

**Applied Analytics Assignment**

Diploma in Cybersecurity & Digital Forensics

Diploma in Data Science

Diploma in Information Technology

Year 2/3 (2025), Semester 3/5

**~~TEAM~~/INDIVIDUAL REPORT**

\*delete where applicable

(40% of AA Module)

**Deadline for Submission:**

**Presentation Slides: 10th August 2025 (Sunday),23:59hrs**

**Report & Code: 10th August 2025 (Sunday),23:59hrs**

|  |  |  |  |
| --- | --- | --- | --- |
| Tutorial Group | : | T01 | |
| Team Number | : |  | |
| Tutor | : | Dr Wang Siqi | |
| Members | : | Student No. | Student Name |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Penalty for late submission:**

10% of the marks will be deducted every day after the deadline.

**NO** submission will be accepted after 15th August 2025, 23:59 hrs.

1. Introduction

Text Classification is a foundational task in Natural Language Processing, with applications ranging from spam detection to sentiment analysis and content recommendation. For this assignment, the aim is to explore text classification techniques by building a model that can predict a movie’s genre based on its text description – focusing on applying text analysis techniques in Python to process, understand and model movie descriptions extracted from IMDb.

## Problem Understanding

The core objective is to classify movies into one of the five genres – action, comedy, documentary, drama or thriller – using the movie’s description. This is a multi-class classification problem where each movie belongs to one of the genres mentioned. The dataset consists of 5000 movie entries, evenly distributed across the five genres, with 1000 in each. Each entry consists of three main fields: Title, Genre and Description. As the focus is on the textual data, the “Description” field will be the primary input feature used for training my models. The text data to understand movie descriptions is often unstructured, noisy and may contain domain specific language, slang or idiomatic expressions. Additionally, genre boundaries can overlap – whereby some action movies may contain thriller elements, thereby adding further complexity to this classification.

Therefore, it is crucial to clean and preprocess the textual data so that it can be understood that it is suitable for to be used by the Machine Learning algorithms.

## Approaches

# Text Data Preprocessing

## Load and Cleanse Text Data

To start our Text Data Preprocessing, I first installed all the required Python packages using “pip”. Using “%” before the command allows us to run shell commands directly inside the notebook cell. These would be the packages that I require moving forward in the assignment.

A screen shot of a computer

AI-generated content may be incorrect.

Next, I import all required modules

A screen shot of a computer program

AI-generated content may be incorrect.