

**Postgraduate Certificate in Software Design with Artificial Intelligence**

**Advanced Machine Learning and Neural Networks - (AL\_KSAIG\_9\_1)**

**Minor Exercise 2**

Student ID: A00267948

Student Name: Daniel Foth

GIT: <https://github.com/DanielsHappyWorks/aml-text-classification>

Brief Description:

Using the Newsgroups dataset and the following machine learning methods

- SVM

- Naive Bayes

- Neural Network

You will need to use the techniques of stopping(removing small insignificant words eg I, the, you etc), stemming(removing the endings of words eg -ed -ing) and use of TF/IDF (Term Frequency over Item Document Frequency) to aid in the classification of the type of news report

This is a task which will require you to do some feature engineering to get decent accuracy

The submission will be the source code which will output a confusion matrix and overall accuracy of each classifier

Dataset: sklearn.datasets import fetch\_20newsgroups

Contents

[Data 3](#_Toc36383843)

[SVM 3](#_Toc36383844)

[Naive Bayes 3](#_Toc36383845)

[Neural Network 3](#_Toc36383846)

[Conclusion 4](#_Toc36383847)

# Data

The data set contains:

Classes 20

Samples total 18846

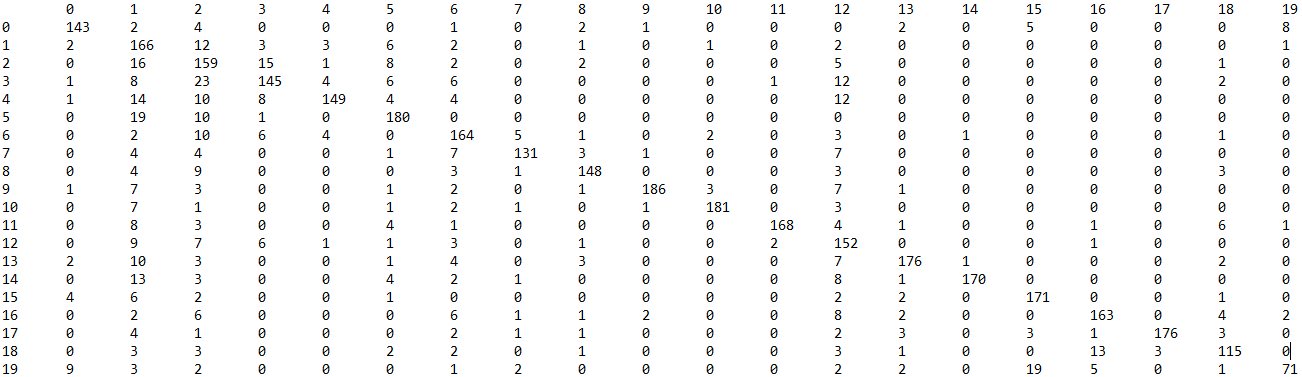
Dimensionality 1

Features text

# SVM

Accuracy 0.8259946949602122

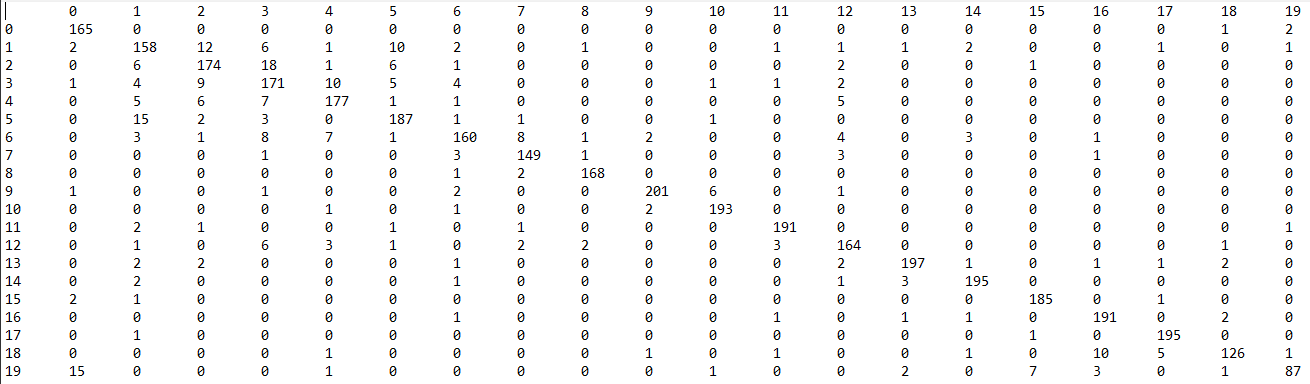
Confusion Matrix



# Naive Bayes

Accuracy: 0.9108753315649868

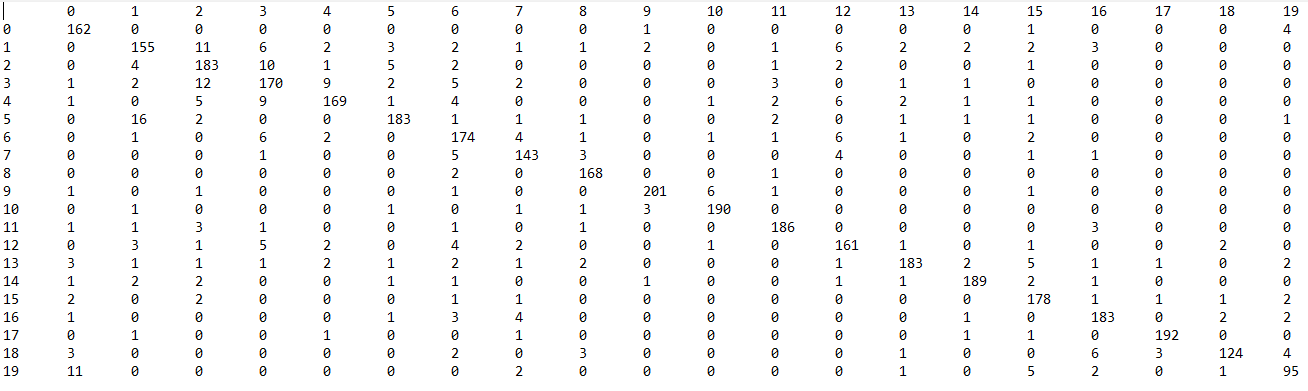
Confusion Matrix:



# Neural Network

Accuracy 0.8989389920424403

Confusion Matrix



# Conclusion

In conclusion processing this much text is very time consuming and makes for altering models very difficult. The best performance was achieved using Naïve Bays, but all of the models are really good as they are all over 80% accurate. Neural network came in at a close second but it’s harder to predict how it works in the background which still makes NB the preferred algorithm.