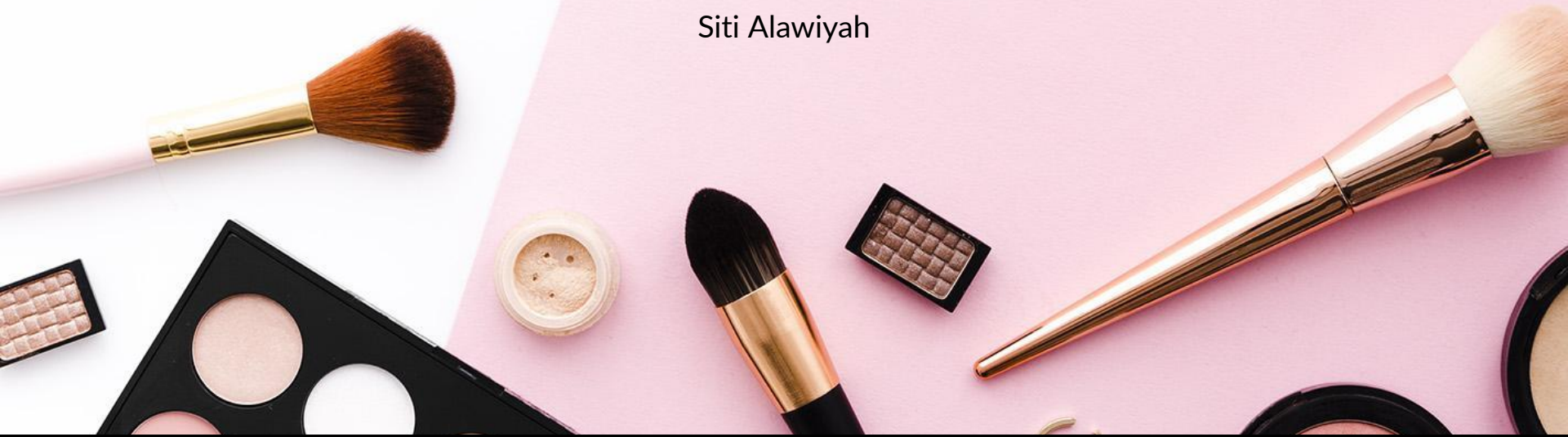




# **Project 3: Classification of 2 Subreddits: Makeup & Fragrance**

---

Siti Alawiyah





# AGENDA



1

**Problem  
Statement**

2

**Process**

3

**EDA**

4

**Modelling**

5

**Results**

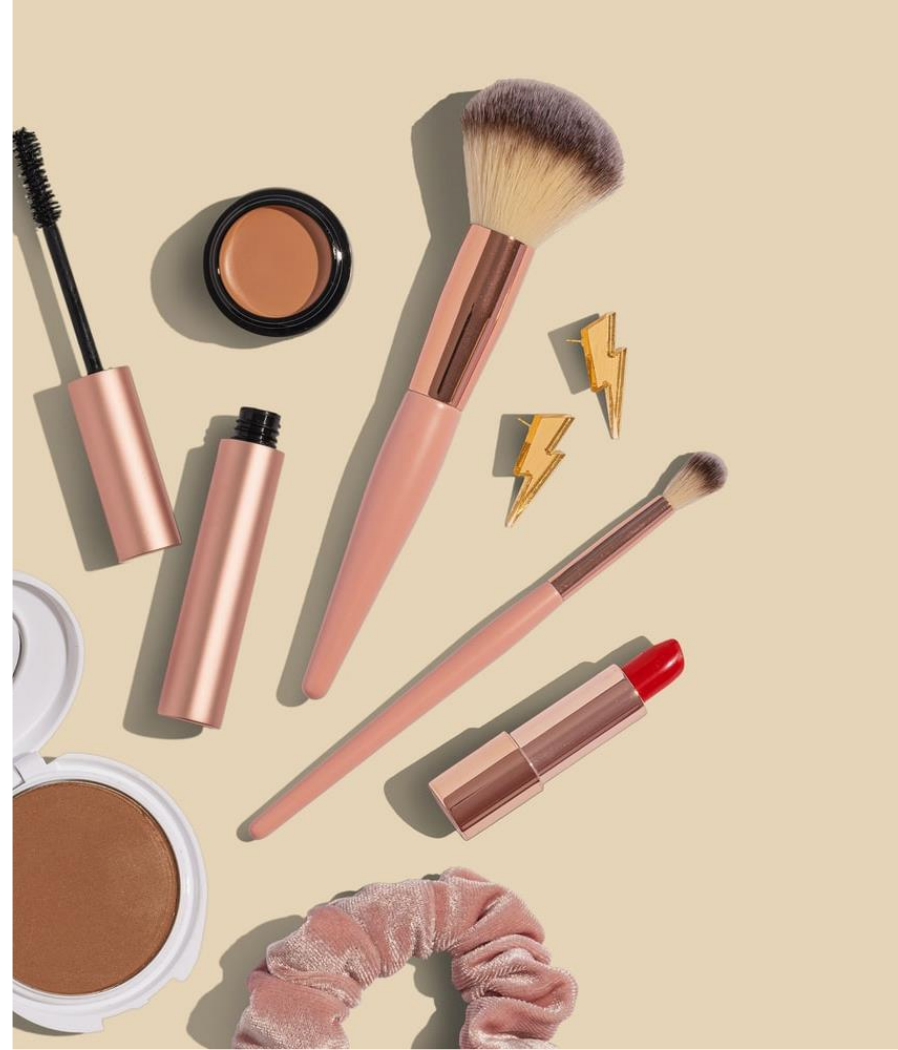
The slide features a central black circle containing the text '01 PROBLEM STATEMENT'. The number '01' is in a light pink