COMP1804 COURSEWORK SPECIFICATIONS

Comp1804 – Applied Machine Learning

2023-24

Module Leader: Dr Stef Garasto.
Deadline Date: 26/03/2024, 23:30 UK time.
Contribution: 100% of module.
Pass mark: 50%+.
Marking: Feedback and grades are normally made available within 15 working days (usually 21 calendar days, unless there are bank holidays) of the coursework deadline. Your coursework may or may not be marked anonymously, depending on requirements.

Learning Outcomes:

1. Rationalise appropriate scenarios for Machine Learning applications and evaluate the choice of machine learning methods for given application requirements.
2. Demonstrate competency in using appropriate libraries/toolkits to solve given real-world Machine Learning problems and develop and evaluate suitable application.
3. Understand and apply the relevant input data preparation and processing required for the Machine Learning models used, and quantitatively evaluate and qualitatively interpret the learning outcome.
4. Recognise and critically address the ethical, legal, social and professional issues that can arise when applying Machine Learning technologies.

Plagiarism

Plagiarism is presenting somebody else's work as your own. It includes: copying information directly from the Web or books without referencing the material; submitting joint coursework as an individual effort; copying another student's coursework; stealing coursework from another student and submitting it as yourown work. Suspected plagiarism will be investigated and if found to have occurred will be dealt with according to the procedures set down by theUniversity. Please see your student handbook for further details of what is / isn'tplagiarism.

1. **All material copied or amended from any source (e.g. internet, books) must be referenced correctly according to the reference style you are using.**
2. **Your work will be submitted for plagiarism checking. Any attempt to bypass our plagiarism detection systems will be treated as a severe Assessment Offence.**

Generative AI tools

We recommend giving lots of thought to whether AI technologies are suitable for use with this assignment. We tried completing the assignment using them and found the results mostly generic or wrong. You might have better luck than us. Using AI is not strictly prohibited if it is for brainstorming or getting past the “blank page” block, but we recommend always critically reflecting on the output, since it is not guaranteed to be factual or correct. If you do use generative AI, you need to reference it and put any text **and code** within quotation marks (otherwise it’s an academic offence).

Coursework Submission Requirements

An electronic copy of your work for this coursework must be fully uploaded on the **Deadline Date of 26th March 2024** using the link on the coursework Moodle page for COMP1804. For this coursework you must submit 3 separate files:

- **A single pdf file named 'report_comp1804.pdf'** which will be the written report. The word count of this written report should be 2500±10%. The word count excludes references, if you have any. **Do not include your name** in the report.
- **A single .zip file** containing a python file (either in .ipynb notebook format or in .py script format). The file should be named 'code_comp1804.ipynb' (or code_comp1804.py). If submitting a .ipynb file, please save it after executing **all** the cells, so that it shows the output. Your code should show all and only all the relevant implementation steps needed to reproduce the content of your report.
- **A single excel file (or equivalent)** with the original dataset and your new labels added to the column "text_clarity". These labels should be one of two options ("clear_enough" and "not_clear_enough") and show the results of your labelling process.

Please note:

- Any text in the document must not be an image (i.e. must not be scanned) and would normally be generated from other documents (from the latex template in this case).
- There are limits on the file size (see the relevant course Moodle page).
- Make sure that any files you upload are virus-free and not protected by password or corrupted otherwise they will be treated as null submissions.
- You must NOT submit a paper copy of this coursework.
- Lecturers and tutors can *only* accept coursework submitted via the appropriate Moodle link. No other way of submitting the coursework will be considered.

The University website has details of the current Coursework Regulations, including details of penalties for late submission, procedures for Extenuating Circumstances, and penalties for Assessment Offences. See http://www2.gre.ac.uk/current-students/regs

Detailed Coursework Specification

The task is to implement ML solutions and produce a written report **individually** on the dataset provided. Please see below for further details.

You need to use the dataset downloaded from the link listed in this coursework specifications. This is because this dataset has been modified to be better tailored to this specific module. **The relevant data file is the .csv file**. The json file has been added to better comply with licensing purposes, but it is not relevant to solve the assigned task.

Please **do *not* use data downloaded from anywhere else**. Any coursework submitted with a different dataset will be given 0 marks.

Data and tasks specifications: predicting topics and clarity from textual data.

