EMATM0067
Text Analytics Coursework

Spring 2025, Lecturer: Edwin Simpson.

Deadline: 13.00 on Monday 28th April

Overview

This coursework is worth 50% of the unit. It will take you through several text analytics tasks to give you experience with applying and analysing the techniques taught during the labs and lectures. The work will be assessed through both your code and your written report, in which you should aim to demonstrate your understanding of text analytics methods, evaluate the methods critically and incorporate ideas from the lectures.

We recommend that you first get a basic implementation for all parts of the required assignment, then start writing your report with some results for all tasks. You can then gradually improve your implementation and results.

Total time required: 40 hours.

Support

The lecturers and teaching assistants are available to provide clarifications about what you are required to do for any part of the coursework. You can ask questions during our lab sessions, post questions on MS Teams, or to the Blackboard discussion forum. If you don't want to share your question with the class, please contact Edwin by email (edwin.simpson@bristol.ac.uk).

Part 1: Climate Sentiment – Jupyter Notebook (max 32%)

Many companies are required to publish 'corporate disclosures' – documents containing useful information about the business and its finances. The information in these disclosures is very useful for investors, regulators and other stakeholders. For example, disclosures may present climate-related developments as risks or opportunities for the business. In this part of the assignment, you will compare classifiers that classify the sentiment of a text as a climate risk, an opportunity, or neutral. We will be working with the ClimateBERT dataset: Webersinke et al., 2022.

Part 1 contains tasks *1.1 (13 marks), 1.2 (8 marks),* and *1.3 (11 marks)*. Please see the accompanying Jupyter notebook text_analytics_part1.ipynb for details, which contains a series of tasks for you to complete. Your answers to tasks 1.1, 1.2 and 1.3 should be saved in the notebook itself, which you will need to submit. Submission details are in the notebook.

Task 2: Climate Sentiment – Report (max. 38%)

2.1. Present an evaluation of the three methods implemented in Part 1. Your evaluation should be presented as the first page of your report. Your evaluation should include the following points:

- Explain the modifications you made to the naïve Bayes classifier in task 1.1c: what did you change, how does it help the classifier, and was there anything you tried that didn't work? (3%, max. 150 words)

- Present a comparison of results in a table or plot, along with your interpretation of how well each method worked. Your discussion should mention concepts from the lectures (e.g., transfer learning) and what could be improved in future work. To inform this discussion, you may want to analyse some examples of misclassified texts.

(10%).

2.2. Using the dataset, can you identify topics that are associated with climate-related risks or opportunities?

- Explain the method you use to identify themes or topics. Make sure to motivate why you chose this approach and discuss its limitations.
- It is important to test and compare different approaches to find out what works best for this dataset. Compare two variations of your method, e.g., by changing an important step or parameter.
- Show your results (e.g., by listing or visualising topics associated with risks or opportunities).
- Interpret the results and summarise the limitations of your approach.

(25%)

Suggested length of report for task 2: 2.5 – 3 pages.

Commented [ES1]: Reword to reduce need to include textbook method descriptions

Task 2: Named Entity Recognition on Twitter (max. 30%)

Social media contains a wealth of information about public opinion and events, but this is often contained in unstructured text data. Your task is to build a tool for named entity recognition from Twitter posts that can help extract information about particular people, organisations and locations. To train and test the NER tagger, we will use the Broad Twitter Corpus (BTC) dataset, published by Derczynski et al., 2016. You can also find useful information on the HuggingFace dataset page.

2.1. Design and run **a sequence tagger** for the BTC dataset. Refer to the labs, lecture materials and textbook to identify a suitable method. You may choose any sequence tagging method you think is suitable, and you may wish to experiment with some variations in the choice of features or model architecture to help justify your design. In your report:

- Briefly explain your chosen method and its main strengths and limitations.
- If your model uses its own tokenizer, explain how you align the tokens with tags (this step is only needed if you use a neural sequence tagger that requires a particular tokenizer).
- Show an example entity span from the dataset, that illustrates how entity spans are encoded as tags in this dataset.
- Detail the features you have chosen, why you chose them, and hypothesise how your choice will affect your results.
- Higher marks are given for good, well-justified model design.

(17 marks)

2.2. Evaluate your method, then interpret and discuss your results. Include the following points:

- Explain your choice of performance metrics and their limitations.
- Describe the testing procedure (e.g., how you used each split of the dataset).
- Show your results using suitable plots and/or tables.