color, while 'PROBLEM STATEMENT' is in white. The background is white and decorated with delicate black line-art floral and leaf patterns in the four corners.

# 01 PROBLEM STATEMENT



# Problem Statement

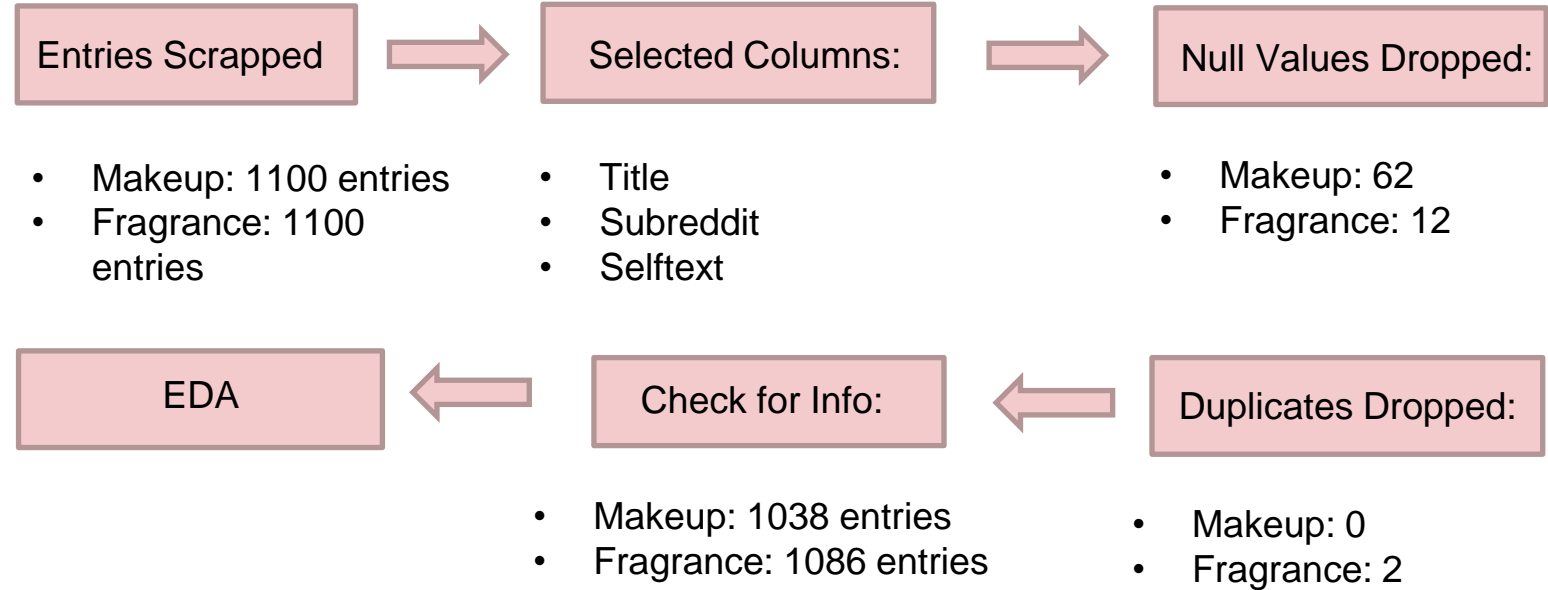
- Use Pushshift API to collect 2 subreddits category: makeup and fragrance
- Use NLP to train a classifier on which subreddit a post has been given to




