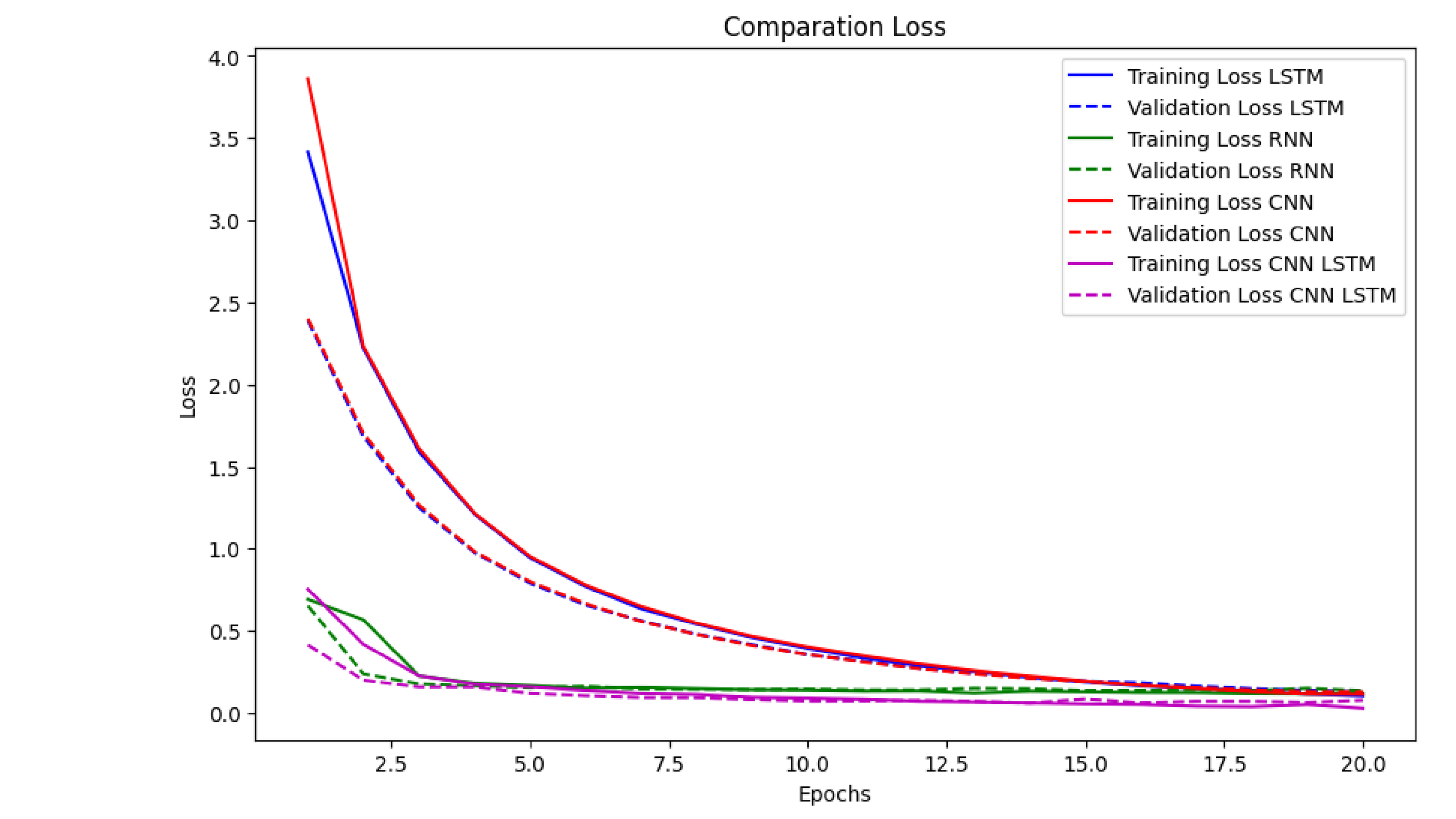
Label Prediction

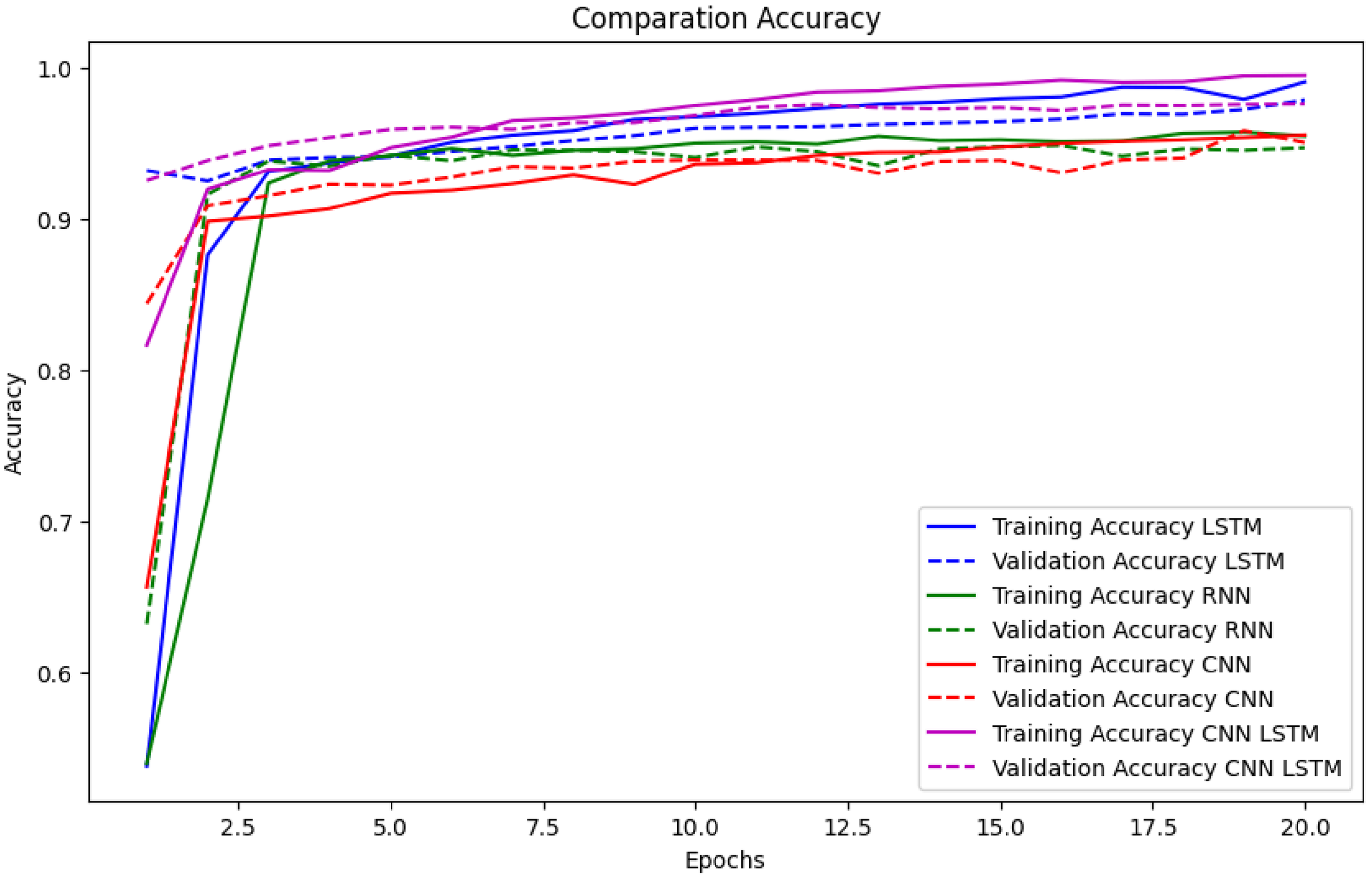
No Recomenddation

Recomenddation

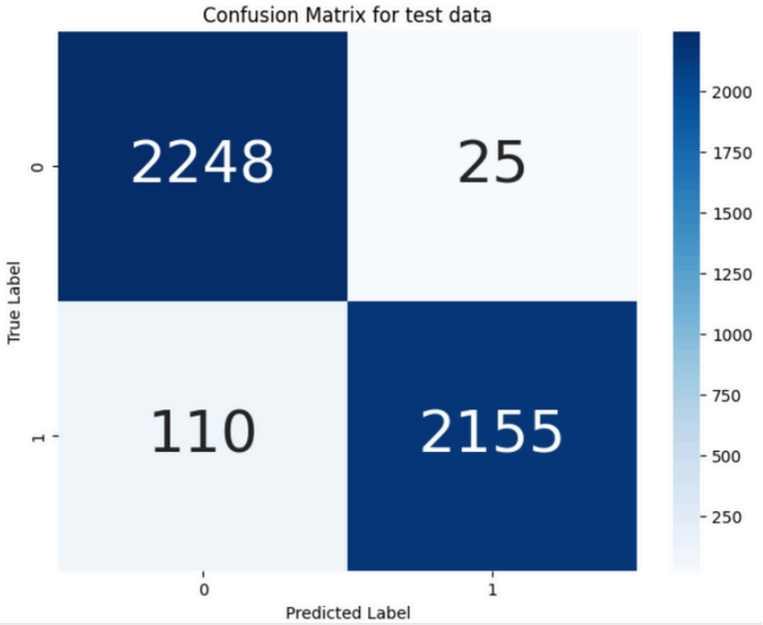
Deep Learning Result with K Means

Model	Accuracy	Precision	Recall	F1-Score	Loss
