SENTIMENT ANALYSIS

ON WOMEN'S E-COMMERCE CLOTHING REVIEWS

ABOUT SENTIMENT ANALYSIS

Sentiment analysis is detecting whether a text has a positive or negative connotation.

This can help companies understand what is working and what is not based on the customer feedbacks.

Positive reviews show what people like.

Negative reviews identify the issues and can help change directions and improve the product or service.

DATA

23,486 DATA POINTS FROM AN ONLINE WOMEN CLOTHING RETAILER DATABASE

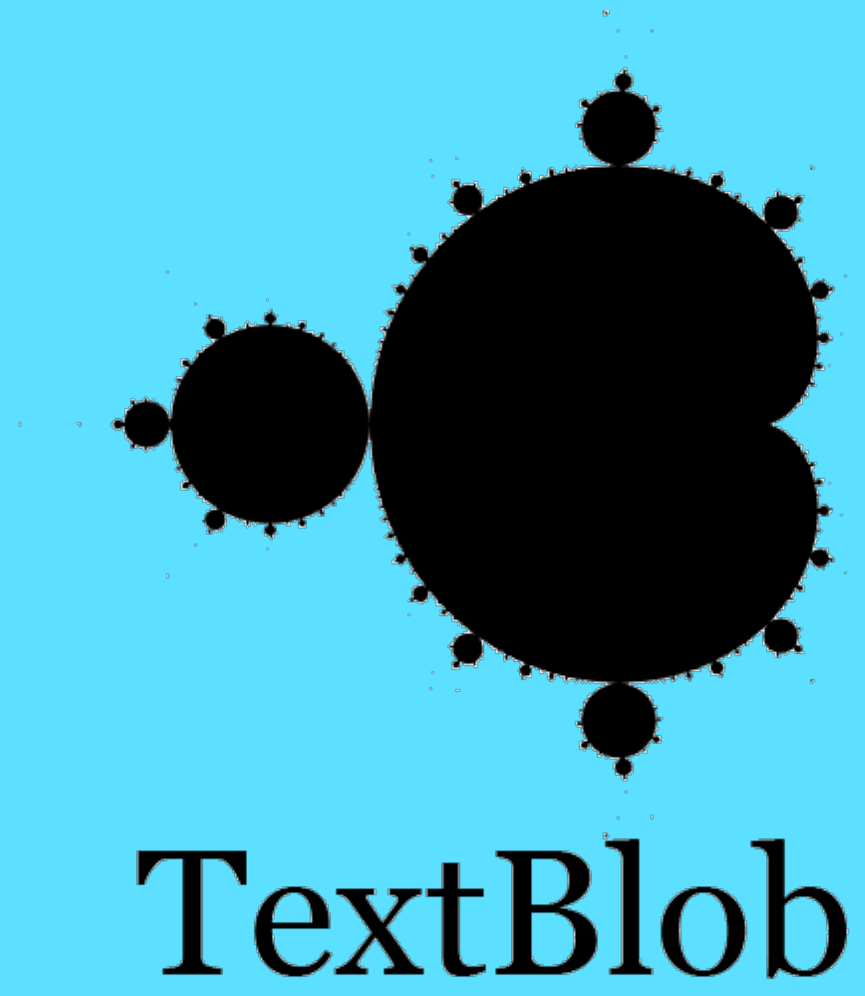
8 COLUMNS : CLOTHING ID, AGE , TITLE, REVIEW TEXT, RATING, RECOMMENDED, DEPARTMENT NAME