The page features a central black circle containing the text '02 PROCESS'. The number '02' is in a light pink color, and the word 'PROCESS' is in white. The background is white, decorated with black line-art floral illustrations in the corners: top-left, top-right, bottom-left, and bottom-right. These illustrations include various leaves, stems, and small flowers.

# 02 PROCESS

# Process

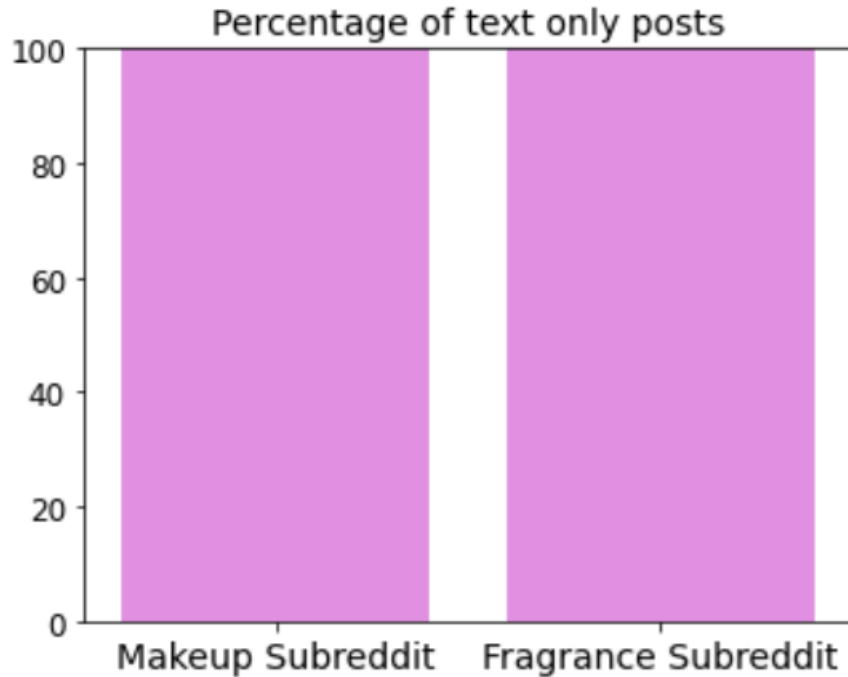


The page features a central black circle containing the text '03 EDA'. The corners of the page are decorated with delicate, black-and-white line drawings of various plants and flowers. In the top-left corner, there are sprigs of leaves and small berries. The top-right corner shows a branch with oval-shaped leaves. The bottom-left corner contains a cluster of leaves and a small flower. The bottom-right corner features a branch with leaves and two small, round fruits.

