

COSMETICS









Data Science team at Sephora



WHO ARE YOU?

Social Media, Marketing and Management team at Sephora





AGENDA

0

INTRODUCTION

Background & Problem Statement

2

EDA

Data Cleaning & EDA

3

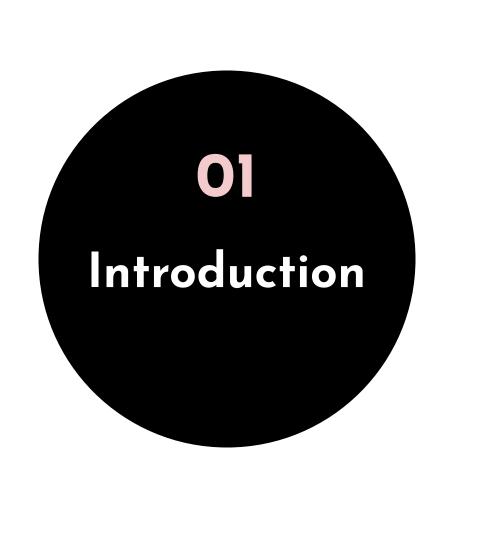
MODELING

Model Evaluation

4

CONCLUSION

Future Enhancement









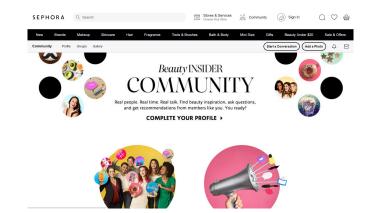






BACKGROUND

- We have recently created our Sephora Beauty Insider Community Forum to help shoppers forge personal connections with like-minded beauty enthusiasts
- Our marketing team has identified the forum as a way to obtain data and insights directly from our customers and improve our understanding of customers' needs and wants.
- From these insights, we can better market and improve our product line according to customers' preferences and drive more sales.







PROBLEM STATEMENT

- Our Social Media department has informed us that users of our newly created Sephora Beauty Insider Community Forum have been posting in the wrong group.
- The makeup and perfumes posts are posted in the General Discussion page and we need to find a way to classify their posts into **Makeup** and **Perfumes** categories for better analysis of their needs and interests.







OBJECTIVES

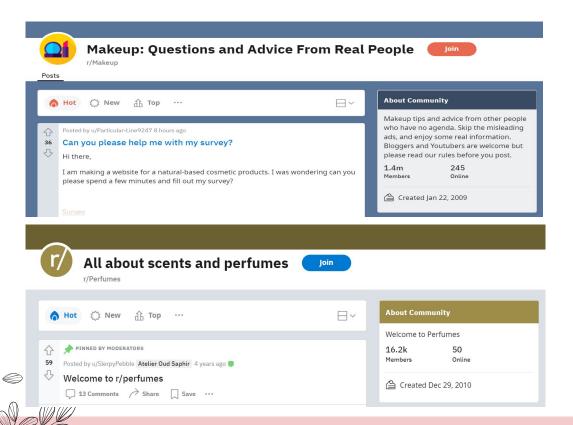
 Use Reddit posts from r/Makeup and r/Perfumes to create a text classifier model that can classify posts into the 2 categories with accuracy more than 50% (baseline accuracy).

We can then use the model to classify the posts in our Sephora forum.

 Find out the trending words in each subreddit to guide our marketing department



SUBREDDITS



• r/Makeup
1.4m members

Community where people share makeup tips and advice

• r/Perfumes
16.2k members

Community where people discuss about scents and perfumes

DATA COLLECTION

- We used Pushshift's API to scrape posts from each of the two subreddits
 - r/Makeup 5000 most recent posts dating back to April 12, 2022 r/Perfumes 5000 most recent posts dating back to January 13, 2021
- We kept columns 'Subreddit', 'Title' and 'Selftext' which contain the category and text required for our EDA and modelling.





















