School of Computing, Napier University

Assessment Brief

1. Module number	**SET11121 / SET11521**
2. Module title	**Data Wrangling**
3. Module leader	*Dimitra Gkatzia*
4. Tutor with responsibility for this Assessment	Dimitra Gkatzia *(D.Gkatzia@napier.ac.uk)*
5. Assessment	Coursework
6. Weighting	100% of module assessment
7. Size and/or time limits for assessment	1700 words plus figures or tables with results and developed code for all questions.
8. Deadline of submission Your attention is drawn to the penalties for late submission	Part A: 08/03/18 at 1500 UK time Part B: 12/04/18 at 1600 UK time
9. Arrangements for submission	Your Coursework must be submitted via Moodle. **Further submission instructions are included in the attached specification, and on Moodle**
10. Assessment Regulations	All assessments are subject to the University Regulations**.**
11. The requirements for the assessment	See Attached
12. Special instructions	See Attached
13. Return of work	Feedback and marks will be provided **within three weeks** of submission.
14. Assessment criteria	Your coursework will be marked using the marking sheet attached as Appendix A. This specifies the criteria that will be used to mark your work. Further discussion of criteria is also included in the coursework specification attached.

Assessment Brief

The assignment aims to cover the learning outcomes specified for the module:

LO1: Critically evaluate the tools and techniques of the data storage, interfacing, aggregation and processing
LO2: Select and apply a range of specialised data types, tools and techniques for data storage, interfacing, aggregation and processing
LO3: Employ specialised techniques for dealing with complex data sets
LO4: Design, develop and critically evaluate data driven applications in Python

The goal of this assignment is to develop a prediction model for Abusive Language Detection.

Data
For this assignment you will require to use the datasets provided on moodle.

Part A - 30%. Deadline: Friday 8 March at 3pm (UK time).

Deliverable 1: You will need to perform a literature review on recent approaches to abusive language detection. You will need to pick 3 new approaches published after 2016. For each approach, you will need to describe the dataset they used, the approach (including the feature selection), a brief description of their result as well as your critical review (are there any issues with the study, how would you improve it? etc.). Your report **must** include an introduction (intro to the topic and described methods), background (description of methods as described previously), a discussion (critical analysis), and a summary of your results from Deliverable 2.

Deliverable 2: Using the **provided** datasets, you will need to:

- Load (in Python) and store the training dataset using one of the approaches you learnt. In the comments explain why you chose to store the data in a particular way.
- Perform some analysis, e.g. find most frequent/infrequent words, number of unique words,

Your references should come from international venues (such as conferences and journals). You can look for papers at Google Scholar or at the university library (online).

Your report must adhere to citation guidelines - any citation style is acceptable. An example guide can be found here:
https://drhazelhall.files.wordpress.com/2013/01/2005_hall_referencing.pdf

You will submit:

Part A consists of two deliverables:

Deliverable 1: One **.pdf file** of 1200 words. The document should include your name, matriculation number and contact details, as well as tables and a short description of your text analysis.

Deliverable 2: Your code with appropriate comments.

Everything must be submitted on moodle only!

Marking: You will be marked on the content (10%), the structure of the report (5%), the criticality (10%) and the quality of code (5%). See the end of the document for a detailed description of the marking scheme.

Part B - 70%. Deadline: Friday 12 April at 3pm (UK time).

For the second part of the assignment you will need **to develop and evaluate abusive language detection models for the given datasets.** You should choose two ML models: one of the ML approaches you were taught in class and one you identified from the literature. You should produce two models and an evaluation metric (metric taken from literature - you need to justify which metric you chose and why). The goal of this exercise is not to produce a state-of-the-art sentiment analysis model. If your chosen model performs poorly by your selected metric, do not worry—this is not what we are testing. Which model you use, and how you evaluate, is up to you. The choice of model is not important (although we will assume that when you choose a model, you understand what it is and how it works) as well as that the evaluation metric is appropriate. Your solution should be sensible - you should be able to explain why it tests something of impact to the problem.

