

# Project 3

## SubReddit Classifier



Muhammad Fadhil

# Contents

- Problem Statement
- Word Clouds
- Model Evaluation
- Conclusion



# Problem Statement


Reddit has identified that they have lost the tags for 2 of their subreddits for a few days and they are **r/buildapc** and **r/Cooking**.

Fortunately, there's a backup for the 2 subreddits. However, posts relating to the 2 subreddits during this period have had no classification.

You have been tasked to classify the tagless posts into their correct subreddits with an appropriate model.

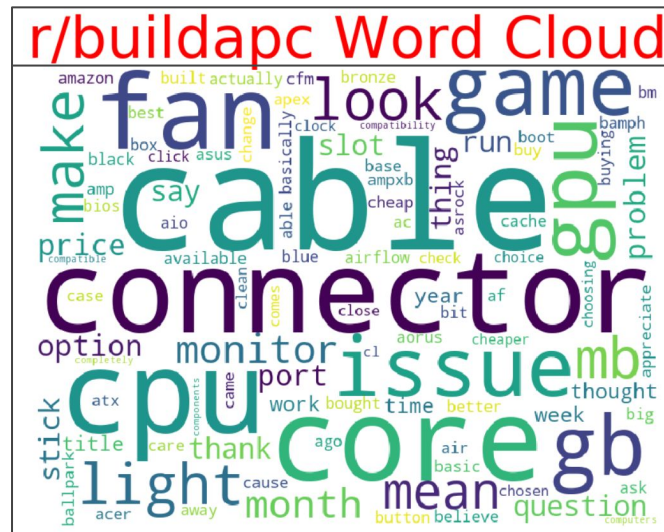
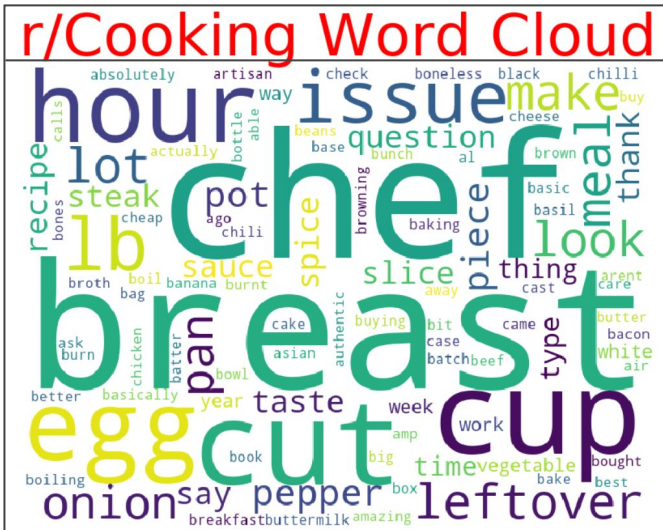
## Objective:

Correctly classify the subreddits, and select a model after comparing the efficacy and limitations of two classification models.



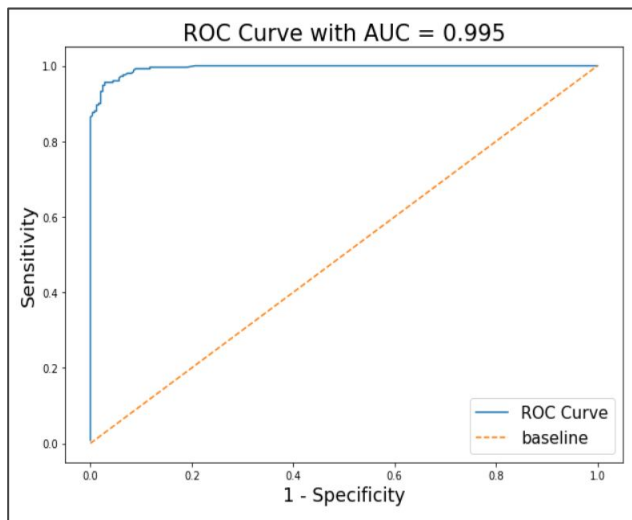
# Word Clouds

There's a stark difference between the most common words between the two subreddits

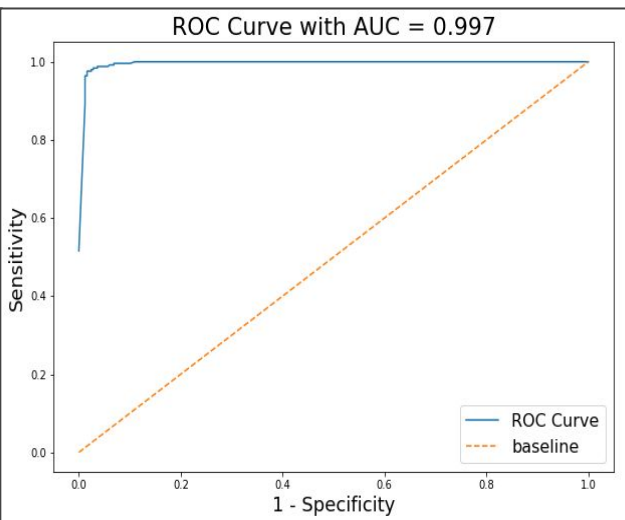


# Model Evaluation

Logistic Regression



Naive Bayes



Better separability does not equate to a better classifier as it does not consider the weights of the features in the model.

Our Naive Bayes had a better AUC score which represents better separability between our two classes.

# Conclusion

Naive Bayes was selected as the preferred Classifier as it had a better accuracy.

- Logistic Regression = 0.9557
- Naive Bayes = 0.9758

Our top 10 key features in Naive Bayes are:

- Cooking: recipe, make, wa, cooking, chicken, pan, time, sauce, making, know
- Buildapc: cpu, gb, case, wa, gaming, fan, amazon, power, know, gpu