1206 CLOTHING ITEMS

GOAL OF THE PROJECT

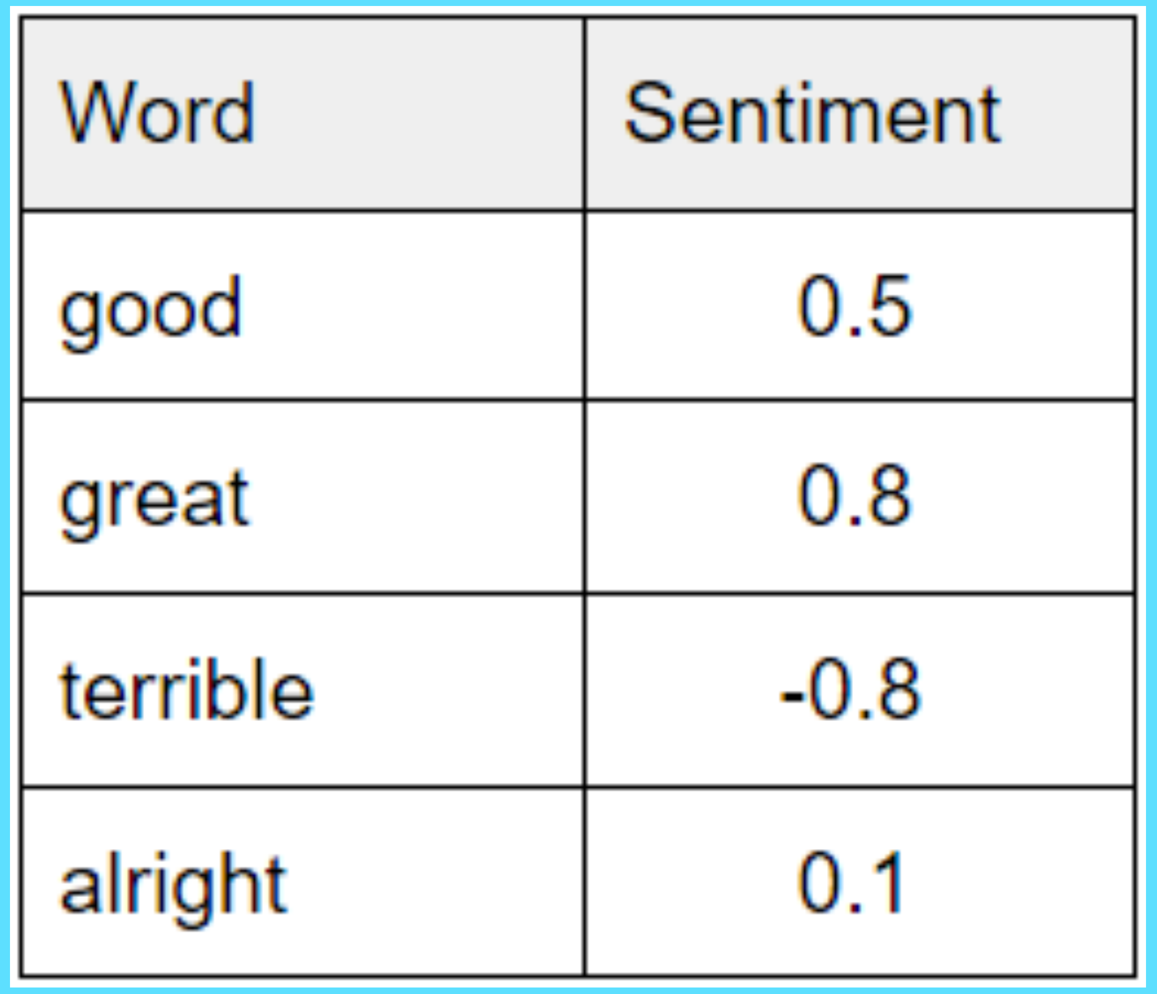
BUILD A SENTIMENT ANALYSIS MODEL BETTER THAN TEXTBLOB

WHAT IS TEXTBLOB ?



‘*TEXTBLOB* IS A PYTHON LIBRARY FOR PROCESSING TEXTUAL DATA AND COMMON NLP TASKS SUCH AS SENTIMENT ANALYSIS, CLASSIFICATION, TRANSLATION, …’

