

Text Mining and Language Analytics

Content and skills covered by the assignment

- Understand advanced concepts of Natural Language Processing (NLP).
- Have a critical appreciation of the main strengths and weaknesses of a range of NLP methods and understand how to use them.
- Have a critical appreciation of how to prepare textual datasets for analysis.
- Understand how to manipulate potentially large datasets in an efficient manner.
- Be able to write computer programs for NLP in Python using industry-standard packages.
- Be able to select appropriate data structures for modelling various NLP scenarios.
- Be able to select the appropriate algorithms for a given NLP problem.
- Be able to prepare, train, evaluate and deploy machine learning models for NLP.
- Effective written communication.
- Planning, organising and time-management.
- Problem solving and analysis.

Requirements:

This assignment requires you to design, explain and justify your proposed solution for a Natural Language Processing (NLP) scenario. The solution should be implemented using the Python (3.x) programming language and also requires a written report to demonstrate how and why you designed the proposed solution, as well as a thorough performance evaluation of it.

Scenario:

In this imaginary scenario, you are a data scientist for a marketing company. Your company has asked you to create machine learning models that given a written review about a product, they can predict the product's rating by online users. To succeed in your assignment, you have to use the provided product reviews' dataset and create suitable machine learning models for predicting the user rating of a product review, as described below. The deliverables for your assignment consist of a Python implementation for training the proposed models and using them for predictions, as well as a written report that justifies your decisions and provides a performance evaluation of your proposed solutions.

Dataset:

The dataset is stored in the “dataset.csv” file and consists of a comma-separated file (csv) with product reviews from Amazon. Each review is annotated with a rating between 1 and 5. The dataset file is organised into two columns, as follows:

Column	Description
Score	Rating as one of the following scores: {1, 2, 3, 4, 5}
Text	The product review's text.

Python Implementation:

You are asked to use the provided dataset in order to develop and evaluate machine learning models for predicting the rating of a product review. Address this problem as a classification problem.

The required tasks are the following:

1. Prepare the dataset by applying any pre-processing or cleaning steps that you consider as necessary. Then, split the dataset into a training set containing 80% of the samples and a test set containing 20% of the samples. Follow an appropriate strategy for the split. You must use these training/test sets for all the models in this coursework. (10%)
2. Implement a Naïve Bayes model for predicting the rating of a product review. Train your model on the training set and test it on the test set. Use an appropriate text representation. (5%)
3. Implement a k-Nearest Neighbours model for predicting the rating of a product review. Train your model on the training set and test it on the test set. Use an appropriate text representation. You must select the best k by examining the performance of the model for $k \in \{1,3,5,7,9\}$, using an appropriate cross-validation approach. Create a plot for k vs. classification performance to justify your choice. (10%)
4. Implement a Convolutional Neural Network (CNN) model for predicting the rating of a product review. The model must have at least two convolutional layers. Train your model on the training set and test it on the test set. Use an appropriate text representation. (13%)
5. Implement a Recurrent Neural Network (RNN) or a Long Short-Term Memory (LSTM) model for predicting the rating of a product review. The model must have at least two RNN/LSTM layers. Train your model on the training set and test it on the test set. Use an appropriate text representation. (12%)
6. Compute the confusion matrix, accuracy, F1-score, precision and recall for each model. (10%)
7. Store the four trained models in files and implement a function “predict_product_rating(text, model)” that given a text string (“text”) and model filename (“model”), it will load the pre-trained model, and predict the product review rating of the input text. The function should be able to work without requiring to rerun all or part of your code. (10%)

Note: You are strongly advised to use the Pandas library for loading and manipulating the dataset.