University of Aberdeen
School of Natural and Computing Sciences
Department of Computing Science
MSc in Artificial Intelligence
2024 - 2025

Assessment 2

CS552J - Data Mining with Deep Learning

This assignment is individually assessed and accounts for 50% of your total mark for the course.

Learning outcomes On successful completion of this component a student will have demonstrated competence in the following areas:

➔ Using a non-trivial dataset, plan, execute and evaluate significant experimental investigations using multiple data mining, visualization and machine learning strategies

Information for Plagiarism and Conduct: Your submitted report and source code may be submitted for plagiarism check (e.g., Turnitin). Please refer to the slides available at MyAberdeen for more information about avoiding plagiarism before you start working on the assessment. Please also read the following information provided by the university: https://www.abdn.ac.uk/sls/online-resources/avoiding-plagiarism/

In addition, please familiarize yourselves with the following document code of practice on student discipline (Academic)

Application Problem Definition: Analysing dialogues between speakers of different ages

The objective of this assessment is to analyze a dataset consisting of transcripts from conversations between speakers from the British National Corpus. This corpus contains dialogues between a range of speakers. Dialogue data is particularly interesting and challenging to analyze due to turn taking patterns, and determining properties of language that is informal and a lot less structured than typical discourse data such as news articles or reviews. We can use NLP techniques to gain insights into various domains including language development, education, discussion forums and much more. Being able to analyze this form of language can be very useful to companies, linguists or simply interesting to various stakeholders who wish to understand this kind of data. Insights can consist of identifying features that correlate with certain properties, in this case, information is provided about the age of the speakers, the number of speakers in the dialogues, and other demographic information.

The dataset can be downloaded from http://cass.lancs.ac.uk/cass-projects/spoken-bnc2014/, you will need to register an account and sign a user agreement, then the file will download with all the metadata. It consists of naturalistic spoken dialogues gathered from volunteers across the UK. Some speakers occur in multiple dialogues, some dialogues contain more than two participants, and all dialogues vary in terms of the age range, and difference in ages between speakers. You will need to download the dataset from the link above, and from there, load and process the data. All the information about the metadata and structure are in the Readme.

The task is to extract interesting insights from these dialogues, which answers questions such as whether there are interesting patterns in terms of speaker age and language use. Is the language used in dialogues between speakers of the same age different from the language used between speakers of different ages? Is the language used in larger groups different from spaller groups of speakers? You will report on the statistics of the dataset, noting the prominent features which differ across speakers, using appropriate visualization and clustering techniques from the practicals. You will then identify the most important aspects of the language used by a particular subset of the data of your choice, either via summarized text, or identification and analysis of main clusters of features as well as justification for the choice of method. You will then train a classifier to predict something interesting from this data: it could be a) a classifier for predicting speaker age, or whether a dialogue contains same vs differently aged speakers (for this you will need to carefully select the relevant subset of data for training & analysis),: this will need a strong baseline, guided by the features you found during the exploratory analysis; or b) some other useful application of the tools you have covered in the practicals (e.g. you may feel ambitious and want to apply other aspects covered in the practicals to this problem). Finally, you will provide some appropriate evaluation.

Feature information in the dataset includes speaker_age, number of participants in the dialogue, utt_id (which indicates the order of the utterances in the dialogue), speaker, (which speaker the utterance_text belongs to) utterance_text (the utterance text itself from which you can derive more features)

No prior knowledge of the domain problem is needed or assumed to fulfill the requirements of this assessment.

Report Guidance & Requirements

Your report must conform to the below structure and include the required content as outlined in each section. Each subtask has its own marks allocated. You must supply a written report (pdf), along with the corresponding source code written in python (single, well-documented jupyter notebook), containing all distinct sections/subtasks that provide a full critical and reflective account of the processes undertaken. **All details and results must be included in the report**, your code is only for evidence to support the information in the report. The report should be written in a formal manner, and results clearly presented and rationale described. Report should be created using the latex template provided, and follow the style guidelines within.
The following task requires you to expand and elaborate upon the principles of data mining, different components of machine learning and some aspects on how such techniques can be used in real-life problems such as in text mining. Use the dataset provided as an example

Programming Task Report: (~max 2000 words/ roughly 4 pages)

The report should contain the *description of data and methods* (1) as well as the *results of the task* (2), reported clearly and concisely and contain the elements outlined in this section as well as the information addressing the subtasks below. All reports should make use of the latex template provided, and be submitted in pdf form.

