COMP90051 Statistical Machine Learning
Project 1 Specification

Due date: 8pm Thursday 21st September 2023 (competition closes 5pm) Melbourne timezone **Weight:** 25%

Competition link: `https://www.kaggle.com/t/efb500b6d4514347ba5b374eb9b88977`

1 Overview

Text generation has become an increasingly popular task with the rise of natural language processing (NLP) techniques and advancements in deep learning. Given a short text prompt written by a human, text generation employs overparameterised models to generate response text: a likely sequence of text that might follow the prompt, based on enormous training datasets of text from news articles, online libraries of books, and from scraping the web. While text generation has a wide range of applications, including chatbots, language translation, and content creation, it also poses a significant challenge in ensuring content authenticity, accuracy, and authoritativeness. This is where text generation detection comes in, which is the process of identifying whether a given text is machine-generated or human-written. Developing effective text generation detection models is important because it can help prevent the spread of fake news, misinformation, and propaganda.

Your task:

Your task is to predict whether given text input instances have been generated by a *human* or a *machine*, given training data (features and labels) and test data (features only). You will participate as part of a group of three students in a Kaggle competition, where you upload your test predictions over the course of the project. Your mark (detailed below) will be based on your test prediction performance and a short report documenting your solution.

You will be provided with training data in two different domains, dataset1 from domain1 and dataset2 from domain2. Each dataset contains both human-generated and machine-generated text data. The machine-generated data from domain2 was generated by 7 different models, while domain1 only uses one model, which is distinct from the models in domain2. You may choose to use this fact in training. You only need to predict whether an instance is generated by human or machine, and not the specific id of the model. The performance of your approach will be evaluated through testing on test data from both domain1 and domain2.

We do not require you to have background experience in NLP, as the datasets have been preprocessed into tokens and mapped to indices in $0, \ldots, 4999$, with special token 0 for *unknown* tokens. This means that the data is represented numerically. Your goal is to focus on ML aspects of the task. To get you started, a popular baseline approach for this type of problem is called a bag-of-words model (though you are not required to implement this approach).

There are two key considerations to this task. **Firstly**, the two datasets are drawn from distinct domains, but you will not be told from which domain a test sample originates. Researchers in machine learning have developed methods for this type of setting, for example, you might want to search online for keywords domain generalisation, domain adaptation, multitask learning, and ensemble learning. **Secondly**, there is a label imbalance in the training data from domain2; you have fewer human-generated samples. The test set has a balanced label distribution for both domains, so you may want to consider how to achieve good classification accuracy on both classes in such a situation. For example, you might search online for keywords: imbalanced classification, over/under sampling, and data augmentation. We encourage you to begin with a simple approach; we don't guarantee that complex approaches will perform better. Ultimately, you are measured on your performance on predicting labels from both domain1 and domain2 in the test data.

2 Dataset

2.1 Training data

We have two training datasets: one coming from domain1, and another from domain2. You can think of these domains as different data sources or data distributions. Each contains both machine-generated and human-authored samples.

The training data is given in newline delimited JSON format, such that each line contains an instance and each instance is a dictionary with keys:

- **text:** the sequence of words, after light preprocessing, where each word has been mapped to an index in $\{0,\ldots,4999\}$
- **label:** a binary label where 0 represents **machine**-generated data and 1 represents **human**-generated.

For the machine-generated data in domain2, an additional attribute **model** is provided, containing an integer ID for the model that generated the text. Of course, this is not provided for the test data.

Two files are provided:

- domain1.json: 19,500 samples (9,750 of each class).
- domain2.json: 14,900 samples (2,150 human-generated samples, 12,750 AI-generated samples).

2.2 Kaggle Submission Format

The test data consists of 1,000 samples, split evenly between datasets and classes (ie. 250 of each class per domain). You will need to submit your predictions on the 1,000 test instances to Kaggle at least once during the project. To accomplish this, you will place your 1,000 predictions (where 0 and 1 represent *machine* and *human* labels, respectively) in a file of a certain format (described next) and upload this to Kaggle.

If your predictions are 0 for first test instance, 1 for the second test instance, and 1 for the third test instance, then your output file should be as follows in CSV format:

```
id,class
0,0
1,1
2,1
```

The test set will be used by the Kaggle platform to generate an **accuracy** measurement for your performance. You may submit test predictions multiple times per day (if you wish). Section 6 describes rules for who may submit—in short you may only submit to Kaggle as a team not individually. During the competition the **accuracy** on a 50% subset of the test set will be used to rank you in the **public leaderboard**. We will use the other 50% of the test set to determine your **final accuracy and ranking**. The split of the test set during/after the competition is used to discourage you from constructing algorithms that overfit on the leaderboard. The training data, the test set, and a sample submission file "sample.csv" will be available within the Kaggle competition website. In addition to using the competition test data, so as to prevent overfitting, we encourage you generate your own validation data from the training set, and test your algorithms with that validation data also.

