ROBERT GORDON
UNIVERSITY ABERDEEN

Assessment Brief

Academic Year	**2022/2023**
Semester	**3**
Module Number	**CMM706**
Module Title	**Text Analytics**
Assessment Method	**Coursework**
Deadline (time and date)	**23rd July Midnight**
Submission	**Assessment Dropbox in the Module Study Area in CampusMoodle.**
Word Limit **(see Assessment Word Limit Statement)**	**1500 words (report)**
Module Co-ordinator	**Ruvan Weerasinghe**

What knowledge and/or skills will I develop by undertaking the assessment?

Students will be able to explore different ways to collect online text data using APIs and/or web scraping, describe the data so collected, clean and prepare the dataset for processing (including identifying spurious records and missing data), determine appropriate ways of extracting features from the data (including the use of dense methods), systematically apply increasingly complex algorithms for classifying the data for

What knowledge and/or skills will I develop by undertaking the assessment?

building a predictive analytics system and identify overfitting and what can be done to mitigate the same. They would also be able to reflect on their learning, including how to use generative AI as a tool to aid their understanding of areas beyond the scope of the course module.

On successful completion of the assessment students will be able to achieve the following Learning Outcomes:

1. *Critically appraise extraction and search models in information retrieval and Natural Language Processing in relation to big data case studies.*
2. *Critically evaluate current research and advanced scholarship in IR and NLP, their role and alternative directions for big data projects.*
3. *Combine methods from NLP, topic modelling and text mining tool–kits to develop new extraction processes for real–world tasks.*
4. *Plan a comparative study to evaluate and interpret results from designing and developing information retrieval and extraction systems for big data.*

Please also refer to the Module Descriptor, available from the module Moodle study area.

What is expected of me in this assessment?

Task(s) - content

The overall goal of the task is to classify news content and explore the possibility of predicting the news source by the content of an article. The rational use of generative AI (e.g. ChatGPT) is encouraged.

What is expected of me in this assessment?

(a) You are required to identify 10 popular Sri Lankan news sources who have an active presence on Twitter. List these 10 twitter handles together with their follower and following counts, and the number of tweets each made during the past 12 months.

(b) Extract all articles indexed by the twitter handles of these news sources and state the dimensions of the resulting article collection (NOT tweet collection) of all news agencies in terms of the total tokens and the unique tokens. State also the total tokens and unique tokens of each of the news agencies separately. In preparation for building a classifier of such news articles, address any potential imbalance in the dataset using at least two (02) different methods.

(c) Use a sparse and a dense vector representation for extracting features for training a classifier for this dataset. Interpret the dimensions of the sparse vector and justify the dimensions of the dense vector used.

(d) Train classifiers with the two (02) representations above using three (03) non-deep learning algorithms, stating your reasons for selecting each algorithm. Compare and contrast the performance of each of the classifiers.

(e) Train also three (03) deep learning classifiers with distinct architectures using two (02) embedding techniques and one (01) contextual embedding technique, justifying the architectures you employ. Compare the performance of each of the models and interpret the results.

Task(s) - format

You need to formulate solutions for each of parts (a) through (e) above, clearly explaining your Python code and specifying the outputs produced by the code for the dataset used in a Jupyter Notebook named Solution_IDNumber.ipynb based on the template given (the IDNumber part of the filename should be

What is expected of me in this assessment?

replaced with your IIT ID number). For each such part, a descriptive summary with an interpretation should be given for the output obtained after each executable cell. The notebook should be compressed as a .zip file. You also need to submit a PDF version of the notebook using a tool described at https://saturncloud.io/blog/how-to-export-jupyter-notebook-as-pdf. The name of the file should be the same as the notebook, except that the extension will be .pdf.

In addition, you should write a comprehensive reflection of not more than 1500 words on how you used generative AI in grasping the concepts of deep learning and transfer learning in particular in carrying out your coursework. You should NOT reproduce the detailed responses of this process, but a step-by-step way that you used it for enhancing your understanding. The PDF version of this report should be named, Report_IDNumber.pdf where your IIT ID number should replace IDNumber.

