Project 3:
Classification of 2 Subreddits:
Makeup & Fragrance

Siti Alawiyah

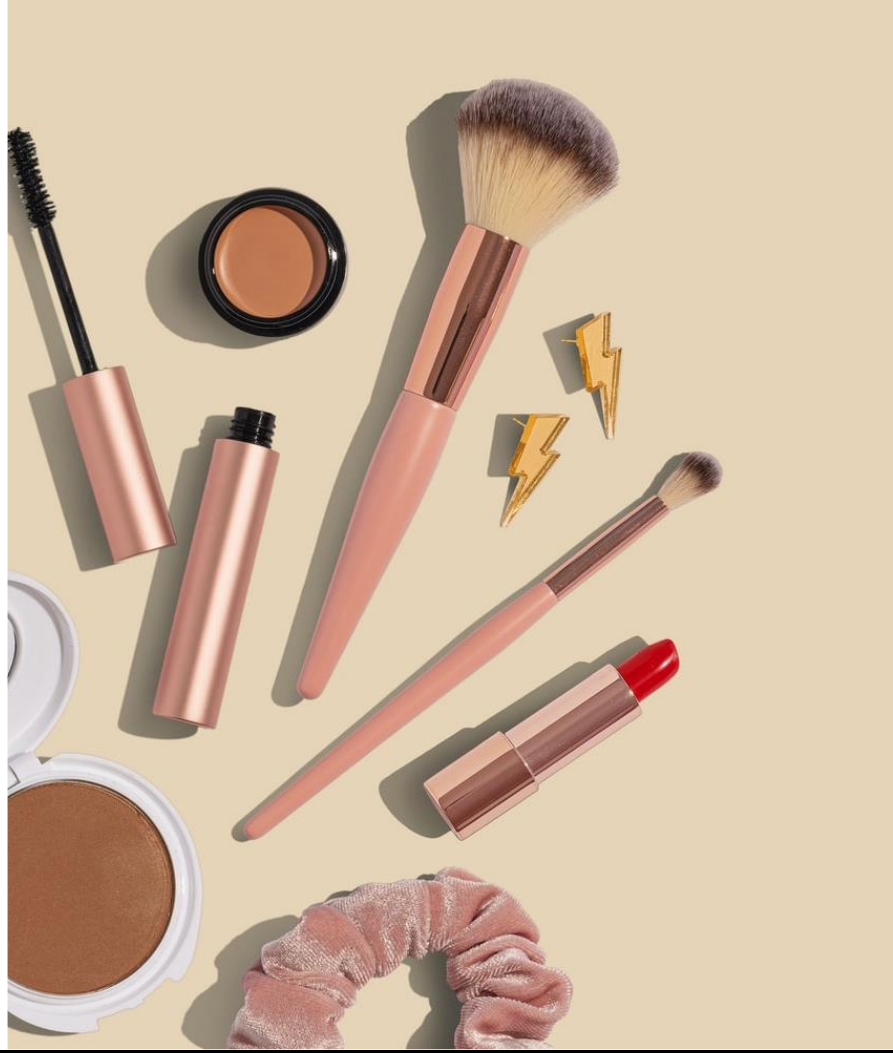
AGENDA

1. Problem Statement
2. Process
3. EDA
4. Modelling
5. Results

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



02

PROCESS

Process

Entries Scrapped

- Makeup: 1100 entries
- Fragrance: 1100 entries

Selected Columns:

- Title
- Subreddit
- Selftext

Null Values Dropped:

- Makeup: 62
- Fragrance: 12

Duplicates Dropped:

- Makeup: 0
- Fragrance: 2

Check for Info:

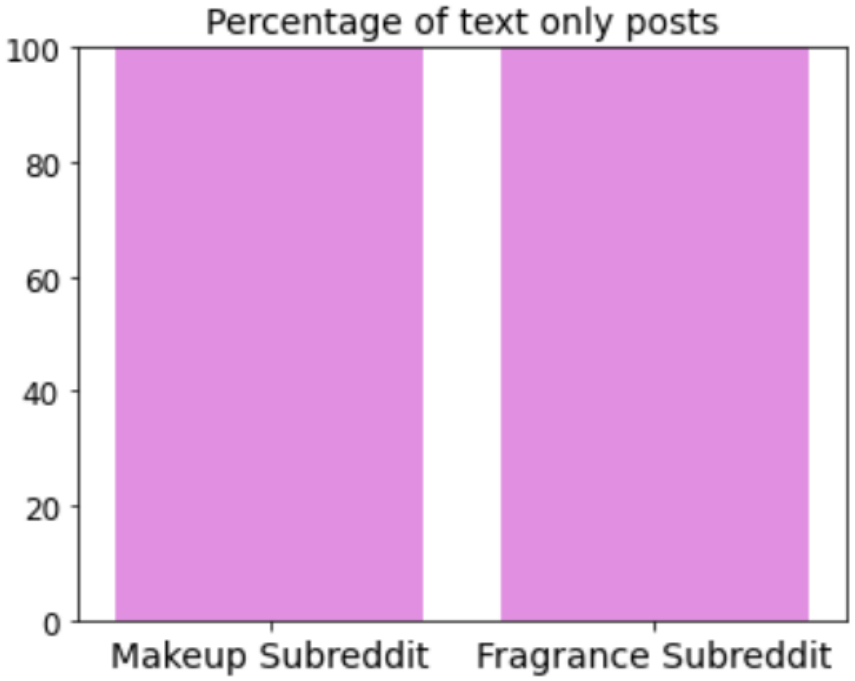
- Makeup: 1038 entries
- Fragrance: 1086 entries

