A picture containing text, font, screenshot, graphics

Description automatically generated

DEPARTMENT OF INFORMATION TECHNOLOGIES

IT 505 Natural Language Processing - Final Exam

Date: 30/05/2025

## **Instructions**

* You have access to standard NLP libraries, and their official documentation.
* Write clear code with brief explanations
* Partial credit will be given for incomplete but correct approaches

## **Question 1: Data Analysis and Preprocessing (25 points)**

You are provided with a movie review dataset ([movie\_reviews.csv](https://drive.google.com/file/d/1QqS-F-PwOnO8gHtNvgy7jWOHOqkLy43k/view?usp=sharing)) containing columns: text and label (positive/negative).

1. **Load and explore the dataset (10 points)**
   * Display basic statistics (number of samples, class distribution)
   * Show 2 examples from each class
   * Calculate the average text length for each class
2. **Basic preprocessing (15 points)**
   * Convert text to lowercase
   * Remove punctuation and numbers
   * Tokenize the text
   * Remove stop words
   * Apply preprocessing to create a new column, processed\_text

## **Question 2: VADER Sentiment Analysis (30 points)**

1. **Implement VADER analysis (20 points)**
   * Use VADER SentimentIntensityAnalyzer on the original text
   * Extract compound scores for all samples
     1. (Compound: It provides us the overall sentiment score and its value ranges from -1 to 1, -1 being the strong negative sentiment and 1 being the strong positive sentiment.)
   * Convert compound scores to binary labels (positive if compound > 0, negative otherwise)
   * Store results in a new column, vader\_prediction
2. **Evaluate VADER performance (10 points)**
   * Calculate accuracy, precision, and recall, against true labels
   * Create and display a confusion matrix

## **Question 3: Machine Learning Implementation (35 points)**

1. **Feature extraction (15 points)**
   * Create TF-IDF features from the processed text
   * Use a maximum of 1000 features
   * Split data into train/test (80/20) using random\_state=42
2. **Train and evaluate a classifier (20 points)**
   * Train a Naive Bayes classifier on the training set
   * Make predictions on the test set
   * Calculate and report accuracy, precision, and recall
   * Show the confusion matrix

## **Question 4: Comparison and Analysis (10 points)**

1. **Compare both approaches (5 points)**
   * Create a comparison table showing accuracy, precision, and recall for both methods
   * Identify which method performed better overall
2. **Critical analysis (5 points)**
   * Write 4-5 sentences discussing:
     + One advantage and one disadvantage of the VADER approach
     + One advantage and one disadvantage of the Machine Learning approach
     + Which method would you choose for a production system and why?

### **Pre-loaded Libraries:**

import pandas as pd

import numpy as np

import nltk

import re

import string

from nltk.corpus import stopwords

from nltk.sentiment import SentimentIntensityAnalyzer

from sklearn.feature\_extraction.text import TfidfVectorizer

from sklearn.naive\_bayes import MultinomialNB

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import accuracy\_score, classification\_report, confusion\_matrix

## **Important Notes**

* **NLTK Data:** Ensure vader\_lexicon and stopwords are downloaded
  + nltk.download('vader\_lexicon', quiet=True)
  + nltk.download('stopwords', quiet=True)
* **Code Comments:** Briefly explain your approach in every step
* **Partial Credit:** Even incomplete implementations can earn points if the approach is correct

**Good Luck!**