03

EDA

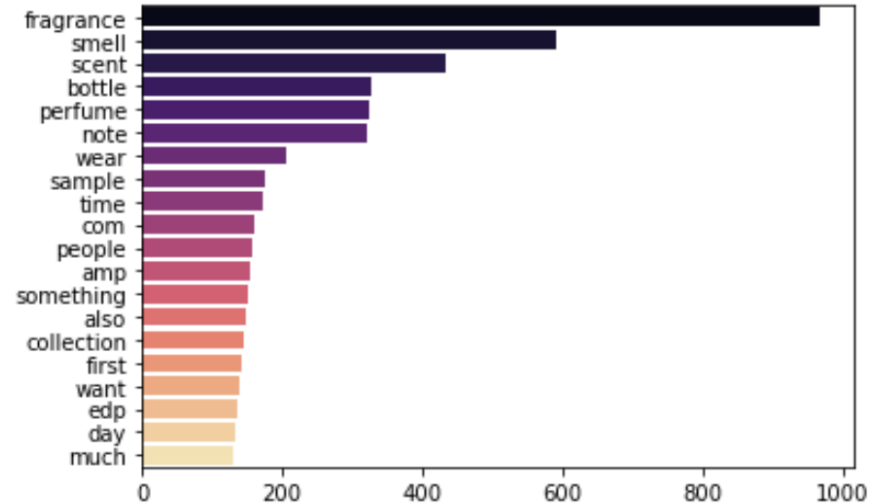
# Texts only posts

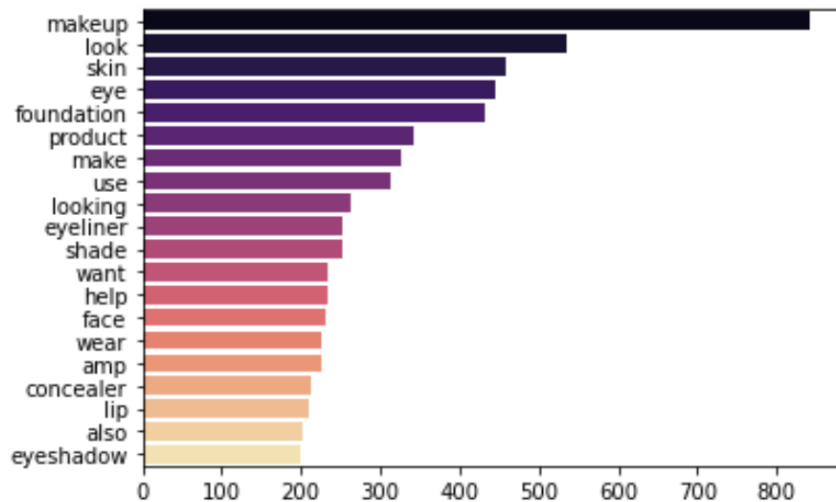


Both subreddits are 100% full of texts which is good as we will need text to be able to classify using our models