EDA

03

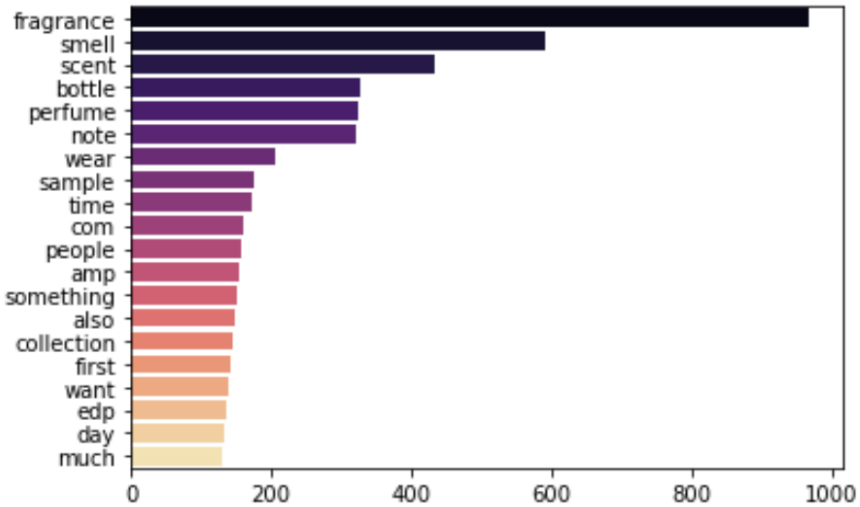
EDA

Texts only posts

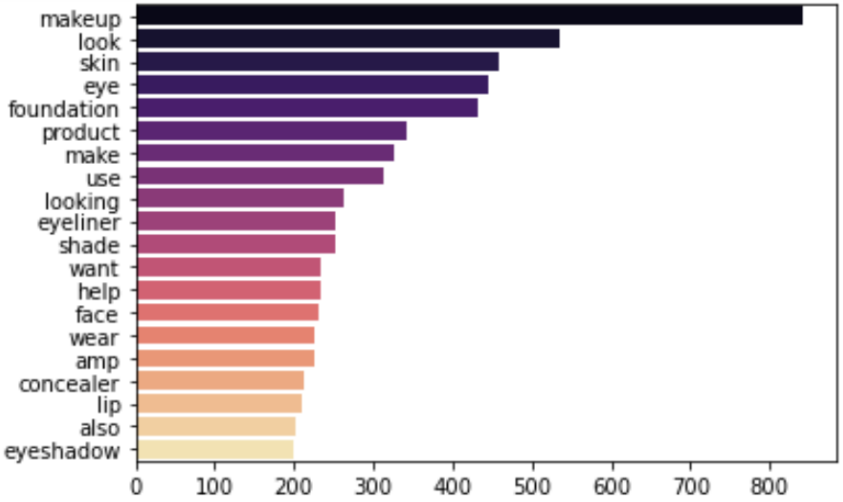


Both subreddits are 100% full of texts which is good as we will need text to be able to classify using our models

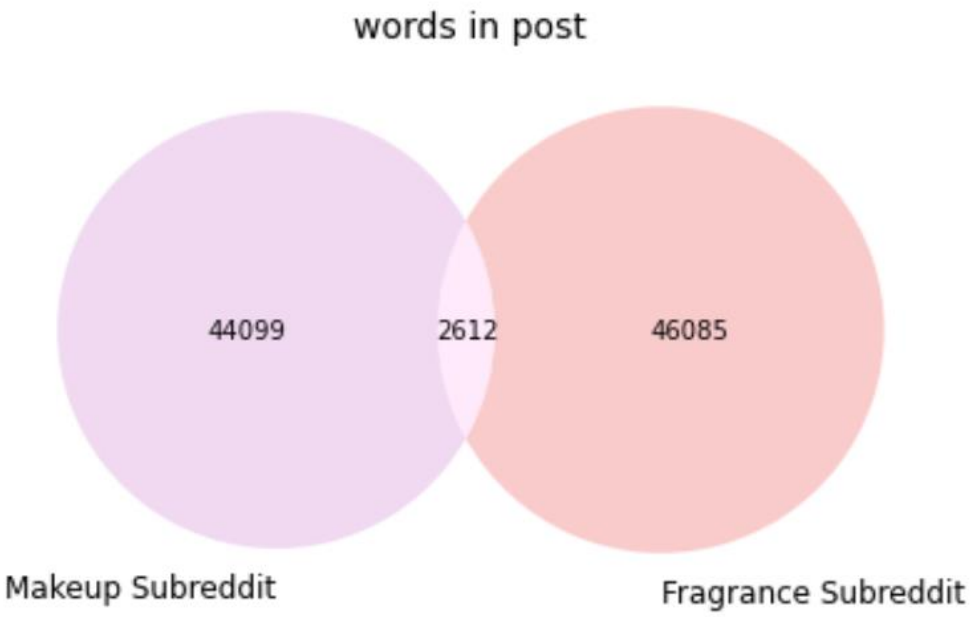
Top words for Fragrance Subreddit



Top words for Makeup Subreddit



Common Words in Subreddits



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

04

Modelling

Pre-Modelling Process

1.

- Combine: Selftext and Title into 1 column called message

2.

- Drop unnecessary columns
- Columns left: Subreddit, Message

3.

Preprocess

4.

Convert the subreddit to binary labels:
0: Fragrance
1: Makeup

5.

Select X and y columns
X = df[‘message']
y = df['subreddit']

6.

- Split the data into training and testing sets
- 70% training set

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

05
Results & Conclusion

Scores

	RF	NB
AUC Train Score	0.987	0.994
AUC Test Score	0.977	0.989
Accuracy Train Score	0.905	0.954
Accuracy Test Score	0.887	0.933

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.

THANK YOU!