- Combine the title and selftext columns
- Remove words such as "[removed]", "[deleted]", "NaN"
- Remove punctuations
- Tokenize the posts
- Remove stop words such as "I", "we", "and" etc
- Lemmatize the words
- Remove words with numbers, non-english words & additional stopwords























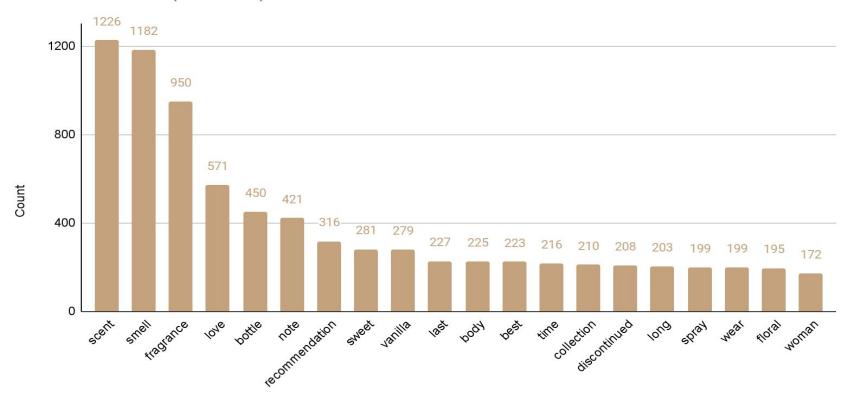
Perfume Subreddit:

- Scent
- **Smell**
- Fragrance
- Love
- Bottle
- Note





Count vs Words (Perfume)









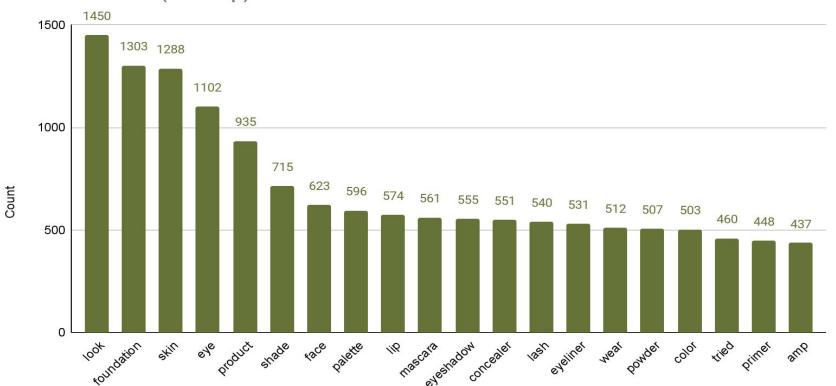
Makeup Subreddit:

- Look
- Foundation
- **Skin**
- Eye
- Product
- Shade





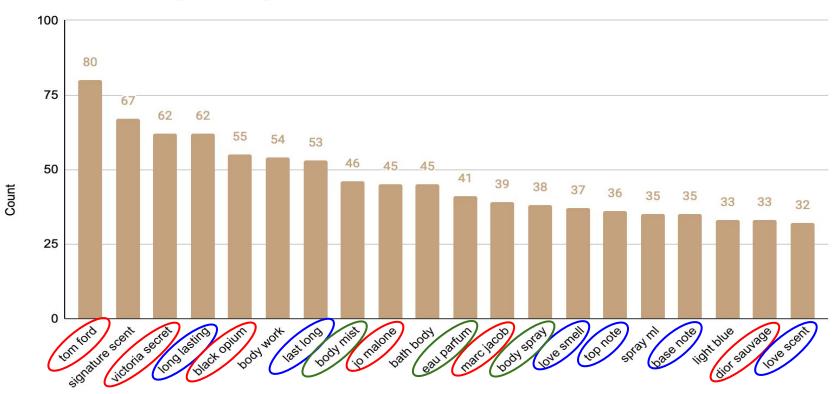
Count vs Words (Makeup)