1. Description of Data and Methods (10 marks)
Using your own words, the lecture material and any other relevant sources, explain specifically:

1. What the dataset consists of and cite its source
2. What basic preprocessing steps you have used to work with this data be needed to work with this data (e.g. tokenization, embeddings etc) and which specific libraries are used

3. What feature extraction or analysis methods did you use to gain insights into the data before modeling? (basic counts like sentence length, proposition of pos tags, using linguistic resources, or clustering)
4. What specific challenges and trade-offs did you consider and why? Add motivating examples and citations to justify choices (e.g. summarisation, making use of transformers, dataset size)
5. Evaluation & Error analysis: How did you evaluate your methods? how did you examine errors? (e.g. metrics of fluency, faithfulness, or use of confusion matrices, precision, recall, F1 score, evaluating performance in different subsets of data)

2. Programming Task: Investigate and report on insights from dialogue data
The problem we aim at tackling has been clearly described and defined earlier. This task includes *six subtasks*, each of which bears its own marks.
Subtasks:

1. Load the dataset provided. You may use CoLab or a jupyter notebook. Please create a table providing summary statistics of the properties in this dataset, i.e. mean values, range, standard deviations, min/max values, median values and 25%/50%/75% percentile values. The main columns of interest to us are the speaker's age demographics in the dialogue, and should also contain basic information about dataset size, diversity, average dialogue length, whether this varies by number of speaker etc. In particular this table should contain statistics of the text data in utterance text, such as average utterance length per speaker, as well as more interesting statistics of the text. *(5 marks).*
2. Visualize the data as you think is appropriate using techniques from the practicals. You should provide at least 3 graphs, each displaying multiple dimensions of the data. Each plot should be referenced in the text, described clearly and supported by statistical analysis (e.g. t-tests or correlations) (example: what are the average age differences between speakers? Do dialogues with speakers of the same age contain more overlapping vocabulary?) *(10 marks)*.
3. Extract a summary of key components of the data: either per age group (you can create age-range bins) or number of speakers per dialogue. Motivate the choice of summarisation techniques, and summary size. In your own words describe insights that can be gained from this summary. (for example, you extract summaries of topics in dialogues between younger vs older speakers. You could additionally employ clustering as a step to help identify the main themes)*(10 marks)*
4. Train a simple classifier (e.g. logistic regression) using features extracted from subtask 1. (e.g. A basic feature would be using the words as features, or the vocabulary overlap between speakers) Split the dataset provided into training (80%) and test (20%) sets. Please use the training set to train your developed model, keeping the test set only for evaluating its performance in unseen data. Report the most informative features for your selected task *(5 marks).*
5. Report on the model's performance, using metrics such as i.e. Precision, Recall, and Accuracy, as well as including error analysis. When reporting performance, please only use the test set created by yourselves *(5 marks).*
6. Repeat steps 4 and 5 but using a more complex classifier this time, compare the accuracies in a table. *(5 marks)*

3 Bonus – Optional

Should you decide to try a bonus task, there will be a reward of *5 marks*. The maximum overall mark for this assessment remains at 50/50; however, attempting the bonus exercise will a) make you practice with an alternative distributed library and b) enhance your chances of getting a higher mark overall. To gain these marks, you will need to show you have synthesized the data mining issues covered in practicals

and lectures and are able to take them further: adapting a model, adopting advanced visualization, coming up with something that is novel.
Possible tasks:

- Include Temporal aspects of dialogue: extract some utterance level statistics and analyze the dialogue data as a time-series
- Visualize: use an appropriate large pretrained language model to encode the utterances and visualize cosine similarity between utterances of different classes (i.e. are child and adult utterances very different? Do clusters accurately reflect age groups? Try clustering them and reporting the most common tokens in each cluster. Report on results. (hint you may want to create balanced age ranges)
- Be creative: pick some aspect of the course that you think can apply to this dataset and try it. Be sure to report the motivation and method clearly, as well as the outcome.

Only attempt to gain bonus marks once you are satisfied you have met the criteria for the rest of the assessment to the best of your ability.

Marking Criteria

- ***Quality of the report***, including structure, clarity, and brevity: is your writing specific and to the point? *- please report word count.*
- ***Reproducibility***. Can another MSc AI student repeat your work based on your report and code?
- ***Quality of your experiments***, including design and result presentation (use of figures and tables for better reporting) Configured to complete the task and the parameter tuning process (if needed)
- ***In-depth analysis of results*** generated, including critical evaluation, insights into data, and significant conclusions
- ***Quality of the source code***, including the documentation of the code

Submission Instructions

You should submit your work via MyAberdeen by **23:59 09/05/2025**. Both the report and the code should be submitted together in the form of a **zipped folder**. The naming convention for the files should be as follows: CS552J_Assessment2_Lastname_Firstname_StudentNumber.zip

Include within Zipped folder

- ➔ **Report** The name of the PDF file should have the same naming convention: For instance, if I was a student with ID number 4568985, my submission file name would be: CS552J_Assessment2_Sinclair_Arabella_4568985.pdf.
- ➔ **Latex source** You should also include the latex source code as evidence you used the template provided to create your pdf.
- ➔ **Python Notebook** submit supplementary material containing the source code of your implementation (as a python notebook ".ipynb"). Your script should use markdown to describe clearly what your code does. It should follow the same naming convention as the other files.

Please **do not submit any training data** on MyAberdeen. Please try to make your submission file less than 20MB as you may have issues when uploading large files to MyAberdeen.

Any questions pertaining to any aspects of this assessment, please address them to the course coordinator Arabella Sinclair (arabella.sinclair@abdn.ac.uk)