Robert Gordon University

CMM706 – Text Analytics 2022	
Module leader	Dr. Ruvan Weerasinghe
Unit	Coursework
Weighting:	70% for Annotated Code 10% for Viva 20% for Comparative Study Report
Learning Outcomes Covered in this Assignment:	LO1 Critically appraise extraction and search models in information retrieval and Natural Language Processing in relation to big data case studies. LO2 Critically evaluate current research and advanced scholarship in IR and NLP, their role and alternative directions for big data projects. LO3 Combine methods from NLP, topic modelling and text mining tool−kits to develop new extraction processes for real−world tasks. LO4 Plan a comparative study to evaluate and interpret results from designing and developing information retrieval and extraction systems for big data.
Handed Out:	25th June 2022
Due Date	24th July 2022, midnight. Each individual will be scheduled a 15 minute time slot to demonstrate their solution on a date(s) to be agreed in class (potentially in the week starting 8th August 2022).
Expected deliverables	One compressed electronic file containing the reports and code specified below.
Method of Submission:	Online via Moodle (see below).
Type of Feedback and Due Date:	Written feedback and marks – within 10 working days after the conclusion of the vivas.

A Sri Lankan startup company wants to setup a sports information service for sports enthusiasts who want to keep up to date with their favourite sports and teams. In order to support their plan, they want to know how they can access every single relevant piece of sports news as soon as possible.

(a) Using the Twitter API, collect at least 10,000 sports news items in the English language from the past year.

(5 marks)

(b) Describe the dataset collected in terms of the number of news items, the total number of words, the number of unique words and the distribution of lengths in tokens of the new items.

(5 marks)

(c) Using a suitable mechanism identify possible duplicates (or near duplicates) and remove them from the dataset. Describe the resulting deduplicated dataset as in (b) above.

(5 marks)

(d) Check the nature of the dataset and take any action required to clean the dataset. Tokenize the news items and preprocess the data for feature extraction.

(5 marks)

(e) Perform feature extraction on the dataset by creating sparse and dense vector representations. Describe the final dataset(s) in terms of number of news items and number of features for each such feature extraction method. Randomly select 300 data points from this final dataset for testing.

(10 marks)

(f) Using machine learning[1], categorize the rest of the news items in the dataset into an appropriate number of groups using any appropriate technique and justify the categorization scheme.

(10 marks)

(g) Manually annotate the 300 randomly selected news items from the dataset into different sports. Check the accuracy of the model learned in (f) using this annotated sample as the ground truth.

(5 marks)

(h) Assume that the categorization you arrived at in (f) is close enough to the 'ground truth' and use the categories learnt as their labels. Take necessary action to remedy any imbalance in the data you may have. Use 3 different non-deep learning algorithms from *scikit-learn* that would help you classify this data, giving reasons for the choice of each such algorithm. Comment on the performance of each of the classifiers.

(10 marks)

[1] From the scikit-learn library.

(i) Use the same assumptions as (h) and explore 2 deep learning architectures[2] to try to improve the sports data classification results. Comment on the performance of the deep learning models.

(10 marks)

(j) How would you detect if your deep learning model has overfitted the training data and what could you do about it, if so?

(5 marks)

The Annotated Code

You need to formulate solutions for each of parts (a) through (j) above, clearly explaining your Python code and specifying the outputs produced by the code for the dataset given in a *Jupyter Notebook* named *Solution_IDNumber.ipynb* based on the template given[3]. For each such part, a descriptive summary with an interpretation should be given for the output obtained after each executable *cell*.

Viva

You need to explain your code and the output it produces using the Jupyter notebook for each part (a) through (j) to demonstrate your understanding.

Comparative Study Report

Many NLP tasks can benefit by the use of transfer learning. Describe what you understand by transfer learning and cite any references of performance improvement achieved in various NLP tasks using transfer learning. Your report should be no longer than **1000 words** and should list the references used at the end.

Use the popular BERT or ELMo architectures for transfer learning on the final dataset used in part (i) above and compare its performance with the classification and deep learning models. Your implementation should be included as part (k) in the above Jupyter Notebook.

The PDF version of the report should be named, *Report_IDNumber.pdf* where your IIT ID number should replace IDNumber.

Submission

Your Jupyter Notebook and comparative study report should be submitted to Campus Moodle *as a single compressed (.zip or .rar) file* with the name *Coursework_IDNumber.zip* (or *.rar*) with the IDNumber replaced by your IIT ID number.

[2] Using the keras/tensorflow library.

[3] The *IDNumber* part of the filename should be replaced with your *IIT ID* number.

Coursework Marking scheme

The Coursework will be marked based on the following marking criteria:

Question	Marks	Marks provided	Comments
(a)	05		
(b)	05		
(c)	05		
(d)	05		
(e)	10		
(f)	10		
(g)	05		
(h)	10		
(i)	10		
(j)	05		
Presentation	10		
Comparative Study Report	20		*15% for documentation/literature* *5% for implementation*
Total	100		