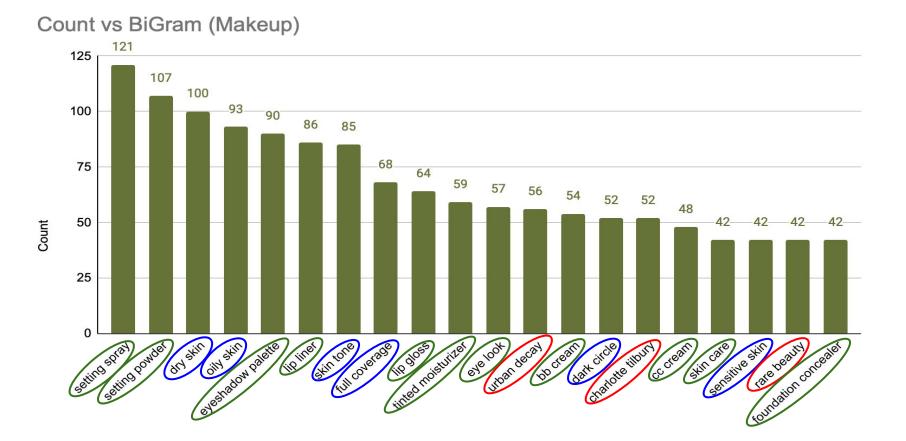
BIGRAM - PERFUME







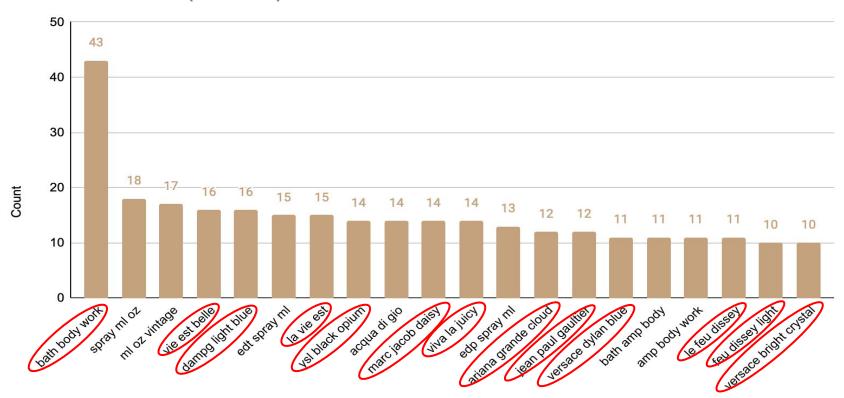






TRIGRAM - PERFUME

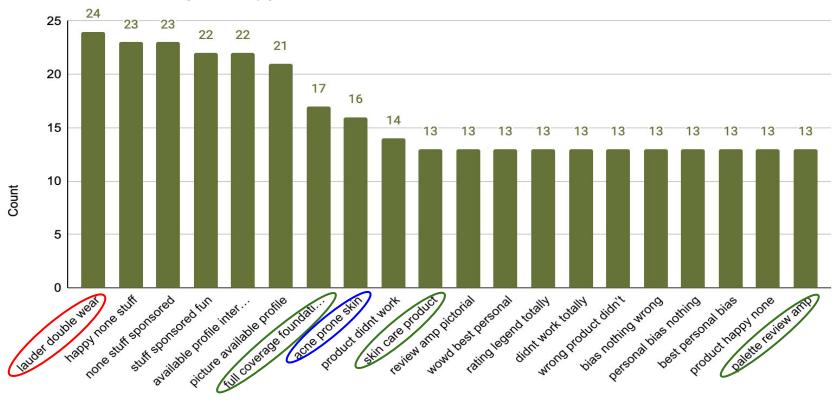














PERFUMES:

Tom Ford

Bath & Body Works

Jean Paul Gaultier

- **Secret**
- OBC Light Blue

Versace Dylan Blue

Black Opium

6 La Vie Est Belle

6 La Feu Dissy Light

Jo Malone

Viva La Juicy

Versace Bright Crystal

Marc Jacob

Marc Jacobs daisy

Oior Sauvage

Ariana Grande Cloud



MAKEUP:

- Setting Spray
- Setting Powder
- Eyeshadow Palette
- Lip Liner
- Lip Gloss
- **M** Tinted Moisturiser

- Eye Look Makeup
- BB Cream
- **CC Cream**
- Skin Care Products
- **Foundation Concealer**
- Full Covereage Foundation



RECOMMENDED FEATURES

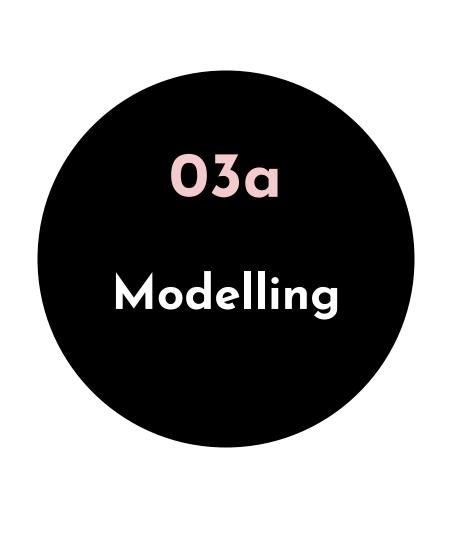


PERFUME:

- Cong Lasting
- **Solution** Love Smell
- Top Note
- Base Note

MAKEUP:

- M Dry Skin
- Oily Skin
- **Skin Tone**
- Dark Circle
- Sensitive Skin
- Acne Prone Skin















Model Selection

- Goal: Classify posts by their group
- Classification Models used:
 - Bernoulli Naïve Bayes
 - Multinomial Naïve Bayes
 - Gaussian Naïve Bayes
 - Logistic Regression
 - K-Nearest Neighbour



Vectorised Data Selection

- 1. Count
- 2. N-grams
 - a. Bigrams
 - b. Trigrams
- 3. TF-IDF
 - a. Term Frequency-Inverse Document Frequency





Vectorisation Type	Model Type	Train Results	Test Results	
Count	Bernoulli NB	0.91082	0.90778	
Count	Multinomial NB	0.96034	0.95417	
Count	Gaussian NB	0.92742	0.92194	
N-Gram	Bernoulli NB	0.87436	0.86889	
N-Gram	Multinomial NB	0.84046	0.83472	
N-Gram	Gaussian NB	0.85581	0.85139	
TF-IDF	Bernoulli NB	0.96027	0.95130	
TF-IDF	Multinomial NB	0.97013	0.96361	
TF-IDF	Gaussian NB	0.96325	0.94222	





Vectorisation Type	Model Type	Train Results	Test Results
Count	Bernoulli NB	0.91082	0.90778
Count	Multinomial NB	0.96034	0.95417
Count	Gaussian NB	0.92742	0.92194
N-Gram	Bernoulli NB	0.87436	0.86889
N-Gram	Multinomial NB	0.84046	0.83472
N-Gram	Gaussian NB	0.85581	0.85139
TF-IDF	Bernoulli NB	0.96027	0.95130
TF-IDF	Multinomial NB	0.97013	0.96361
TF-IDF	Gaussian NB	0.96325	0.94222





Vectorisation Type	Model Type	Train Results Test Resul	
Count	Bernoulli NB	0.91082	0.90778
Count	Multinomial NB	Multinomial NB 0.96034 0.95	
Count	Gaussian NB	0.92742	0.92194
N-Gram	Bernoulli NB	0.87436	0.86889
N-Gram	Multinomial NB	0.84046	0.83472
N-Gram	Gaussian NB	0.85581	0.85139
TF-IDF	Bernoulli NB	0.96027	0.95130
TF-IDF	Multinomial NB	0.97013	0.96361
TF-IDF	Gaussian NB	0.96325	0.94222