Your Jupyter Notebook files (the zipped notebook and the converted pdf) and the comparative study report should be submitted as three separate files to Campus Moodle. Note that the PDF files should NOT be compressed. Submissions which do NOT adhere to this formatting and naming convention may be marked later causing delays in the release of your grades.

How will I be graded?	
A grade will be provided for each criterion on the feedback grid which is specific to the assessment.	
A	At least 50% of the feedback grid to be at Grade A, at least 75% of the feedback grid to be at Grade B or better, and normally 100% of the feedback grid to be at Grade C or better.
B	At least 50% of the feedback grid to be at Grade B or better, at least 75% of the feedback grid to be at Grade C or better, and normally 100% of the feedback grid to be at Grade D or better.
C	At least 50% of the feedback grid to be at Grade C or better, and at least 75% of the feedback grid to be at Grade D or better.
D	At least 50% of the feedback grid to be at Grade D or better, and at least 75% of the feedback grid to be at Grade E or better.
E	At least 50% of the feedback grid to be at Grade E or better.
F	Failing to achieve at least 50% of the feedback grid to be at Grade E or better.
NS	Non-submission.

Feedback grid

GRADE	A	B	C	D	E	F
DEFINITION / CRITERIA (WEIGHTING)	**EXCELLENT** Outstanding Performance	**COMMENDABLE/VERY GOOD** Meritorious Performance	**GOOD** Highly Competent Performance	**SATISFACTORY** Competent Performance	**BORDERLINE FAIL**	**UNSATISFACTORY** Fail
CRITERION 1 **(20 %)** **Grade:** □	Student has identified 10 top news sources, extracted their news items over the past 12 months, described them visually and addressed class imbalance.	Student has identified 10 top news sources, extracted their news items over the past 12 months, described them and tried to address class imbalance.	Student has identified 10 top news sources, extracted a sufficient number of news items, described them and tried to address class imbalance.	Student has identified good news sources, extracted a sufficient number of news items, tried to describe them and made some effort to address class imbalance.	Student has identified Sri Lankan news sources, extracted some news items, failed to describe them adequately and not addressed class imbalance adequately.	Students hasn't identified relevant news sources, extracted tweets instead of news articles, inadequately described the data or not addressed class imbalance.
CRITERION 2 **(10 %)** **Grade:** □	Student has clearly identified and carried out feature extraction, justifying each and clearly explaining the resulting shape of the dataset.	Student has identified and carried out feature extraction and clearly explained the resulting shape of the dataset.	Student has identified and carried out feature extraction and explained the resulting shape of the dataset.	Student has identified and carried out feature extraction but only stated the resulting shape of the dataset.	Student has identified and carried out some feature extraction, but failed to state the final shape of the dataset.	Student has carried out basic feature extraction and appears not to have grasped the purpose or the significance of this process.
CRITERION 3 **(15 %)** **Grade:** □	Student has used and justified the choice of algorithm used clearly, interpreted the results and compared them including testing for overfitting.	Student has used and tried to justify the choice of algorithm used, interpreted the results and compared them including testing for overfitting.	Student has used and tried to justify the choice of algorithm used, compared the results including testing for overfitting.	Student has used and tried to justify the choice of algorithm used, compared the results but not tested for overfitting.	Student has used three algorithms and compared the results but not tested for overfitting.	Student has used algorithms done in class and stated the results but not tested for overfitting.
CRITERION 4 **(25 %)** **Grade:** □	Student has justified the use of significantly different deep learning architectures, using different embedding methods and demonstrating a good understanding of the process including in dealing	Student has justified the use of different deep learning architectures, using different embedding methods and demonstrating some understanding of the process including in dealing with the problem of overfitting.	Student has used different deep learning architectures, using different embedding methods and demonstrated some understanding of the process including in dealing with the problem of overfitting.	Student has used different deep learning architectures, using different embedding methods and dealt with the problem of overfitting.	Student has used some deep learning algorithms, embedding methods but has little understanding of how to deal with the problem of overfitting.	Student has used some deep learning algorithms and embedding methods fairly mechanically and has little to no understanding of how to deal with the problem of overfitting.