Tips and Clarifications

We are not looking for models that performs well: we are looking to see that you can build sensible models, i.e. choose meaningful features and perform a sensible evaluation. If you are struggling to make something work with the volume of data present, you can subsample (for instance, randomly pick a proportion of the dataset). You must use Python and its libraries to tackle this task. You are strongly encouraged to make use of third-party libraries for model building and evaluation, rather than writing your own, unless you specifically need to do something with no library support.

You will submit:

1. The code of your solution, and a 500 words **.pdf document** explaining the data pre-processing, model features and evaluation as well as a discussion of your results and suggestions for improvement. If you do any pre-processing to the data, please also include the script you use to do this (or a list of the commands run).

Marking:

40% for method / model, 15% for evaluation, 15% for report and reflection. See Appendix A for more explanations.

Appendix A: Marking Scheme

	No Submission	Very poor	Inadequate	Adequate	Good	Very good	Excellent	Outstanding
A1 Content **10%**	No work submitted	Literature not described adequately, i.e described only the topic or the data, or sources are not relevant	Literature not described adequately, leaving most work unexplained	Literature described partially: half of its elements covered	Literature described partially	Literature described almost fully	Literature fully described, covering everything	Literature fully described and additional investigation was performed
A2 Structure **5%**	No work submitted	Report does not follow the guidelines or word limit	The structure of the report requires more work	The structure of the report is ok, but some part is missing	The structure of the report is overall good but there is room for improvement	The structure of the report is very good, naming of titles could improve	The structure of the report is excellent	The structure of the report is outstanding and professional
A3 Criticality **10%**	No work submitted	The lit has not been criticised	The lit review has not been criticised adequately, e.g. no mentioning of specific drawbacks	Not all sources has been criticised.	The lit review has been criticised but not thoroughly enough	The lit review has been criticised thoroughly and good insights has been provided	The lit review has been criticised thoroughly and valuable insights has been provided	The lit has been criticised thoroughly with excellent suggestions for improvement
A4 Code and explanation **5%**	No work submitted	Code with bugs	Code with bugs but good explanations or questions answered partly	Code without bugs but inadequate explanation	Code without bugs and good but not thorough explanation	Code without bugs and explanations almost complete	Excellent code and thorough explanations	Outstanding code and thorough and thoughtful explanations.

	No Submission	Very poor	Inadequate	Adequate	Good	Very good	Excellent	Outstanding
B1 Methods/ Models **40%**	No work submitted	Code with bugs and algorithm / model not well described	Code with bugs but algorithm / model well described	Code with a minor bug but algorithm /model not well described and justified	Code with a minor bug but algorithm /model well described and justified	Code without bugs but algorithm / model not described or justified	Code without bugs but algorithm / model not described and justified in great detail	Code without bugs and algorithm / model described and justified in detail
B2 Evaluation **15%**	No work submitted	Not appropriate evaluation metric chosen	Neither the evaluation setup nor the results are described appropriately	Evaluation setup is not justified but almost correctly executed and results are mentioned	Evaluation setup is not justified but correctly executed and results are mentioned	Evaluation setup is somewhat justified and results are somewhat mentioned and discussed	Evaluation setup is somewhat justified and results fully described and discussed	Evaluation setup is justified and results fully described and discussed
B3 Reflection **15%**	No work submitted	Reflection and future work suggestions did not make sense	Not adequate reflection provided neither suggestions for future work	Either only reflection or suggestions for future work submitted	Average reflection and suggestions for future work	Good reflection and suggestions for future work	Very good reflection and suggestions for future work	Excellent reflection and suggestions for future work

Late submission policy

Coursework submitted after the agreed deadline will be marked at a maximum of 40% (undergraduate) or P1 (postgraduate). Coursework submitted over five working days after the agreed deadline will be given 0% (although formative feedback will be offered where requested).

Extensions

If you require an extension, please contact the module leader **before** the deadline. Extensions are only provided for exceptional circumstances and evidence may be required. See the Fit to Sit regulations for more details.

Plagiarism

Plagiarised work will be dealt with according to the university's guidelines: http://www2.napier.ac.uk/ed/plagiarism/