You are consulting for a non-profit organization, called NotTheRealWikipedia, on whether machine learning can help them automatically analyse new content added to their site.

You are provided with a dataset with sample texts, their associated topic and other potentially useful information. The client is focusing on paragraph-sized text (still of varying length), because new articles and edits created on their website can be short.

From this dataset, they want you to do **two tasks**.

1) First, the client wants to know **whether it is possible to use machine learning to identify the topic of a paragraph of text, for some specific topics of interest**. Specifically, they want to know whether a given piece of text is about "artificial intelligence", "movies about artificial intelligence", "programming", "philosophy" or "biographies" (5 classes in total). They also provide a feature indicating whether each paragraph contains references to a person, and organisation and/or a product, which they think might provide further relevant information.
    a. For the avoidance of doubt, **the topic to predict is in the column titled "category". The input features to use are: "paragraph" and "has_entity"**.
    b. The client will consider the results successful if the model is better than a trivial baseline, if it does not overfit to the training dataset **and** if, for each class, no more than 10% of the paragraphs get misclassified into an unrelated class ("artificial intelligence" text being misclassified as "programming" is the one error they are willing to overlook).
    c. They also want to know which other scalar performance metric would be most informative to understand how well the algorithm is performing in general.
2) Then, the client want to explore whether it would be possible to automatically detect if a given paragraph is written clearly enough. They are planning to use the results to automatically reject edits and additions to the website's knowledge base if they are not clear enough. However, they do not have any labels for this task outside of the first few rows of the dataset. So, they want you to **build a prototype by first labelling a subset of the data** (they give an optional suggestion of 100 data points), **and then building a machine learning algorithm to predict these labels from the text and any other feature, as relevant**. Specifically, they want you to use two labels: "clear_enough" and "not_clear_enough", to denote the level of text clarity. You should add your labels in the column called "text_clarity". This column will then be your output feature.
    a. They have heard there is now lots of interest about responsible use of machine learning. So, they would like you to review the ethical implications and risks of using an algorithm to automatically reject users' work (for example, in terms of potential bias). Depending on the risks identified, they are open to consider applying the algorithm in a different way and are looking for suitable suggestions.
    b. The client will develop the prototype further if the algorithm produces results that do not overfit on the training data and are better than simply guessing the majority class all the time. They also want to know your top suggestion for improvement.
    c. The client is particularly interested in a prototype that includes some more advanced techniques (the main suggestions given are being able to make use of

both labelled and unlabelled data points or using pre-trained word embeddings).

They want you to write the results of your analysis and implementation in a report. More details about what to include in the report are provided below.

The dataset can be downloaded from this Moodle link. The dataset has been adapted to the requirements of this module; the original textual content is licensed under the GNU Free Documentation License (GFDL) and the Creative Commons Attribution-Share-Alike 3.0 License by Wikipedia.[1]

The table below gives some background information about the dataset features. For task 2, you will need to carefully select which extra features, in addition to the "paragraph" column, can be helpful to use as input.

FEATURE NAME	*BRIEF DESCRIPTION*
par_id	Unique identifier for each paragraph to classify.
paragraph	Text to classify.
has_entity	Whether the text contains a reference to a product (yes/no), an organisation (yes/no), or a person (yes/no).
lexicon_count	The number of words in the text.
difficult_words	The number of difficult words[2] in the text.
last_editor_gender	The gender of the latest person to edit the text.
category	The category into which the text should be classified.
text_clarity	The clarity level of the text. Very few data points are labelled at first.

Report

You should submit a report detailing the work you did to solve the machine learning (ML) tasks above, including any experiment performed to optimise the models. Report templates (Latex and Word) can be downloaded from Moodle.

The report should contain the following sections:

0. Executive summary.

Briefly summarize (**no more than 100 words**) for each task: how much of it you solved, the algorithm(s) used, to what extent the results meet the client's definition of success and why.

1. Data exploration and assessment.

Describe the exploratory data analysis performed and the dataset's main characteristics.

[1] Wikimedia Foundation. *Wikimedia Downloads*. https://dumps.wikimedia.org.
Some text may be available only under the Creative Commons license; text written by some authors may be released under additional licenses or into the public domain. For details, see Wikimedia Foundation's Terms of Use and Copyright licensing information. The coursework dataset is released under CC BY-SA 4.0 Deed.

