

Classification of Subreddit Post

Problem Statement

Building a classification model to predict the correct subreddit between two subreddit post (Cooking & OutoftheLoop)



Data Extraction

Data extraction through the API call.

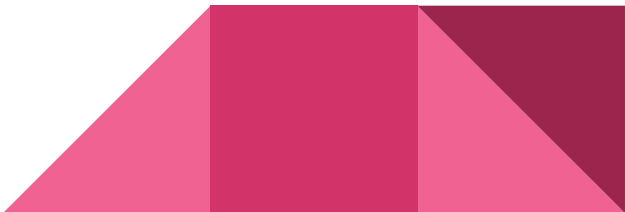
We collected nearly 2000 subreddit from two category

```
Outofloop.shape
```

```
(2465, 101)
```

```
Cooking.shape
```

```
(2492, 102)
```



Data Cleaning / Preprocessing

- ❖ Lemmatize
- ❖ Remove Stopwords
- ❖ Stemming
- ❖ Remove weblinks 'https'
- ❖ Remove Special Characters
- ❖ Remove email id
- ❖ Remove Duplicates



EDA - Word Cloud



The ouoftheloop subreddit doesnt have most frequent word like in cooking subreddit

Naive Bayes

Count Vectorize the text in the subreddit posts after preprocessing & merging datasets and use Multinomial Naive Bayes

	precision	recall	f1-score	support
0	0.97	0.93	0.95	469
1	0.94	0.97	0.96	559
accuracy			0.95	1028
macro avg	0.95	0.95	0.95	1028
weighted avg	0.95	0.95	0.95	1028

True Negatives: 434

False Positives: 35

False Negatives: 15

True Positives: 544

Classification report for the Naive Bayes Model



Logistic Regression - Optimization

```
parameters = {'C': [0.001, 0.01, 0.1, 1, 10],  
              'class_weight': [None, 'balanced'],  
              'penalty': ['l1', 'l2']}
```

Optimized the logistic regression using Grid search to find the best parameters

	precision	recall	f1-score	support
0	0.98	1.00	0.99	469
1	1.00	0.98	0.99	559
accuracy			0.99	1028
macro avg	0.99	0.99	0.99	1028
weighted avg	0.99	0.99	0.99	1028

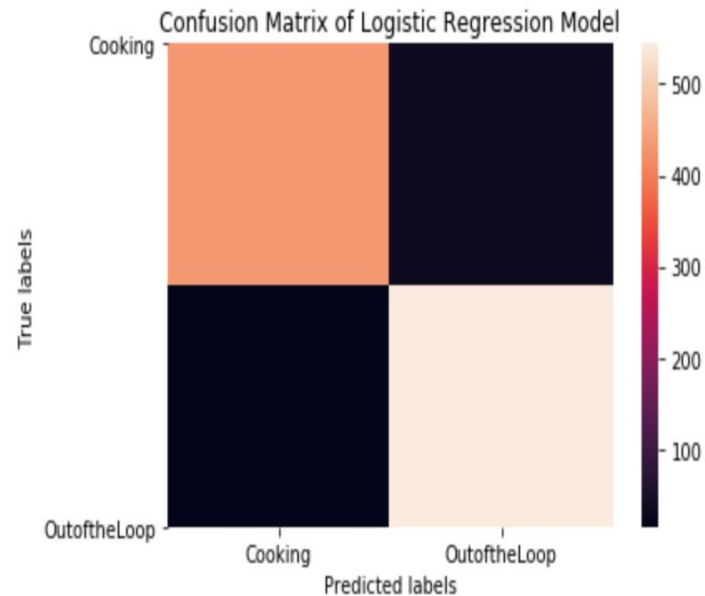
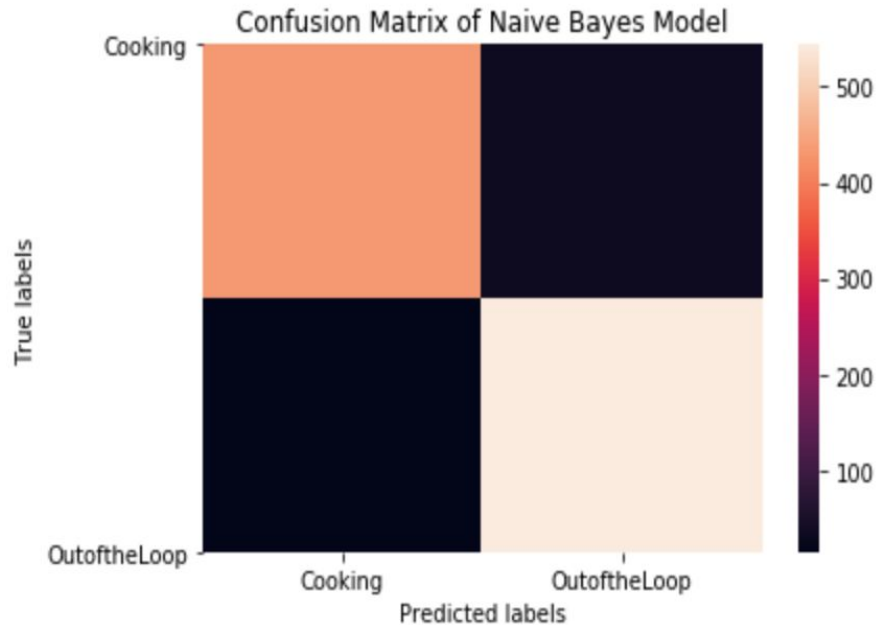
True Negatives: 469

False Positives: 0

False Negatives: 9

True Positives: 550

Confusion Matrix - Outcomes of two Models



Conclusion & recommendations

- ❖ Current Model with Accuracy of 99% can be deployed
- ❖ deploying the model on more unseen posts to further validate its accuracy.