Vectorised Data Selection

Model Type	Train Results	Test Results
Bernoulli NB	0.91082	0.90778
Multinomial NB	0.96034	0.95417
Gaussian NB	0.92742	0.92194
Bernoulli NB	0.87436	0.86889
Multinomial NB	0.84046	0.83472
Gaussian NB	0.85581	0.85139
Bernoulli NB	0.96027	0.95130
Multinomial NB	0.97013	0.96361
Gaussian NB	0.96325	0.94222
	Bernoulli NB Multinomial NB Gaussian NB Bernoulli NB Multinomial NB Gaussian NB Bernoulli NB Multinomial NB	Bernoulli NB 0.91082 Multinomial NB 0.96034 Gaussian NB 0.92742 Bernoulli NB 0.87436 Multinomial NB 0.84046 Gaussian NB 0.85581 Bernoulli NB 0.96027 Multinomial NB 0.97013



Model Optimisation

Model	Before	After
Bernoulli Naïve Bayes	binarize = 0.0	binarize = 0.2
Multinomial Naïve Bayes	fit_prior = True	fit_prior = False
Gaussian Naïve Bayes	n/a	n/a
Logistic Regression	solver = 'liblinear'	n/a
K-Nearest Neighbour	n = 3	n = 2

Model Evaluation

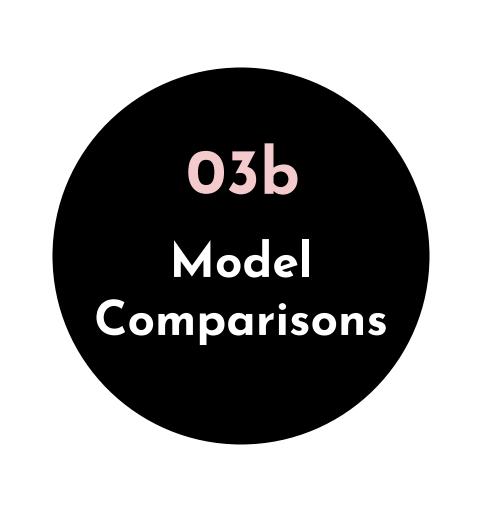
Vectorisation Type	Model Type	Train Results	Test Results	
TF-IDF	Bernoulli NB	0.96027	0.95130	
TF-IDF	Multinomial NB	0.97013	0.96361	
TF-IDF	Gaussian NB	0.96325	0.94222	
TF-IDF	Logistic Regression	0.98020	0.97000	
TF-IDF	K-Nearest Neighbour	0.95701	0.75278	



Model Evaluation

Vectorisation Type	Model Type	Train Results	Test Results	
TF-IDF	Bernoulli NB	0.96027	0.95130	
TF-IDF	Multinomial NB	0.97013	0.96361	
TF-IDF	Gaussian NB	0.96325	0.94222	
TF-IDF	Logistic Regression	0.98020	0.97000	
TF-IDF	K-Nearest Neighbour	0.95701	0.75278	