Commented [ES2]: Reword to reduce need to include textbook method descriptions

- Do your methods make any particular kinds of error? Show some examples of mislabelled sentences and suggest how the methods could be improved in future.

(13 marks)

Suggested length of report for task 2: 2 pages.

Implementation

The lab notebooks provide useful example Python code, which you may reuse. You may libraries introduced in the labs, or others of your choice. For tasks 2 and 3, you may write your code in either Jupyter notebooks or standard Python files.

Report Formatting

- Absolute maximum 5 pages
  - References do not count toward the page limit.
  - Aim for quality rather than quantity: you will receive higher marks if you write concisely.
- To set the page layout, fonts, margins, etc., we recommend using the template from an academic conference, such as LREC-COLING 2024 if writing the report in Latex
  - You can use this template directly to write in Latex[1] or follow the formatting style in Word, Libreoffice, etc.
  - You don't need to include an abstract or introduction or conclusion.
  - Please number your answers to each task clearly so that we can find them.
  - No less than 11pt font
  - Single line spacing
  - A4 page format
- The text in your figures must be big enough to read without zooming in.

Citations and References

Make sure to cite a relevant source when you introduce a method or discuss results from previous work. You can use the citation style given in the LREC-COLING 2024 style guide above. The details of the cited papers must be given at the end in the references section (no page limits on the references list). Please only include papers that you discuss in the main body of the report.

Google Scholar and similar tools are useful for finding relevant papers. The 'cite' link provides bibtex code for use with latex and references that you can copy, but beware that this often contains errors.

Submission

- Deadline for report + code: please see first page.
- On Blackboard under the "assessment, submission and feedback" link.

Please upload the following **three files:**

1. Your submission for task 1: please see the details in the Jupyter notebook. It should be submitted to the submission point "Text Analytics Part 1 Notebook".

[1] Latex is the most common tool for writing published papers in Computer Science and AI research. A good way to get started with Latex is to use https://www.overleaf.com/.

2. Your report for tasks 2 and 3 as a **PDF with filename <student_number>.pdf,** where "<student_number>" is replaced by your student number from eVision (starting with a '2', not your username).
    - Upload this to the submission point marked "Turnitin submission point - Text Analytics Coursework".
    - **Please don't include your name in the report itself:** to ensure fairness, we mark the reports anonymously.
3. Your code for tasks 2 and 3 a **single zip file with filename <student_number>.zip.**
    - Inside the zip file there should be a single folder containing your code, with your student number as the folder name.
    - Please remove datasets and other large files to minimise the upload size.
    - Upload this file to the submission point "Text Analytics Parts 2 & 3 Code".
    - For tasks 2 and 3, your marks will be based on the contents of your report, rather than for good code structure or style. Assessment Criteria

To gain high marks, your report will need to demonstrate a thorough understanding of the tasks and the methods used, backed up by a clear explanation of your results and analysis or errors. Marks will be awarded for appropriately including concepts and techniques from the lectures.

Avoiding Academic Offences

Please re-read the university's plagiarism rules to make sure you do not break any rules. Academic offences include submission of work that is not your own, falsification of data/evidence or the use of materials without appropriate referencing. Note that sharing your report with others is also not allowed. These offences are all taken very seriously by the University.

Do not copy text directly from your sources – always rewrite in your own words and provide a citation.

Work independently – do not share your code or reports with others.

Do not use AI to generate your answers – this includes automatically translating passages of text from another language.

Suspected offences will be dealt with in accordance with the University's policies and procedures. If an academic offence is suspected in your work, you will be asked to attend an interview with senior members of the school, where you will be given the opportunity to defend your work. The plagiarism panel can apply a range of penalties, depending on the severity of the offence. These include a requirement to resubmit work, capping of grades and the award of no mark for an element of assessment.

Extensions and Exceptional Circumstances

If the completion of your assignment has been significantly disrupted by serious health conditions, personal problems, or other serious issues, you can apply for consideration in accordance with the normal university policy and processes. Students should refer to the guidance and complete the application forms as soon as possible when the problem occurs. Please see the guidance below and discuss with your personal tutor for more advice:

https://www.bristol.ac.uk/students/support/academic-advice/assessment-support/request-a-coursework-extension/

https://www.bristol.ac.uk/students/support/academic-advice/assessment-support/exceptional-circumstances/