[2] Difficult words are words with at least 3 syllables that are not in the Dale-Chall list of 3000 common words.

Also comment on any issues found in the dataset. For example, you may look at missing data.

For each figure, table or any other piece of information included here, it should be clear why it is relevant for the machine learning task(s) and what the implications are.

2. Data splitting and cleaning.

Describe in detail the steps performed for data splitting (training/validation/test) and cleaning (for example, imputation of missing values). Provide some justifications, based on theory and/or experiments, for your design choices. At the end of this section, you can include updated information on the dataset characteristics, if necessary (for example, if the label distribution has changed).

3. Data encoding.

Describe in detail the steps performed to encode each feature (for example, and where appropriate, text encoding, normalization/standardization, feature encoding, over/under-sampling). Provide some justifications, based on theory and/or experiments, for your design choices. Some of these justifications may be given in section 4, depending on your setup.

4. Task 1: topic classification.

4a. Model building.

Describe your solution to the topic classification task using ML techniques.

You should describe the final model hyper-parameters in details, ideally in a table, and give a brief explanation of why you chose this specific algorithm. You may include here hyper-parameters from section 3, if relevant.

Describe the experiments you did to optimise your model (specifically hyper-parameters optimisation; comparisons with one other model can be included if relevant). These experiments should be rigorous and follow best practice. Provide justifications, based on theory and/or experiments, for your design choices, including the choice of which hyper-parameters to test.

4b. Model evaluation.

Evaluate the model performance using a confusion matrix, a classification report and other relevant metrics, based on the characteristics of the dataset and the client's specifications. You should include a comparison with one "trivial" baseline (for example, random guess or majority class). You should also comment on how the metrics used are suitable for addressing the client's requirements.

Results should be presented in well-formatted figures and tables.

4c. Task 1 Conclusions.

You should provide two short bullet points with the following:

- State if the model is successful according to the client's definition of success (refer to point 1b from the task specifications).
- State which other scalar performance metric (one metric only!) you would recommend the client uses to keep track of the algorithm's performance.

5. Task 2: text clarity classification prototype.

5a. Ethical discussion.

Review and discuss the ethical implications and risks of using an algorithm to automatically reject users' work based on predicted text clarity. The discussion could include issues of bias, issues with the data, issues with performance, and others. Depending on the risks identified, make suitable suggestions on if and how to apply the algorithm to minimise ethical risks. You may use the Data Hazard Labels to help think through potential risks and also consider which communities and people may be affected by this algorithm.

5b. Data labelling.

You should describe your labelling process by giving details about your labelling criteria: the goal is to allow someone else to be able to label more data in a way that is consistent with yours. You should provide your own evaluation of how certain/uncertain you are about the final labels.

Also provide the final label statistics (how many data points were labelled, and how many ended up in each category), plus two examples, one from each category.

5c. Model building and evaluation.

Describe your solution to the text clarity classification task using ML techniques.

You should describe the final model hyper-parameters in details, ideally in a table.

Describe the advanced techniques used for this task and any experiments you did to optimise your model, if any. Note that, since you are required to build a **prototype**, the experiments do not need to be as thorough as in task 1, but, if performed, still need to be rigorous and follow best practice.

Provide justifications, based on theory and/or experiments, for your design choices, and give a brief explanation of why you chose this specific algorithm/technique.

Evaluate the model performance using a confusion matrix, a classification report and any other relevant metrics, based on the characteristics of the dataset and the client's specifications. You should include a comparison with the majority class baseline. You should also comment on how the evaluation criteria used are suitable for addressing the client's requirements.

Results should be presented in well-formatted figures and tables.

5d. Task 2 Conclusions.

You should provide three short bullet points with the following:

- State if the model is successful according to the client's definition of success (refer to point 2b from the task specifications).
- State which other scalar performance metric (one metric only!) you would recommend the client uses to keep track of the algorithm's performance.
- Your top suggestion for improvements and why.

6. Self-reflection

Here, you should identify a previous section that you think could have been improved,

being as specific as possible in commenting on what exactly could have been improved in that section. You should then comment on how you would improve the section and what you would need to be able to do so (for example: more time? better understanding of a specific concept?). This section should be **maximum 50 words**.