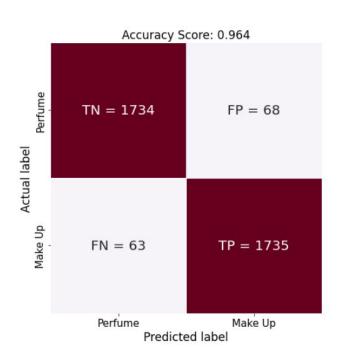


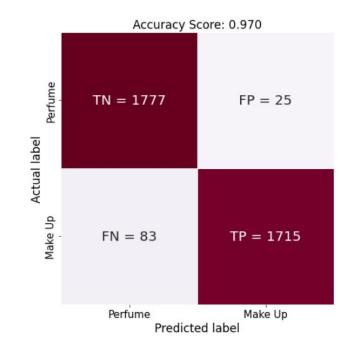




Confusion Matrix

Multinomial NB









Classification Report

Multinomial NB

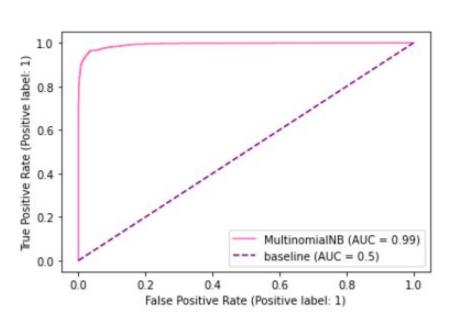
	precision	recall	f1-score	support	precision	recall	f1-score	support
Perfume	0.9649	0.9623	0.9636	1802	0.9554	0.9861	0.9705	1802
Make Up	0.9623	0.9650	0.9636	1798	0.9856	0.9538	0.9695	1798
accuracy			0.9636	3600			0.9700	0.97
macro avg	0.9636	0.9636	0.9636	3600	0.9705	0.9700	0.9700	3600
weighted avg	0.9636	0.9636	0.9636	3600	0.9705	0.9700	0.9700	3600

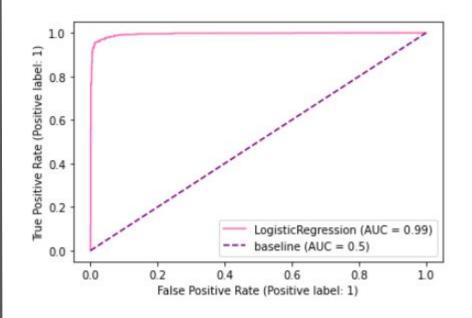




ROC Curve

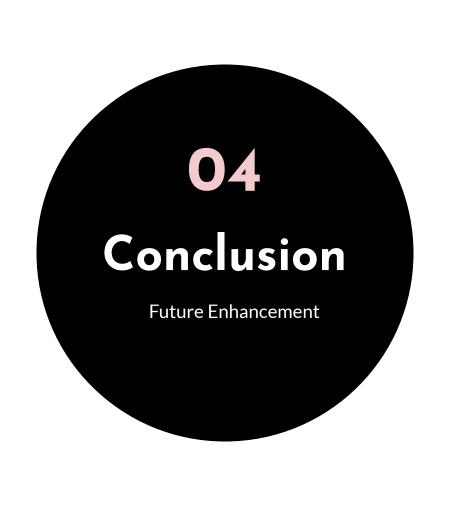
Multinomial NB

















Conclusion



Logistic Regression

Improve Efficiency and Effectiveness of Marketing & Social Media Team



Trending Words

Develop Marketing Strategies as per trending words



Top Predictors

Identify the key words that can improve our product and services









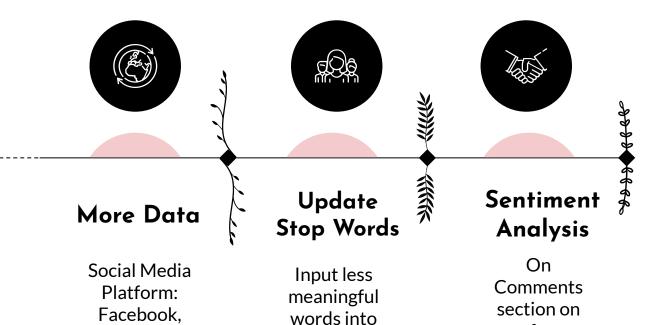
PERFUMES:

- **Tom Ford**
- Victoria Secret
- Black Opium
- Jo Malone
- Marc Jacob
- Oior Sauvage

FEATURES:

- Dry Skin
- Oily Skin
- Skin Tone
- O Dark Circle
- Sensitive Skin
- (Acne Prone Skin





stop words

list

Twitter,

Instagram

our forum

Future Enhancements

THANKS!