3 Report

A report describing your approach should be submitted through the Canvas LMS **by 8pm Thursday 21st September 2023** . It should include the following content:

1. A brief description of the problem and introduction of any notation that you adopt in the report;

2. Description of your final approach(s) to the generation detection problem, the motivation and reasoning behind it, and why you think it performed well/not well in the competition; and

3. Any other alternatives you considered and why you chose your final approach over these (this may be in the form of empirical evaluation, but it must be to support your reasoning—examples like "method A, got accuracy 0.6 and method B, got accuracy 0.7, hence I use method B", with no further explanation, will be marked down).

4. A discussion on addressing the differences in performance of different methods in the two different domains. Provide your analysis and insights on how the domain may have affected your results, and discuss any strategies or techniques you employed to mitigate the impact of the two domains in your approach.

Your description of the algorithm should be clear and concise. You should write it at a level that a postgraduate student can read and understand without difficulty. If you use any existing algorithms, *please do not rewrite the complete description, but provide a summary* that shows your understanding and references to the relevant literature. In the report, we will be interested in seeing evidence of your thought processes and reasoning for choosing one algorithm over another.

Dedicate space to describing the features you used and tried, any interesting details about software setup or your experimental pipeline, and any problems you encountered and what you learned. In many cases these issues are at least as important as the learning algorithm, if not more important.

Report format rules. The report should be submitted as a PDF, and be no more than three pages, single column. The font size should be 11pt or above and margins should be at least 1.5cm on all sides, i.e., like this document. If a report is longer than three pages in length, we will only read and assess the report up to page 3 and ignore further pages. (Don't waste space on cover pages. References and appendices are included in the page limit—you don't get extra pages for these. Double-sided pages don't give you extra pages—we mean equivalent of three single-sided. *Three pages means three pages total.* Learning how to concisely communicate ideas in short reports is an incredibly important communication skill for industry, government, and academia alike.)

4 Submission

The final submission will consist of three parts:

- **By 5pm Thursday 21st September 2023** , submitted to the Kaggle competition website: A valid submission to the Kaggle competition. This submission must be of the expected format as described above, and produce a place somewhere on the leaderboard. Invalid submissions do not attract marks for the competition portion of grading (see Section 5).

- **By 8pm Thursday 21st September 2023** , submitted to the Canvas LMS:

  – A written research report in PDF format (see Section 3).

  – A zip archive[1] of your source code[2] of your generation detection algorithm including any scripts for automation, and a README.txt describing in just a few lines what files are for. Again, do not submit data. You may include Slack/Github logs if you wish. (We are unlikely to run your code, but we may in order to verify the work is your own, or to settle rare group disputes.) **We require your team to meet at least 3 times** and have included a template for recording high-level minutes if you wish, which can be included in the zip file.

[1] Not rar, not 7z, not lzh, etc. Zip! Substantial penalties will be applied if you don't follow this simple instruction.

[2] We would encourage you to use Python, but we will also accept submissions in Matlab, R, or otherwise. You are welcome to use standard machine learning libraries, such as sklearn, pytorch, etc, but the code submitted should be your own. Do not submit data as Canvas has limited storage space.

– Your Kaggle team name—without your exact Kaggle team name, we may not be able to credit your Kaggle submission marks which account for almost half of project assessment.

The submission link will be visible in the Canvas LMS prior to deadline.

Note that after about a week into Project 1 you will need to also submit a **Group Agreement**. While not a formal legal contract, completing the agreement together is a helpful way to open up communication within your team, and align each others' expectations. This agreement doesn't attract marks. Finally, we will conduct an unmarked]**Project 1 group evaluation quiz** after the project is completed, to allow students to provide feedback to staff on their own performance and the performance of their team mates. Should significant team disputes be identified through this quiz, we reserve the right to follow-up with students and consider differential marking. This is a rare occurrence and we find that almost all teams work well together.

5 Assessment

The project will be marked out of 25. Note that there is a hurdle requirement on your combined continuous assessment mark for the subject, of 25/50, of which Project 2 will contribute 25 marks. **Late report submissions will incur a deduction of 2 marks per day—it is not possible to mark late competition entries.**

The assessment in this project will be broken down into two components. The following criteria will be considered when allocating marks.

Based on our experimentation with the project task, we expect that all reasonable efforts at the project will achieve a passing grade or higher. Therefore, please have fun with this project, and focus on your learning!