7. References

If needed, cite references and sources used at the end. These can be academic papers, blogs, code repositories, and more. Remember to give credit if someone else's work has helped you complete your coursework.

Reproducibility and report presentation

The report should be written in a professional style, in English, with clear language and a logical flow (report presentation). Double check your spelling, punctuation and grammar.

In principle, the reader should be able to recreate your final machine learning solution based on the report alone, without needing to look at the code (reproducibility). To achieve this, all the necessary implementation details must be included in the report.

The report must include only and all the sections detailed above. Marks may be removed if different sections are used. You are allowed to add sub-sections. To be safe, you are highly recommended to use one of the templates provided. You can download the coursework templates (Latex and Word) from Moodle.

Make sure that all figures have legible text and an appropriate label – use the same label to reference the figure in the text.

Do not include any screenshots of your code.

Code

The code must be developed in python and submitted as part of the coursework. It should show the implementation leading to all the results described in the coursework and should be well documented.

If needed, tutors should be able to run the code without errors. Tutors may review your code to check that the implementation is consistent with what is written in the report. Inconsistencies between the report and the code may be penalized. However, the code itself is not marked. Only what is described in your report counts towards your final mark.

Deliverables

For the submission to be admissible, all coursework submission requirements as specified in the Coursework Submission Requirements section above should be uploaded by the Deadline Date using the link on the coursework Moodle page for COMP 1804. If any of the required files is missing, or uploaded in a wrong format (for example, if the report is uploaded as a zip file), marks may be deducted.

Assessment Criteria

Marks breakdown

The marks breakdown into detailed assessment criteria is given below. Generally, both the solution implemented and the quality of explanation in the report will be taken into account.

- Data preparation (30 marks in total), split into:
  - Data exploration and assessment (10 marks).
  - Data splitting and cleaning (10 marks).
  - Data encoding (10 marks).
- Topic classification model (25 marks in total), split into:
  - Topic classification: model building (15 marks).
  - Topic classification: model evaluation (10 marks).
- Text clarity classification (32 marks in total), split into:
  - Text clarity classification: ethical discussion (7 marks).
  - Text clarity classification: data labelling (5 marks).
  - Text clarity classification: model building and evaluation (20 marks).
- Self-reflection (3 marks).
- Writing clarity and report presentation (10 marks).

Please see the table at the end for further details on how marks will be allocated.

Grading Criteria

A mark in the range of 80 to 100 will typically require the following:

- An outstanding and fully working implementation, showing a complete machine learning system for one task and at least an almost complete one for the other. The evaluation is comprehensive and in-line with the dataset's characteristics and the client's specifications. System design is rigorous. The solutions implemented demonstrate work that has gone beyond the material taught in the class.
- An outstanding report, clear and with all necessary details. It details thoughtful and comprehensive design choices and demonstrates a thorough understanding of machine learning applications and their ethical implications.

A mark in the range of 70 to 79 will typically require the following:

- A great and fully working implementation, showing a complete machine learning system for one task and a system completed to a good degree for the other. The evaluation is comprehensive and in-line with the dataset's characteristics and the client's specifications. System design is rigorous.
- A great report, clear and with all necessary details. It details thoughtful design choices and demonstrates a great understanding of machine learning applications and their ethical implications.

A mark in the range 60 to 69 will typically require the following:

- A good implementation, showing a complete or almost complete ML system for one

task and some good attempts at the other. The evaluation is appropriate and considers the datasets characteristics and the client's specifications to at least some extent. System design is mostly rigorous.
- A good report, sufficiently clear and with most necessary details. It includes some thoughtfulness in the design choices and demonstrates a good understanding of machine learning applications and their ethical implications.

A mark in the range 50 to 59 will typically require the following:

- An implementation showing a minimal but mostly working and sufficiently rigorous machine learning implementation for one task, with at least minimal engagement with the other. The evaluation is appropriate.
- An adequate report, sufficiently clear and with enough details. It shows some understanding of machine learning application and their ethical implications.

A mark in the range 0 to 49 will typically require the following:

- A system that fails to implement a working machine learning system for each of the tasks, and/or contains fundamental omissions or mistakes. The evaluation is not appropriate.
- An unsatisfactory and unclear report, showing little understanding of machine learning applications and their ethical implications.