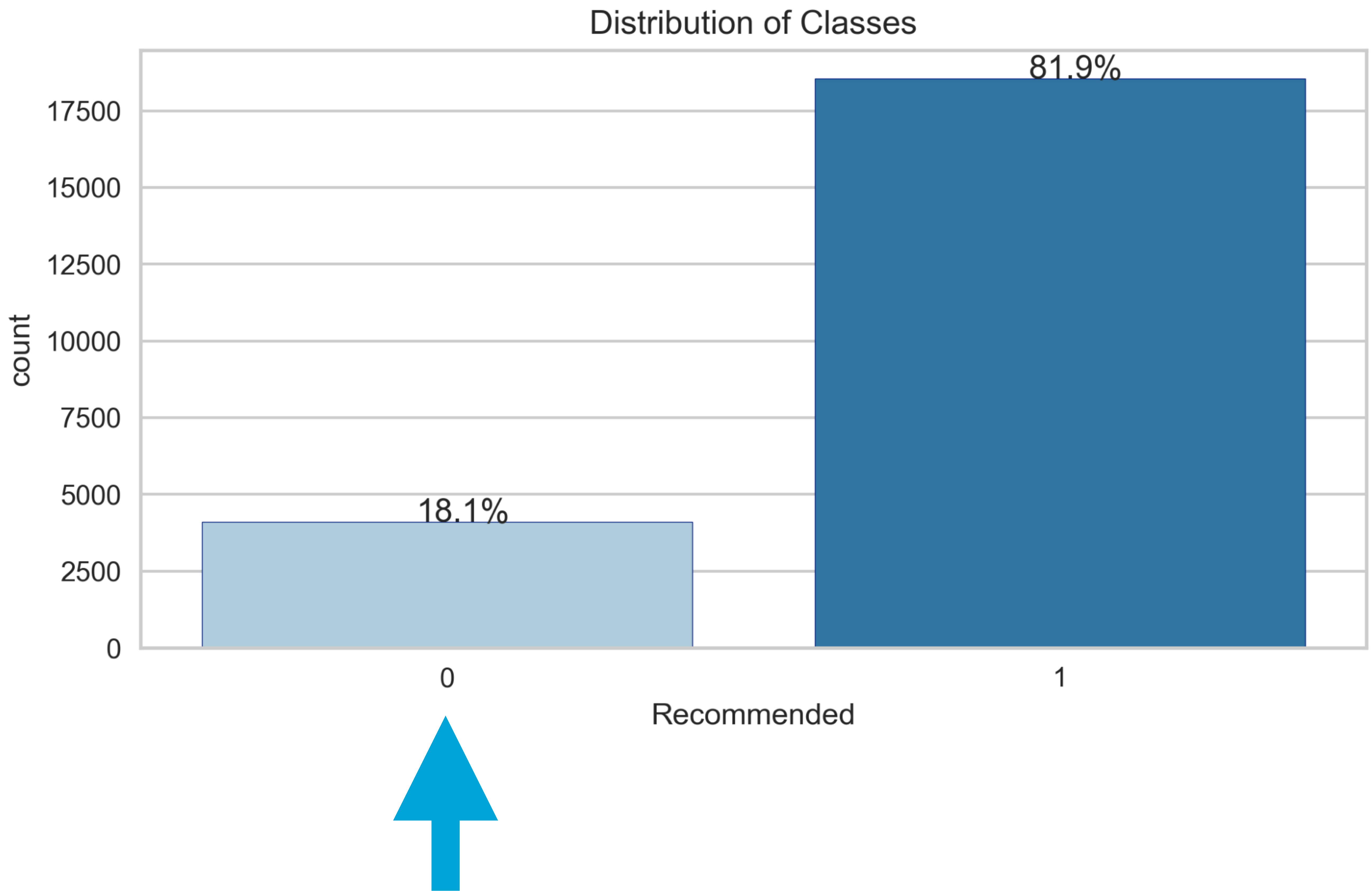
-1 <= POLARITY =<+1



Word	Sentiment
good	0.5
great	0.8
terrible	-0.8
alright	0.1

EDA

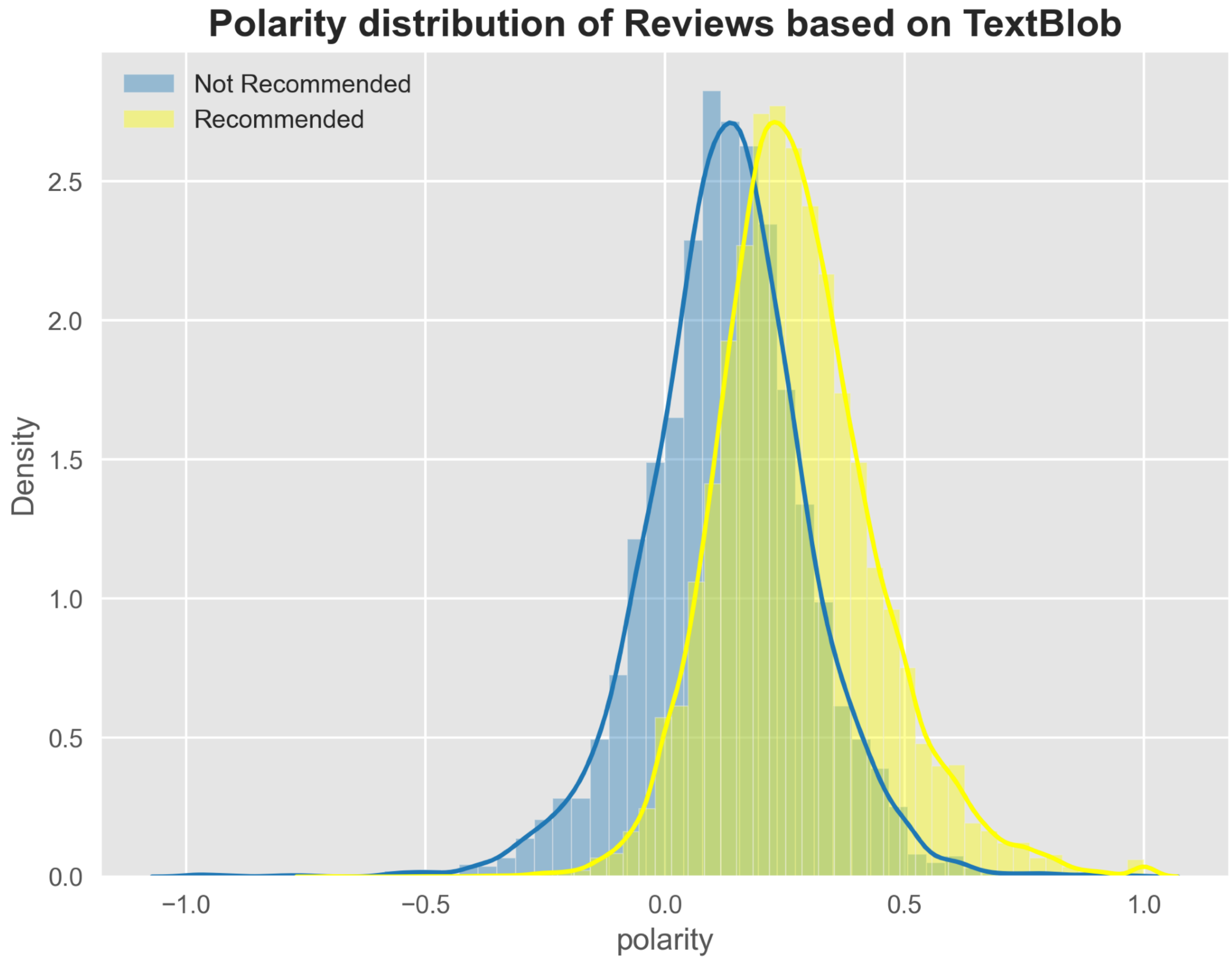
- The majority of reviewers recommend the clothing items.
- The focus of the study is to detect the negative reviews.
- The outcome will provide feedback for the product team to increase future sales.



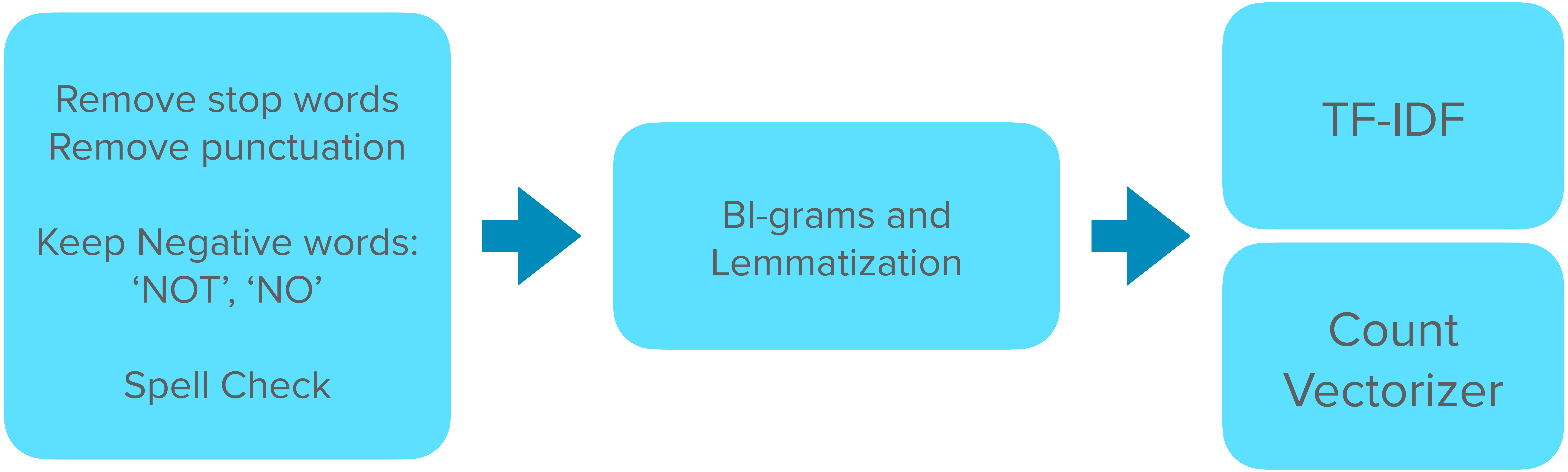
Main Focus

EDA

- Most reviews fall in the positive spectrum .
- Very close distributions between recommended and not recommended items.
- TextBlob doesn't seem to be a helpful tool to classify the sentiment of this dataset.