Kaggle Competition (**12/25**):

Your final mark for the Kaggle competition is based on your rank in that competition, calculated on a held-out 50% portion of the test set which is distinct from the data used to calculate the public leaderboard throughout the competition. Assuming N teams of enrolled students compete, there are no ties and your team comes in at R place (e.g. first place is 1, last is N) with an accuracy of $ACC \in [0,1]$ then your mark is calculated as[3]

$$9 \times \frac{\max\{\min\{ACC, 0.9\} - 0.5, 0\}}{0.4} + 3 \times \frac{N-R}{N-1} \ .$$

Ties are handled so that you are not penalised by the tie: tied teams receive the rank of the highest team in the tie (as if no entries were tied). This expression can result in marks from 0 to 12. For example, if teams A, B, C, D, E came 1st, 4th, 2nd, 2nd, 5th, then the rank-based mark terms (out of 3) for the five teams would be 3, 0.75, 2.25, 2.25, 0.

This complicated-looking expression can result in marks from 0 all the way to 12. We are weighing more towards your absolute ACC than your ranking. The component out of 9 for ACC gives a score of 0/9 for ACC of 0.5 or lower; 9/9 for ACC of 0.9 or higher; and linearly scales over the interval of ACC [0.5, 0.9]. We believe that a mid-way ACC is achievable with minimal work, while results around 0.9 are good, but will require more effort. *An ACC of 0.9 for a student coming last would yield 9/12; or 10.5/12 if coming mid-way in the class.*

External unregistered students may participate, but their entries will be removed before computing the final rankings and the above expression, and will not affect registered students' grades. We do not however actively invite such participation.

The rank-based term encourages healthy competition and discourages collusion. The other ACC-based term rewards students who don't place in the top but none-the-less achieve good absolute results. Therefore you can achieve a high H1 grade overall irrespective of your placing in the ranking.

Note that invalid submissions will come last *and* will attract a mark of 0 for this part, so please ensure your output conforms to the specified requirements.

[3]Note that Kaggle is set to take two "scored private submissions" per team. These mean that by default, your top two submissions based on public leaderboard score are chosen, then after competition close these two submissions will be scored on the entire test set, and the best total test scoring submission will make up your accuracy *ACC*. If you prefer non-default submissions for your two "scored private submissions" Kaggle permits you to select others prior to completion of the competition.

Report (**13/25**):
The marking rubric in Appendix A outlines the criteria that will be used to mark your report.

6 Additional Competition Rules

Teams: You are required to form a team of three students including yourself. The Ed Discussion category "Assignment - A1" is available to help with this—e.g. you can post if you're looking there. It isn't required that teams sit in one workshop.

Group account submissions to Kaggle: Only submissions from your group account are permitted on Kaggle. You should not attempt to submit from your individual account or create additional accounts to circumvent the daily submission limit. We have set the submission limit high so no advantage is gained from circumvention. Moreover submitting too many times is likely to risk overfitting to the leaderboard portion of the test data.

Auxiliary data prohibited: The use of any additional datasets to help with your solution is prohibited. You should only use the data that we have supplied for you. You should not search for original data sources, and we have deployed a number of obfuscation measures to prevent this.

Plagiarism policy: You are reminded that all submitted project work in this subject is to be your own individual work. Automated similarity checking software will be used to compare submissions. It is University policy that academic integrity be enforced. For more details, please see the policy at `http://academichonesty.unimelb.edu.au/policy.html`.

See more over page

A Marking scheme for the Report

Critical Analysis (Maximum = 8 marks)	**Report Clarity and Structure** (Maximum = 5 marks)
8 marks Final approach is well motivated and its advantages/disadvantages clearly discussed; thorough and insightful analysis of why the final approach works/not work for provided training data; insightful discussion and analysis of other approaches and why they were not used	5 marks Very clear and accessible description of all that has been done, a postgraduate student can pick up the report and read with no difficulty.
6.4 marks Final approach is reasonably motivated and its advantages/disadvantages somewhat discussed; good analysis of why the final approach works/not work for provided training data; some discussion and analysis of other approaches and why they were not used	4 marks Clear description for the most part, with some minor deficiencies/loose ends.
4.8 marks Final approach is somewhat motivated and its advantages/disadvantages are discussed; limited analysis of why the final approach works/not work for provided training data; limited discussion and analysis of other approaches and why they were not used	3 marks Generally clear description, but there are notable gaps and/or unclear sections.
3.2 marks Final approach is marginally motivated and its advantages/disadvantages are discussed; little analysis of why the final approach works/not work for provided training data; little or no discussion and analysis of other approaches and why they were not used	2 mark The report is unclear on the whole and the reader has to work hard to discern what has been done.
1.6 mark Final approach is barely or not motivated and its advantages/disadvantages are not discussed; no analysis of why the final approach works/not work for provided training data; little or no discussion and analysis of other approaches and why they were not used	1 mark The report completely lacks structure, omits all key references and is barely understandable.