------------------------------------------------------------------------------------------------------------------------------------------

Lab Assignment: 2

Full name: Gardyan Priangga Akbar

Student number: 7695883

Email address: [gardyanakbar99@gmail.com](mailto:gardyanakbar99@gmail.com)

------------------------------------------------------------------------------------------------------------------------------------------

**Text Classification with Naïve Bayes Classifier**

This lab assignment aims to explore the use of Naïve Bayes classifier for the purpose of text classification. This includes developing an understanding of Bayes’ formula and the preprocessing steps needed to train a Naïve Bayes model for text classification. The dataset used for this assignment is the 20-newsgroup dataset which consists of 20,000 texts divided into 20 classes. This assignment was conducted using a machine with the following specifications:

Processor: Intel i7-9700K

Graphics Card: Nvidia RTX 2070

RAM: 32 GB

Answers to questions:

1. Dd
2. D
   1. The first step is to load the dataset and split them into training and test sets. Next is the configuration of stopwords, which are words that are either too common or too rare and carries no significant weight or meaning in providing context to classify the texts. A helper function is then created to remove punctuations from the text. This is done because punctuations such as full stop or an indent is not needed to classify a text. Additionally, numerical strings, words with only 1 or 2 characters, and empty strings are to be removed as well for the same reason. The step after that is creating a helper function to tokenize the text. Tokenizing means taking the sentences and breaking them down into a list of words. Removing any metadata such as \n is the next step as, again, they do not mean anything for the classification task.