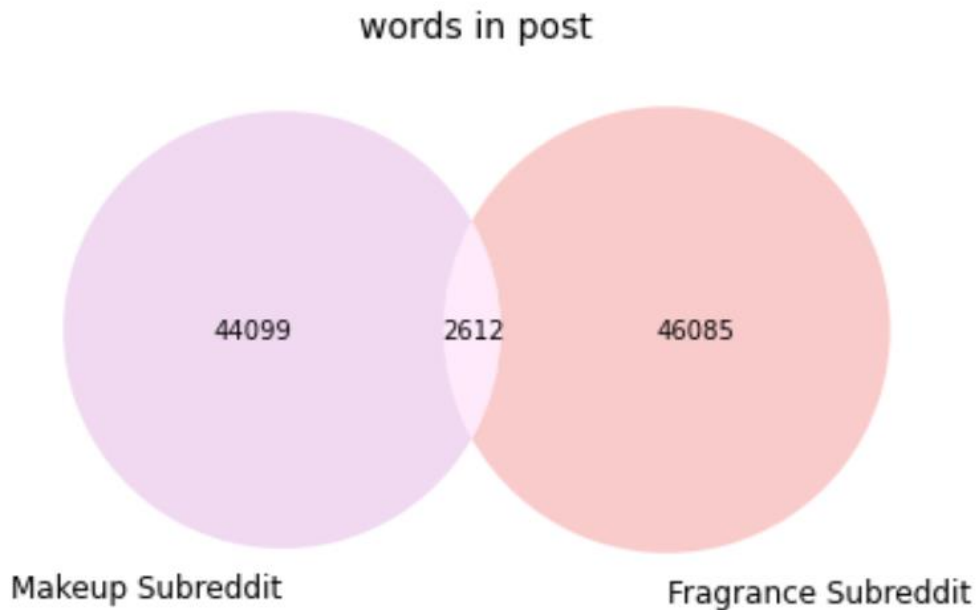


# Top words for Fragrance Subreddit





# Common Words in Subreddits



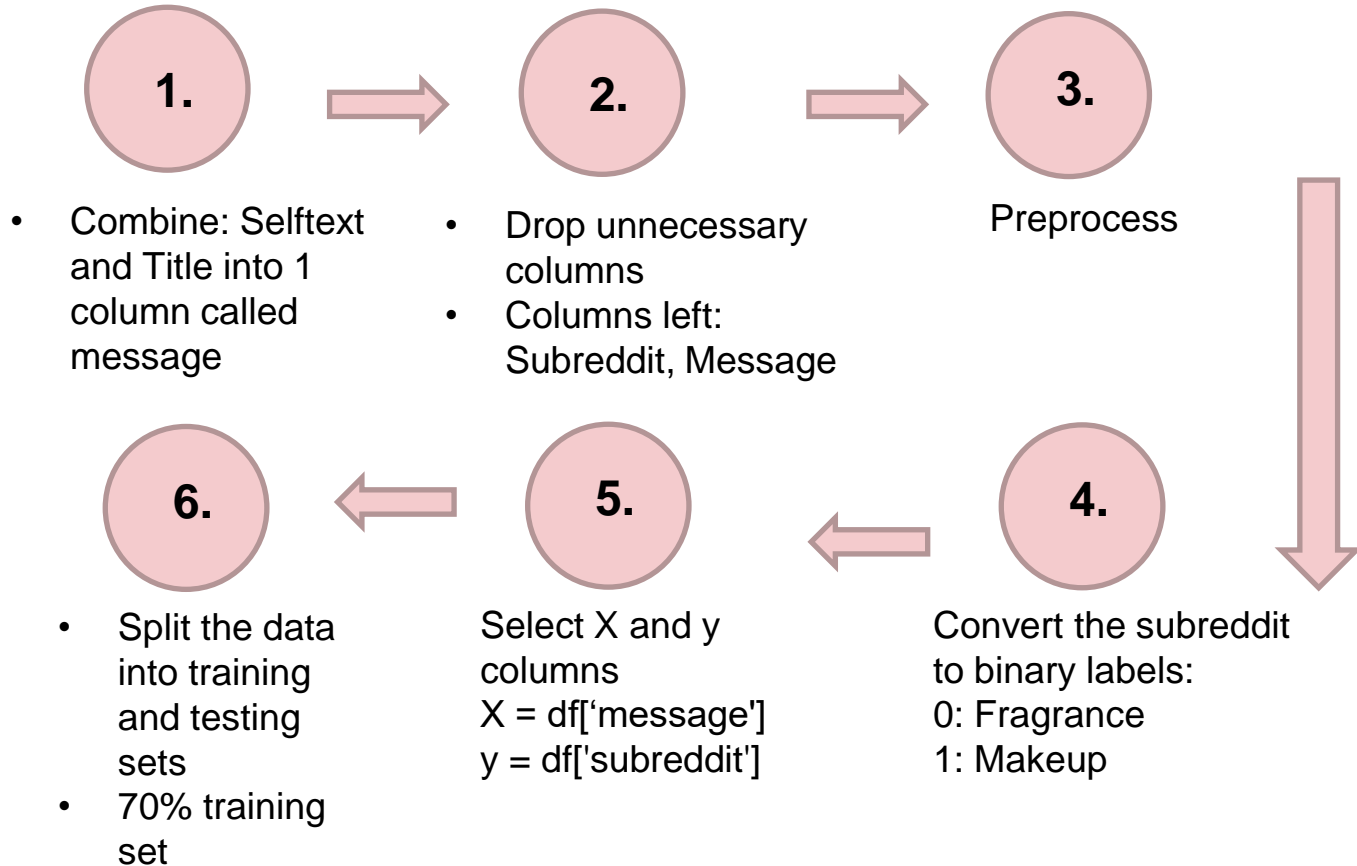
There are 2612 words that are common in both makeup and fragrance subreddit