TEXT PRE-PROCESSING



Remove stop words
Remove punctuation

Keep Negative words:
‘NOT’, ‘NO’

Spell Check

BI-grams and
Lemmatization

TF-IDF

Count
Vectorizer

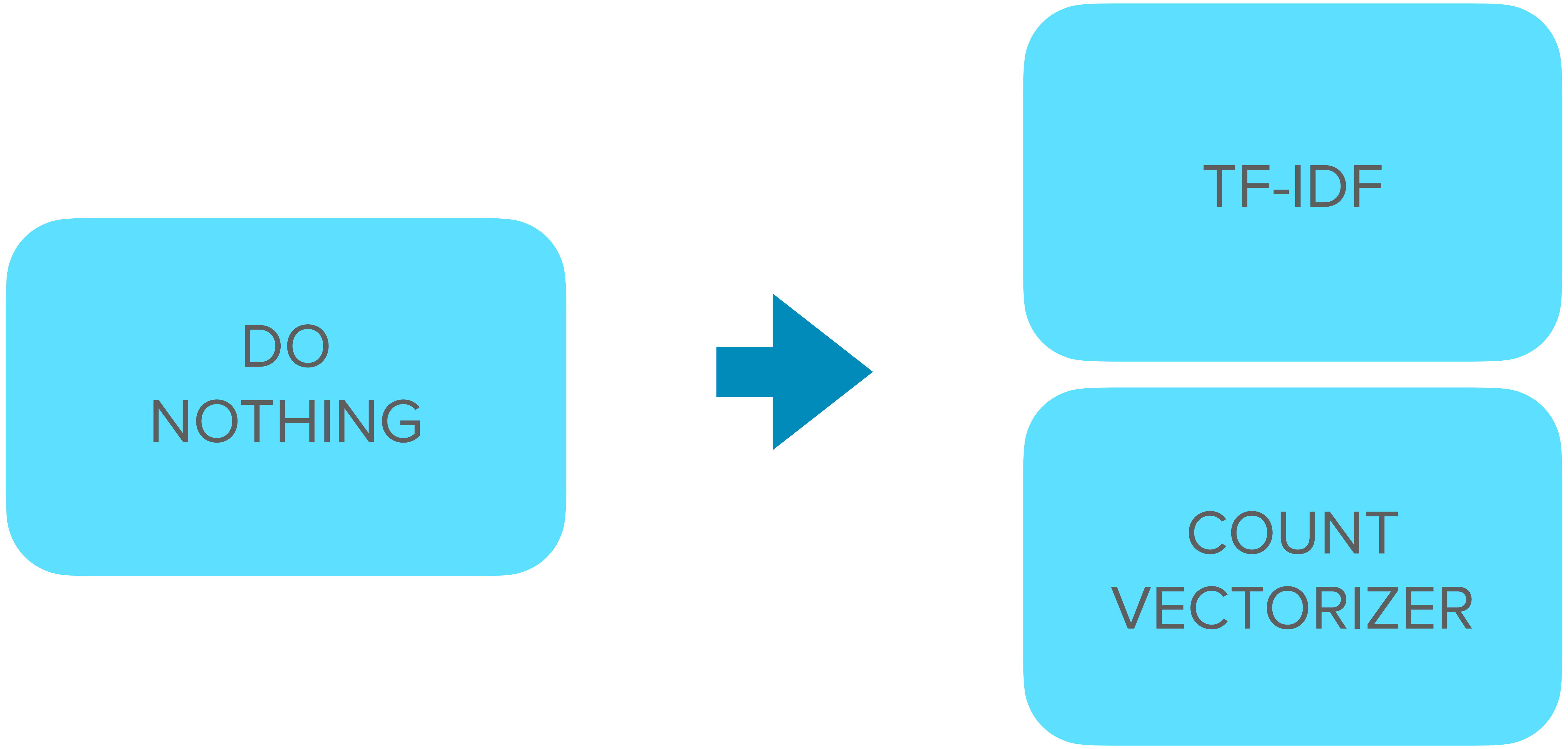
TF-IDF WITH PRE-PROCESSED TEXT

	Accuracy	Recall (Minority Class)	Average Percision Score
Random Forest	0.83	0.09	0.83
Logistic Regression	0.88	0.77	0.93
XGBoost	0.88	0.48	0.89
Naive Bayes	0.75	0.36	0.85

CV WITH PRE-PROCESSED TEXT

	Accuracy	Recall (Minority Class)	Average Percision Score
Random Forest	0.87	0.31	0.86
Logistic Regression	0.89	0.71	0.92
XGBoost	0.89	0.52	0.96
Naive Bayes	0.82	0.19	0.84