GRADE	A	B	C	D	E	F
DEFINITION / CRITERIA (WEIGHTING)	**EXCELLENT** Outstanding Performance	**COMMENDABLE/VERY GOOD** Meritorious Performance	**GOOD** Highly Competent Performance	**SATISFACTORY** Competent Performance	**BORDERLINE FAIL**	**UNSATISFACTORY** Fail
	with the problem of overfitting.					
CRITERION 5 **(10 %)** **Grade:**	Student confidently answers all questions asked to verify understanding of data collection, cleaning, dealing with imbalance, feature extraction, algorithm selection, model building, model diagnostics and interpretation.	Student confidently answers most questions asked to verify understanding of data collection, cleaning, dealing with imbalance, feature extraction, algorithm selection, model building, model diagnostics and interpretation.	Student is able to answer most questions asked to verify understanding of data collection, cleaning, dealing with imbalance, feature extraction, algorithm selection, model building, model diagnostics and interpretation.	Student answers many of the core questions asked to verify understanding of data collection, cleaning, dealing with imbalance, feature extraction, algorithm selection, model building, model diagnostics and interpretation.	Student provides partial answers to the core questions asked to verify understanding of data collection, cleaning, dealing with imbalance, feature extraction, algorithm selection, model building, model diagnostics and interpretation.	Student is unable to answer many core questions asked to verify understanding of data collection, cleaning, dealing with imbalance, feature extraction, algorithm selection, model building, model diagnostics and interpretation.
CRITERION 6 **(20 %)** **Grade:**	The report provides a clear and concise reflection of the student’s problem-solving process including aspects of deep learning and transfer learning yet to be covered by the deadline for submission.	The report provides a good reflection of the student’s problem-solving process including aspects of deep learning and transfer learning yet to be covered by the deadline for submission.	The report provides a good reflection of the student’s problem-solving process including some aspects of deep learning and transfer learning yet to be covered by the deadline for submission.	The report provides an adequate reflection of the student’s problem-solving process including some aspects of deep learning and transfer learning yet to be covered by the deadline for submission.	The report provides a sketchy reflection of the student’s problem-solving process and only a shallow understanding of deep learning and transfer learning.	The report fails to provide a reflection of the student’s problem-solving process and no evidence of understanding of deep learning and transfer learning.

Coursework received late, without valid reason, will be regarded as a non-submission (NS) and one of your assessment opportunities will be lost.

What else is important to my assessment?

What is plagiarism?

"Plagiarism is the practice of presenting the thoughts, writings or other output of another or others as original, without acknowledgement of their source(s) at the point of their use in the student's work. All materials including text, data, diagrams or other illustrations used to support a piece of work, whether from a printed publication or from electronic media, should be appropriately identified and referenced and should not normally be copied directly unless as an acknowledged quotation. Text, opinions or ideas translated into the words of the individual student should in all cases acknowledge the original source" (RGU 2022).

What is collusion?

"Collusion is defined as two or more people working together with the intention of deceiving another. Within the academic environment this can occur when students work with others on an assignment, or part of an assignment, that is intended to be completed separately" (RGU 2022).

For further information please see Academic Integrity.

What is the Assessment Word Limit Statement?

It is important that you adhere to the Word Limit specified above. The Assessment Word Limit Statement lists what is included and excluded from the word count, along with the penalty for exceeding the upper limit.

What if I'm unable to submit?

- The University operates a Fit to Sit Policy which means that if you undertake an assessment then you are declaring yourself well enough to do so.
- If you require an extension, you should complete and submit a Coursework Extension Form. This form is available on the RGU Student and Applicant Forms page.
- Further support is available from your Course Leader.

What else is important to my assessment?

What additional support is available?

- RGU Study Skills provide advice and guidance on academic writing, study skills, maths and statistics and basic IT.
- RGU Library guidance on referencing and citing.
- The Inclusion Centre: Disability & Dyslexia.
- Your Module Coordinator, Course Leader and designated Personal Tutor can also provide support.

What are the University rules on assessment?

The University Regulation ‘A4: Assessment and Recommendations of Assessment Boards‘ sets out important information about assessment and how it is conducted across the University.

Date created: Aug 2022
Version: 2