Golve CNN	0.9703	0.9703	0.9703	0.9703	0.1658
Golve RNN	0.9726	0.9726	0.9726	0.9726	0.1397
Golve LSTM	0.9735	0.9735	0.9735	0.9735	0.1298
Glove CNN LSTM	**0.9764**	**0.9764**	**0.9764**	**0.9764**	**0.0688**

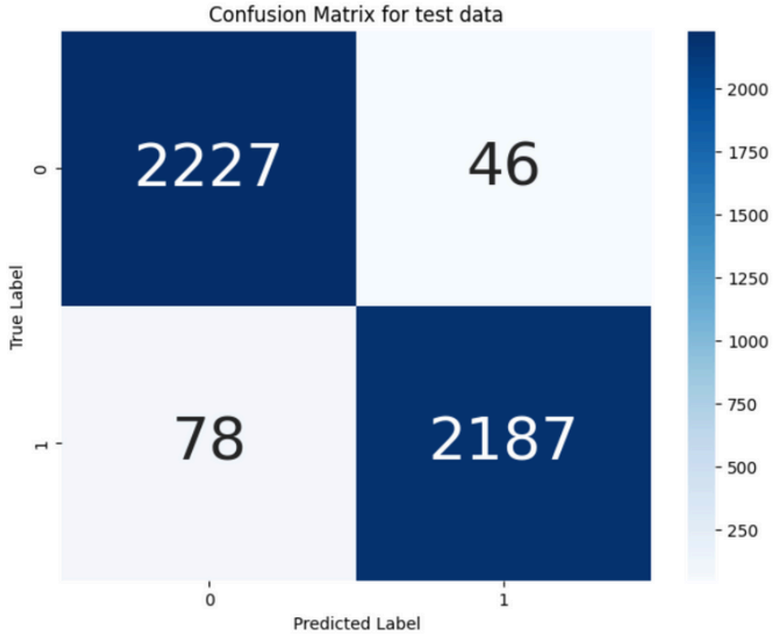

Comparation Loss
Training Loss LSTM
Validation Loss LSTM
Training Loss RNN
Validation Loss RNN
Training Loss CNN
Validation Loss CNN
Training Loss CNN LSTM
Validation Loss CNN LSTM
Loss
Epochs
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0