The slide features a central black circle containing the text '04 Modelling'. The number '04' is in a light pink color, while 'Modelling' is in white. The background is white and decorated with delicate black line art of various plants and flowers in the corners: top-left, top-right, middle-left, middle-right, bottom-left, and bottom-right.

04

**Modelling**

# Pre-Modelling Process



# Types of Models

**RandomForestClassifier**

**Naïve  
Bayes**

**Pipeline**

CountVectorizer, RandomForestClassifier  
(random\_state=42)

("vec", None), ("model", MultinomialNB())

**PipeParameters**

```
'cvec__max_features': [800, 900, 1_000],  
'cvec__stop_words': [None, "english"],  
'cvec__min_df': [2, 3],  
'cvec__max_df': [.9, .95],  
'cvec__ngram_range': [(1, 1),(1, 2),(2,3)],  
'rf__n_jobs': [-1],  
'rf__n_estimators': [100,150,200],  
'rf__max_depth': [None, 1, 2, 3, 4, 5, 6],
```

```
"vec": [CountVectorizer(), TfidfVectorizer()],  
'vec__max_features': [800, 900, 1_000],  
'vec__stop_words': [None, "english"],  
'vec__min_df': [2, 3],  
'vec__max_df': [.9, .95],  
'vec__ngram_range': [(1, 1),(1, 2),(2,3)]
```

**GridSearchCV**

cv =3, scoring = AUC, accuracy

cv=3, scoring = AUC, accuracy

The slide features a central black circle containing the text '05 Results & Conclusion'. The number '05' is in a light pink color, while 'Results & Conclusion' is in white. The background is white and decorated with black line-art floral illustrations in the corners: top-left, top-right, bottom-left, and bottom-right.

# 05 Results & Conclusion

# Scores

RF

NB

AUC Train Score	0.987	<u>0.994</u>
AUC Test Score	0.977	<u>0.989</u>
Accuracy Train Score	0.905	<u>0.954</u>
Accuracy Test Score	0.887	<u>0.933</u>



# Best Parameters

Best Params	RF Classifier	NB
Cvec max df	0.9	0.9
Cvec max features	1000	1000
Cvec min df	3	3
Ngram_range	1,1	1,1
Stopwords	english	english
Max depth	6	-
N estimators	200	-
Top Features:	Look, smell, eye, foundation, perfume, scent, bottle, eyeshadow, face, concealer	



# Conclusion

- All models can classify which subreddit a post has been given to.
- Best model recommendation: Naive Bayes
- difference between the test/training AUC score and accuracy score is approximately 0.01
- AUC score and accuracy score is also the highest
- Future steps: explore other models such as GradientBoost. Try removing the top features and see if the classifiers are able to classify accurately.

A decorative border on the left side of the slide, featuring a pink circular background with black line-art illustrations of leaves and small berries.

**THANK YOU!**

A decorative border on the right side of the slide, featuring a pink circular background with black line-art illustrations of leaves and small berries.