NO TEXT PRE-PROCESSING



DO NOTHING

TF-IDF

COUNT VECTORIZER

BEST MODEL

LOGISTIC REGRESSION

BEST VECTORIZER

TF-IDF & COUNT VECTORIZER

BEST PRE-PROCESSING METHOD

DO NOTHING

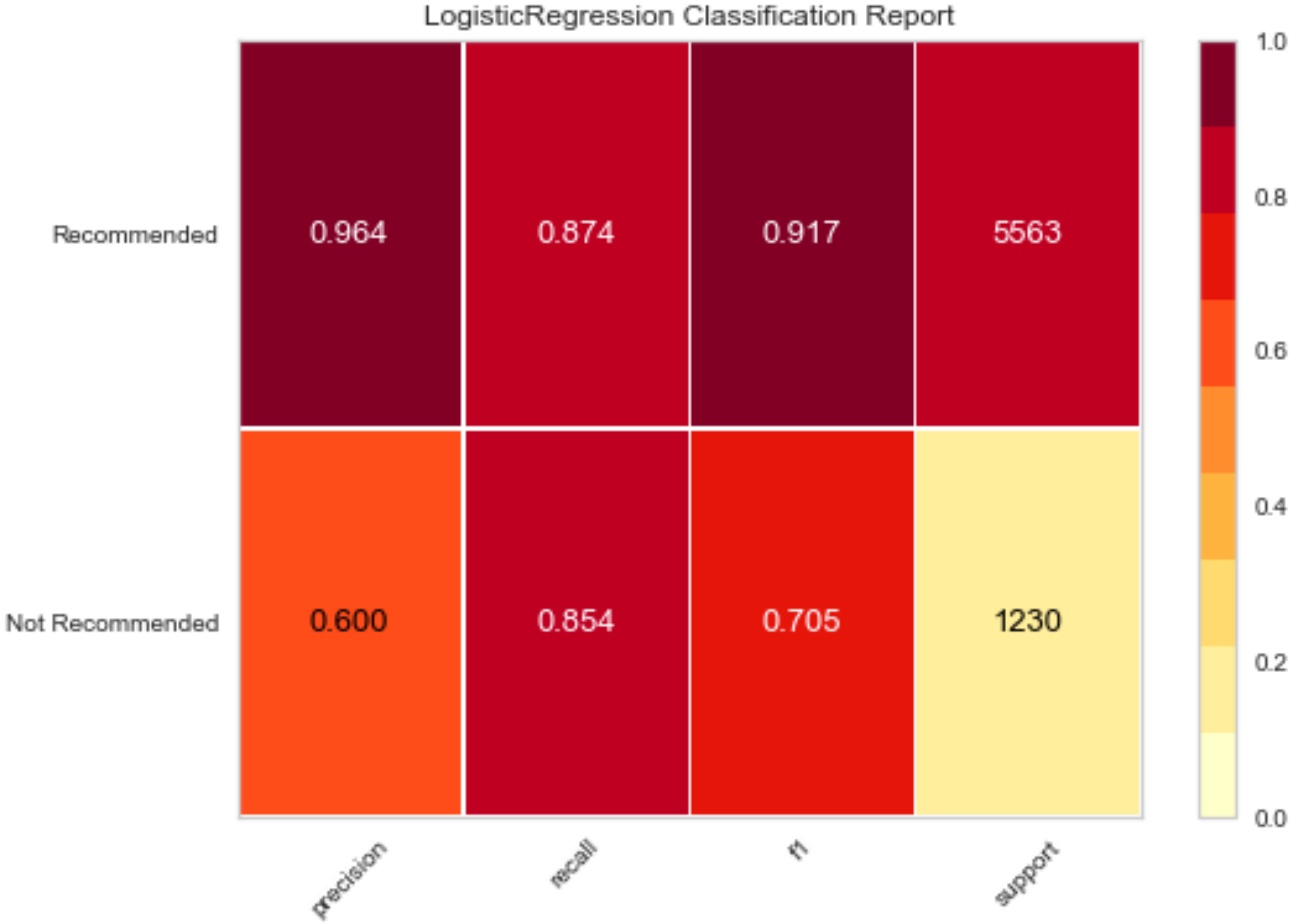
TF-IDF WITH NO TEXT PREP

	Accuracy	Recall (Minority Class)	Average Percision Score
Random Forest	0.84	0.12	0.83
Logistic Regression	0.87	0.82	0.94
XGBoost	0.88	0.49	0.89
Naive Bayes	0.60	0.49	0.83

CV WITH NO TEXT PREP

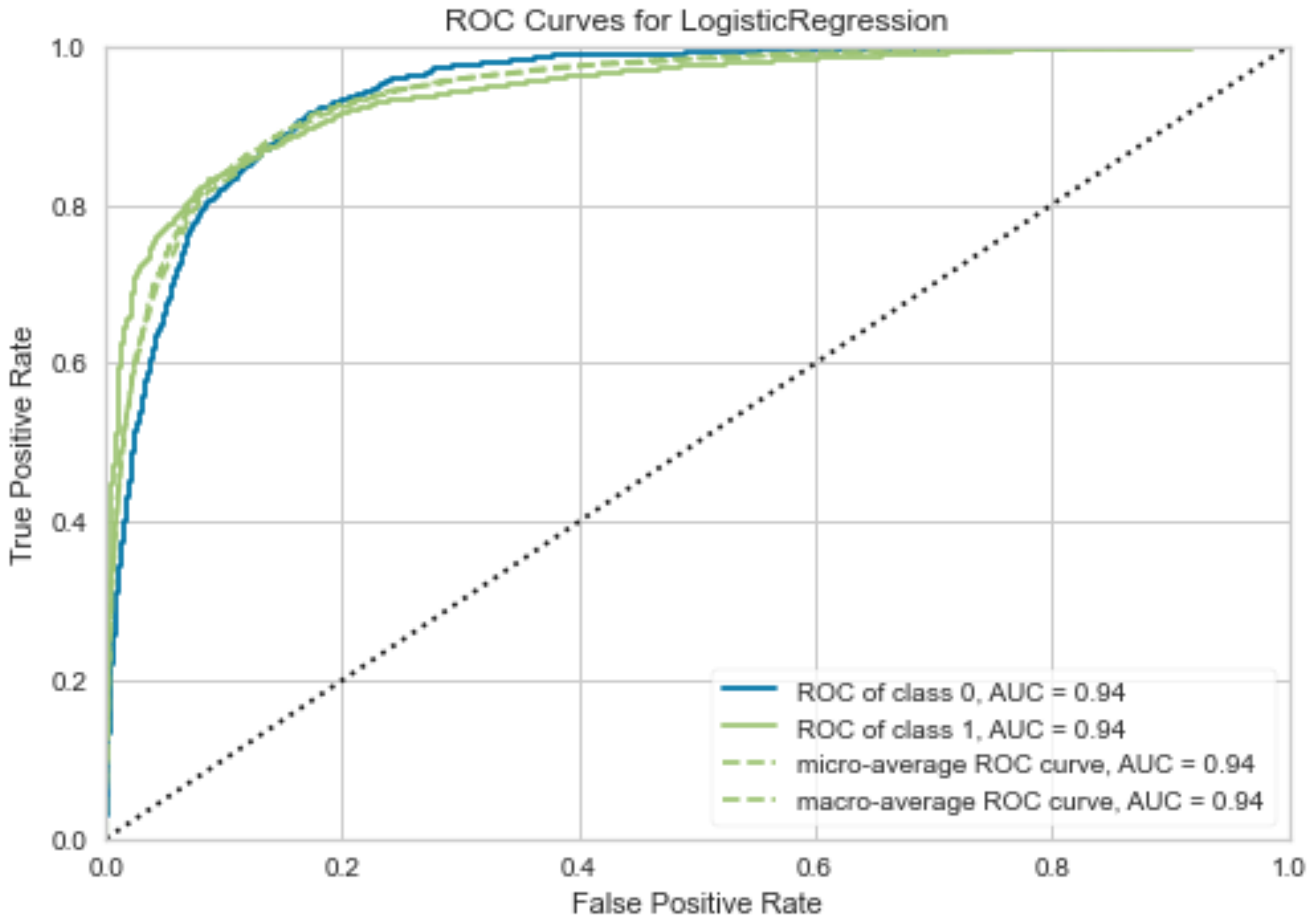
	Accuracy	Recall (Minority Class)	Average Percision Score
Random Forest	0.84	0.11	0.83
Logistic Regression	0.87	0.73	0.92
XGBoost	0.88	0.52	0.89
Naive Bayes	0.60	0.51	0.83

