Sentiment Analysis – Take home Assignment

Handed out:	14th May 2023
Due Date:	24th May 2023, Midnight
Expected Deliverables:	One ZIP folder containing, • Report on findings in PDF format. • Well commented Python code in .ipynb format
Method of Submission:	Online via Moodle(see below)

Assignment Description

IMDb, a popular online movie database, is eager to gather feedback from its customers through movie reviews posted on their platform. They want to analyze the sentiment of these reviews and, more importantly, categorize any suggestions provided by customers to improve their movie offerings. Your task is to develop a solution that enables IMDb to achieve this goal efficiently.

You are required to produce Python code in a Jupyter Notebook (or Google Colaboratory) to do the following. And compile the findings into a report.

(a) Use the IMDB movie reviews dataset provided. Read the csv data file to a Pandas dataframe and take a sample of 10,000 reviews as your main dataset used for Sentiment Analysis. Use stratified sampling to ensure that the classes are balanced. Clean the data in appropriate ways. Print the number of reviews which are positive and the number that are negative in order to gauge the dimensions of the dataset.

(b) Create bag-of-words and TF-IDF representations of the reviews in the main-dataset above and use two relevant supervised learning algorithms to classify future reviews according to their sentiment. Print the confusion matrices of the four (04) resulting combinations for a held-out (test) dataset. (Split the data into training and testing sets, using 80% of the data for training and the rest for testing.)

(c) Suggest any strategies you may use to improve the performance of the above classifier (apart from using deep learning). Implement your suggestions as improvements to the above models and print the confusion matrix of the best representation and model you get.

(d) Write a report summarizing your findings and including the following sections:

   a. Introduction: Briefly explain the purpose of the assignment, the dataset used, and the goal of sentiment analysis.

b. Data Preparation: Explain how you cleaned and tokenized the data and created the bag-of-words and TF-IDF representations. Provide a brief explanation of the bag-of-words and TF-IDF techniques.
c. Classification Models: Evaluate the performance of your model using the test set. Compare and contrast the performance of the TF-IDF and bag-of-words methods. Provide a comprehensive analysis and comparison of the performance metrics achieved by the two models on the testing set.
d. Conclusions and Discussion: Suggest the best-performing model for conducting sentiment analysis and provide reasons for your choice. Reflect on your results and provide insights into the strengths and limitations of your sentiment analysis model. Discuss any issues with the data and models used (without implementing solutions) and suggest improvements to make the overall model more useful and enhance its performance.

Ensure the report is well-written, clear, and concise. Include any relevant visualizations, tables, or graphs to support the findings and the Python code should be well-commented and organized.

Submission

You need to formulate solutions for each parts (a) through (c) above, clearly explaining your python code specifying the outputs produced by the code for the dataset given in a Jupyter Notebook named *Solution_IDNumber.ipynb* based on the template given. As per the instructions (d) the prepared report should include your findings and answers to the above parts from (a)-(c) Your Jupyter Notebook and Report should be submitted to the Moodle as a compressed (.zip or .rar) file with the above naming(e.g., *Solution_IDNumber.zip*).