Comparation Accuracy
Accuracy
Epochs
1.0
0.9
0.8
0.7
0.6
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Training Accuracy LSTM
Validation Accuracy LSTM
Training Accuracy RNN
Validation Accuracy RNN
Training Accuracy CNN
Validation Accuracy CNN
Training Accuracy CNN LSTM
Validation Accuracy CNN LSTM

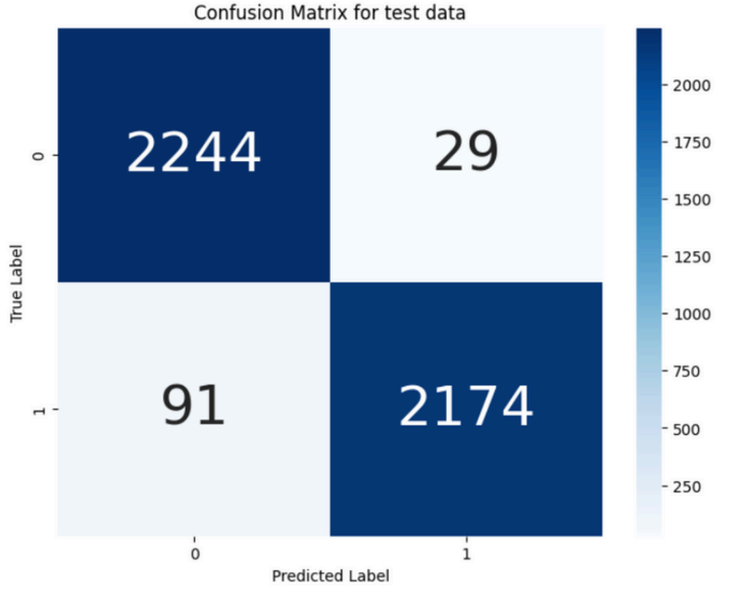
CNN



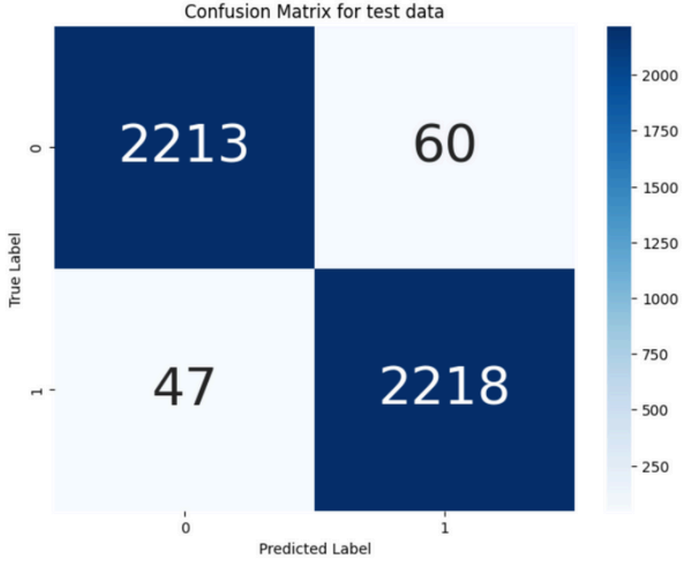
RNN



LSTM



CNN LSTM



Compare with SOTA

Model	Accuracy	Precision	Recall	F1-Score	Loss
TF Logistic Regression	0.958	0.952	0.962	0.957	0.1272
TF-IDF Logistic Regression	0.958	0.968	0.944	0.956	0.1297
Glove 50d Logistic Regression	0.961	0.979	0.939	0.959	0.1398
Golve LSTM	0.9735	0.9735	0.9735	0.9735	0.1298
Glove CNN LSTM	**0.9764**	**0.9764**	**0.9764**	**0.9764**	**0.0688**

Conclusion

- Clustering is needed to overcome data imbalance so that better prediction results are obtained than before clustering
- Of the 10 features available, the most influential features in this prediction are the review text, title text, and rating features which are useful for weighting review sentiment and title text
- Machine Learning models using Logistics succeeded in predicting labels well, but Deep Learning models were still superior for predicting labels in sentiment analysis for women's clothing recommendations