BEST MODEL REPORT



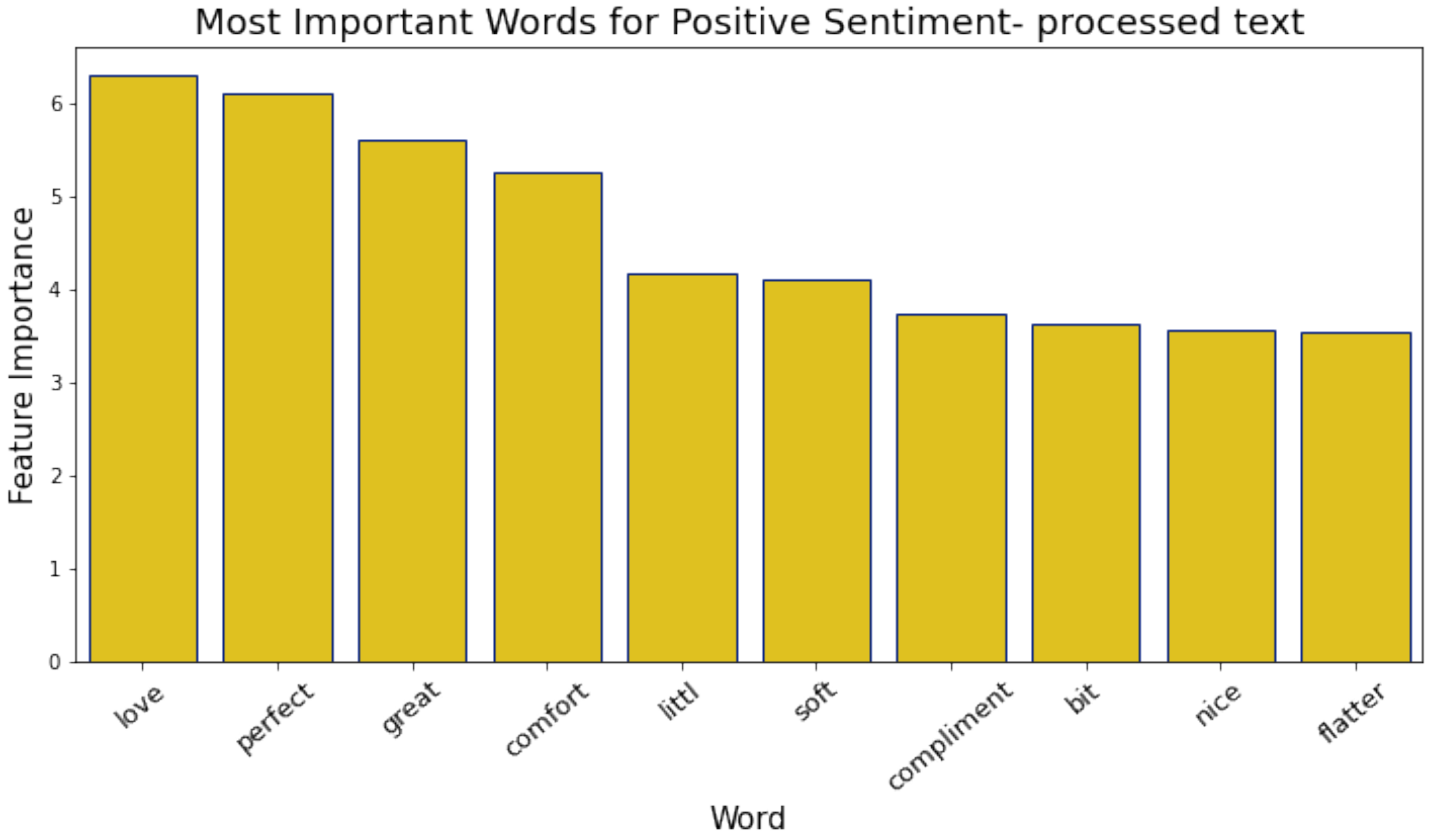
LogisticRegression Classification Report

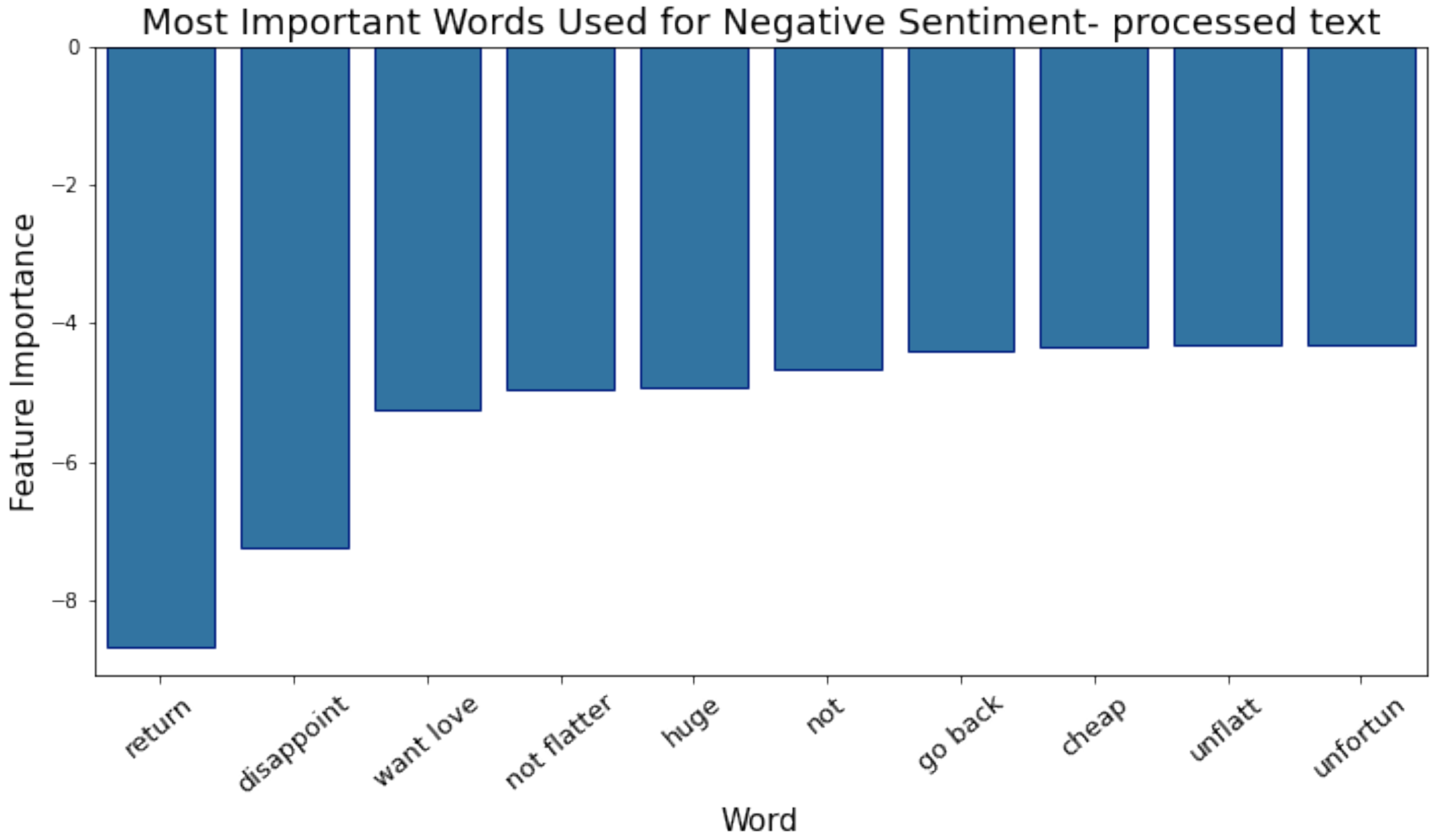
	precision	recall	f1	support
Recommended	0.964	0.874	0.917	5563
Not Recommended	0.600	0.854	0.705	1230