Notes

Generally, marks will be given for:

- Features implemented. That is, the extent to which a comprehensive, rigorous, optimized, clear and properly justified machine learning solution was implemented. The complexity and appropriateness of the implementation are also taken into account. Note that rigorous, clear and justified can be achieved in many ways: it could be based on theoretical or empirical reasoning, it can refer to knowledge acquired before, during and after the (often iterative) implementation. Also, optimised means showing how you tried to improve the model and why.
- There is not a direct relationship between the achieved accuracy and the final mark. You are expected to do your best to get a good prediction accuracy, but we recognise that the tasks are challenging. Your mark will depend much more strongly on the soundness of your approach, evaluation and analysis than on the final accuracy.
- Critical understanding of relevant concepts, appropriate explanation, recommendations and discussion that are backed up by the results.
- Quality of the report: Are all the required sections included and completed properly? Is the executive summary clear and concise? Is the report clear, well formatted and easy to read? Does it have a logical structure? Does it have a discussion on design decisions? Is the evaluation realistic, does it show that you have really thought about your system and the client's requirements?
- An ethical discussion that is not generic but tailored to the specific task, that identifies

potential risks (with reasoning given) and provides suitable suggestions.

Detailed marks breakdown

Data exploration and assessment.	
No or little data exploration or assessment.	0 to 4.
Sufficient data exploration and assessment, but with mistakes/omissions.	5 to 8.
Great data exploration and assessment, with little to no mistakes/omissions.	9 to 10.
Data splitting and cleaning.	
No or little data cleaning or no or incorrect data splitting.	0 to 4.
Sufficient data splitting and cleaning, but with mistakes/omissions.	5 to 8.
Great data splitting and cleaning, with little to no mistakes/omissions.	9 to 10.
Data encoding.	
No or little data encoding.	0 to 4.
Sufficient data encoding, but with mistakes/omissions.	5 to 8.
Great data encoding, with little to no mistakes/omissions.	9 to 10.
Topic classification: model building.	
There an insufficient machine learning solution.	0 to 6.
There is a minimal but working machine learning solution, which may contain some major mistakes/omissions.	7 to 8.
There is a working machine learning solution, with attempts at optimisation (even successful), but with mistakes and/or omissions.	9 to 11.
There is an optimised machine learning solution, with only some to no minor mistakes/omissions.	12 to 15.
Topic classification: model evaluation.	
There is an insufficient model evaluation.	0 to 4.
The evaluation framework is appropriate, but is superficial and/or has little engagement with the client's specifications. May contain major mistakes.	5 to 6.
The evaluation framework is appropriate and has some engagement with the client's specifications, but may contain minor mistakes.	7 to 8.
The evaluation framework is appropriate and comprehensive, fully engaging with client's specifications.	9 to 10.
Text clarity classification: ethical discussion.	
Missing or highly unsatisfactory ethical discussion.	0 to 2.
There are sufficient attempts at ethical discussion, but has major to minor	3 to 5.

issues (for example, it is generic and/or irrelevant).	
The ethical discussion is specific to the task and has little to no issues.	6 to 7.
Text clarity classification: data labelling.	
No engagement with data labelling.	0.
Some engagement with data labelling, but incomplete, incorrect and/or superficial.	1 to 3.
Great and comprehensive engagement with data labelling.	4 to 5.
Text clarity classification: model building and evaluation.	
There is an insufficient machine learning solution, with no or insufficient evaluation.	0 to 5.
There is a minimal but working machine learning solution and evaluation, which may contain some major mistakes/omissions. No advanced techniques are used.	6 to 8.
There is a minimal but working machine learning solution and evaluation, which may contain some major mistakes/omissions. There are attempts at advanced techniques, but with major mistakes/omissions.	9 to 11.
There is a working machine learning solution and evaluation, with a working advanced technique, but with minor mistakes and/or omissions.	12 to 14.
There is a working machine learning solution and evaluation, with a working advanced technique, and little to no mistakes/omissions.	15 to 20.
Self-reflection.	
No self-reflection.	0.
A correct and/or thoughtful self-reflection.	1 to 3.
Writing clarity and report presentation.	
Unsatisfactory clarity and structure. Report is often hard to follow.	0 to 4.
Overall clear, with a good structure and enough details. A minority of sections may be hard to follow.	5 to 8.
Fully clear, with a good structure and all relevant details.	9 to 10.