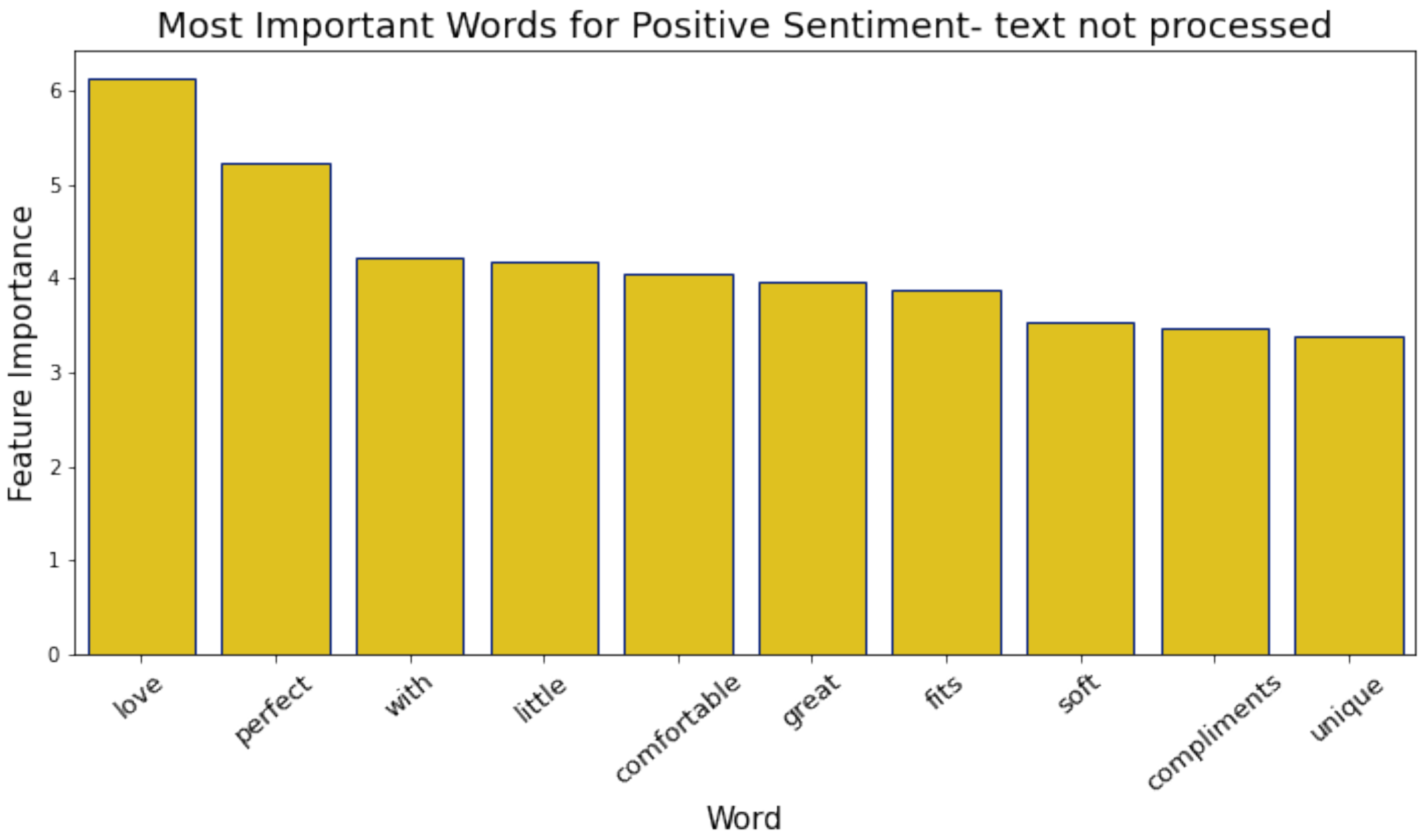
ROC Curves for LogisticRegression

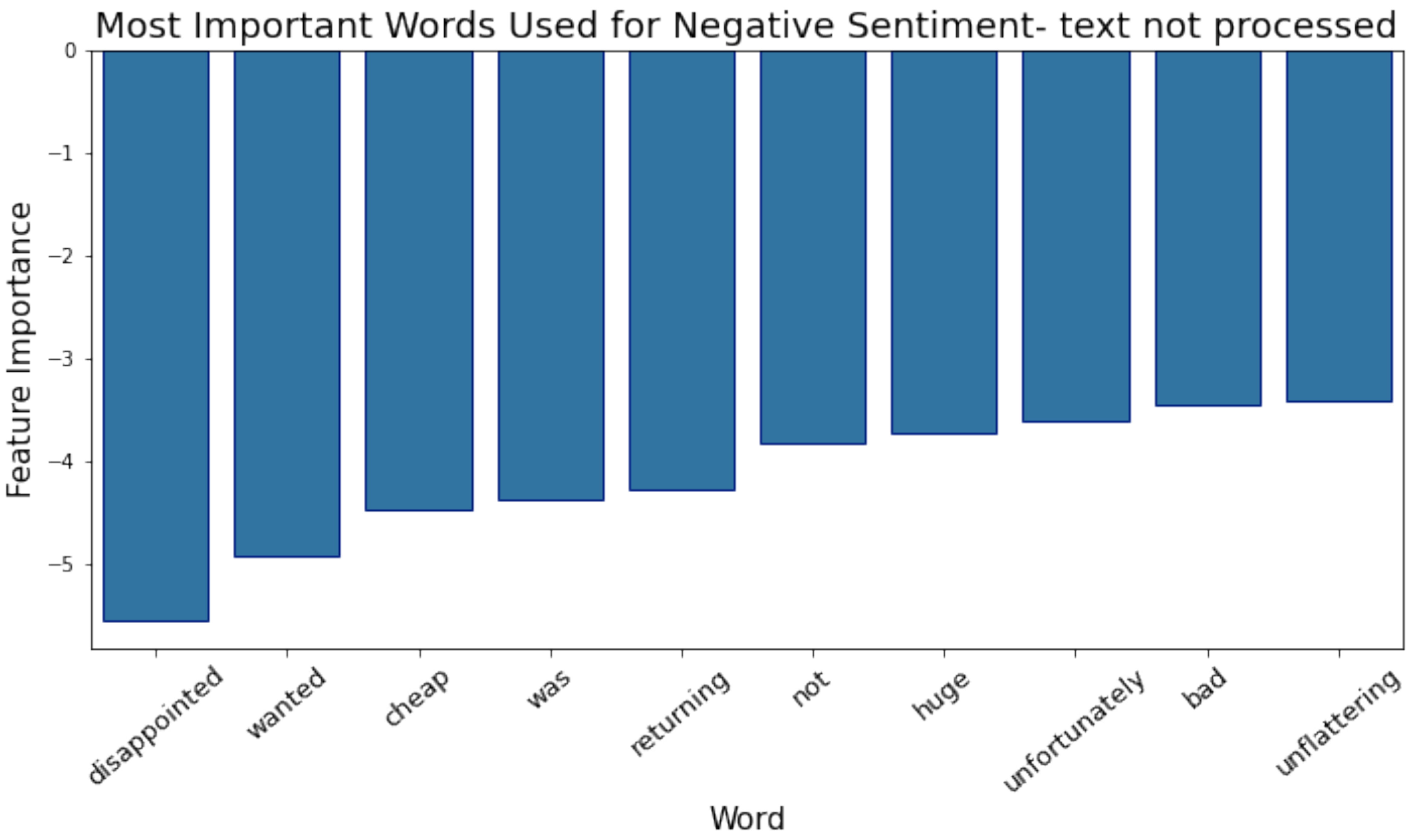
MOST IMPORTANT WORDS FOR SENTIMENT CLASSIFICATION - PRE PROCESSED TEXT





MOST IMPORTANT WORDS FOR SENTIMENT CLASSIFICATION - TEXT NOT PROCESSED

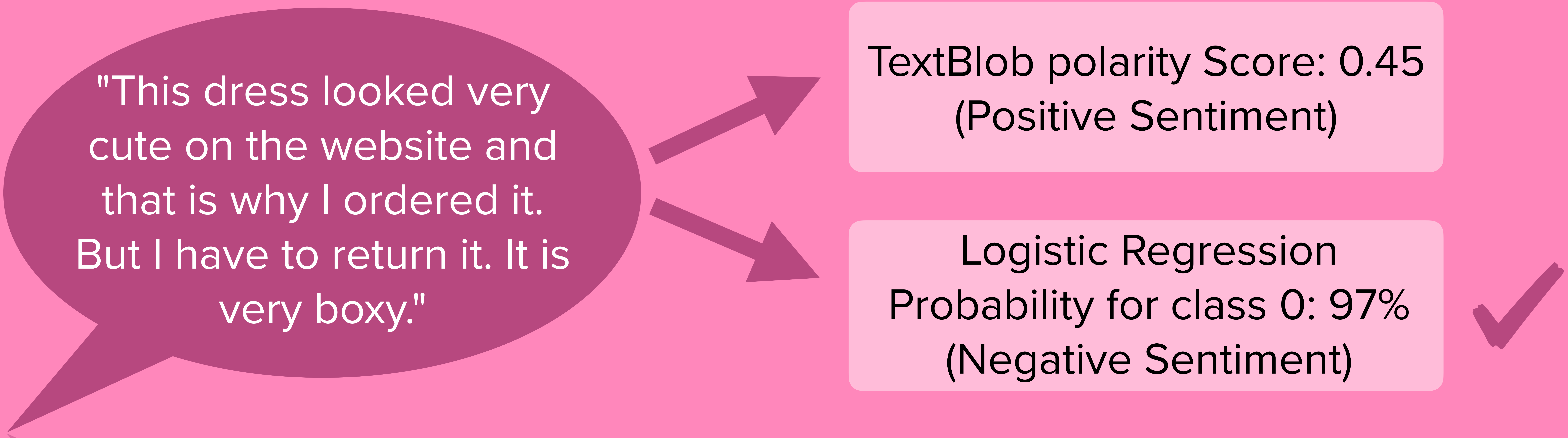




- The important words on processed and not processed texts are very similar!

TEXT BLOB AND LOGISTIC REGRESSION COMPARISON

Unseen text:



"This dress looked very cute on the website and that is why I ordered it. But I have to return it. It is very boxy."

TextBlob polarity Score: 0.45
(Positive Sentiment)

Logistic Regression
Probability for class 0: 97%
(Negative Sentiment) ✓

TAKE AWAYS

1. Always start with the simplest method!!
2. Noise reduction techniques are not helpful with sentiment analysis of this dataset

NEXT STEPS

1. Try different text datasets
2. Tune other hyper parameters
3. Try a deep learning model
4. Build an interactive sentiment analyzer which allows user-inputted reviews and give predictions on its sentiment where the users can help the model learn when it makes a wrong prediction

‘THANK YOU’

Bahar Biazar

baharbiazar@gmail.com

linkedin.com/in/bahar-biazar

github.com/baharbiazar