EMATM0067
Text Analytics Coursework

Spring 2024, Lecturer: Edwin Simpson.

Deadline: 13.00 on Wednesday 22nd May

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 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).

Task 1: Emotion Classification in Tweets (max 59%)

People often express opinions and feelings on social media sites and processing them automatically can help to identify patterns and trends, from medical symptoms to market sentiment or the popularity of a product. A key challenge is to recognise the emotions that the authors express.

Your task is to design, run and evaluate an emotion classifier for social media posts using the ***TweetEval Emotion*** dataset, which contains English tweets tagged with (0) anger, (1) joy, (2) optimism or (3) sadness. You may use any existing classifier implementations in libraries such as Scikit-learn, Gensim, NLTK and Transformers to achieve this. We provide a copy of the data and a 'data_loader_demo' Jupyter notebook containing code for loading the data. The notebook is available in our Github repository. Further information about the dataset is available on Huggingface and in the paper by *Barbieria et al., "TWEETEVAL: Unified Benchmark and Comparative Evaluation for Tweet Classification", Findings of EMNLP 2020.*

1.1. Train **one non-neural**[1] method for classifying emotions in tweets. Refer to the labs, lecture materials and textbook to identify a suitable method. In your report:

- Briefly explain how your chosen method works and its main strengths and limitations.

[1] A method that does not use a neural network or 'deep learning'.

- Describe the preprocessing steps and the features you use to represent each text instance.
- Explain why you chose those features and preprocessing steps and hypothesise how they will affect your results.
- Higher marks are given for good, well-justified classifier design.

(10 marks)

1.2. Train **one neural network-based** method for classifying emotions in tweets. Refer to the labs, lecture materials and textbook to identify a suitable method. In your report:

- Briefly explain how your method works, including details of the model architecture and how you chose this configuration.
- Discuss any use of model transfer or transfer learning in your approach.
- State the method's strengths and limitations in comparison to your previous method.
- Describe any preprocessing steps needed to prepare the data.
- Plot the changes in the losses during training as learning curves. Explain what the learning curves show and how this information can be used during training.
- Higher marks are given for good, well-justified classifier design.

(15 marks)

1.3. Evaluate both methods, then interpret and discuss your results. Include the following points:

- Define your performance metrics and state their limitations.
- Describe the testing procedure (e.g., how you used each split of the dataset).
- Show your results using suitable plots or tables.
- How could you improve the method or experimental process? To inform this discussion, you may want to analyse some examples of misclassified texts.

(14 marks)

1.4. Using the dataset, can you identify topics that people appear to be optimistic or joyful about?

- Explain the method you use to identify themes or topics.
- Show your results (e.g., by listing or visualising example topics or themes).
- Interpret the results and summarise the limitations of your approach.

(20 marks)

High performance figures are less important for getting high marks than motivating your method well and implementing and evaluating it correctly.

Suggested length of report for task 1: 4 pages.

Task 2: Named Entity Recognition (max. 41%)

In scientific research, information extraction can help researchers to discover relevant findings from across a wide body of literature. As a first step, your task is to build a tool for named entity recognition in scientific journal articles. We will be working with the ***Bio Creative V*** dataset containing sentences from articles on PubMed, a database of biomedical research literature. Each sentence is annotated with mentions of chemicals and diseases. We provide a cache of the data and code for loading the data in 'data_loader_demo' in our Github repository. The data can be sourced from HuggingFace. More information can be found in *Wei, Chih-Hsuan, et al.* "Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task." *Database 2016 (2016).*

2.1. Design and run **a sequence tagger** for tagging chemicals and diseases in Bio Creative V. 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:

- Explain how your chosen method works 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).
- Briefly explain how entity spans are encoded as tags for each token in a text.
- 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.

(15 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.
- 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.

(14 marks)

2.3. This task requires you to apply techniques for computing similarity between words or phrases.

- Select one disease entity from the test set as a "query".
- Use two techniques to identify five similar and five dissimilar diseases to your query.
- Explain and compare the results from each technique. You may wish to use tables or figures to support your discussion.
- Marks are given for correct use of the techniques, your understanding of them, and your interpretation of the results. If it supports your interpretation, you may include more than one query entity.

(12 marks)

Suggested length of report for task 2: 3 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. You may write your code in either Jupyter notebooks or standard Python files.

Report Formatting

- Absolute maximum 8 pages
  - References do not count toward the page limit.

    - Aim for quality rather than quantity: you do not have to use the maximum number of pages and will receive higher marks if you write concisely and clearly.
- 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[2] 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: 13:00 (GMT+1) on 22nd May.
- On Blackboard under the "assessment, submission and feedback" link.

Please upload the following **two files:**

1. Your report as a **PDF with filename <student_number>.pdf,** where "<student_number>" is replaced by your student number (not your username). Upload this to the submission point "Text Analytics Coursework (Turnitin)".
2. Your text analytics code inside 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 – we only need the code itself. Upload this file to the submission point "Code for Text Analytics Coursework".

We will briefly review your Python code by eye – we do not need to run it. Your marks will be based on the contents of your report, with the code used to check how you carried out the experiments described in your report. We will **not** give marks for the coding style, comments, or organisation of the code.

Please do not include your name in the report text itself: to ensure fairness, we mark the reports anonymously.

[2] 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/.

Assessment Criteria

Your coursework will be evaluated based on your submitted report containing the presentation of methods, results and discussions for each task. 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 (including figures) of your results and error analysis. The exact structure of the report and what is included in it is your decision and you should aim to write it in a professional and objective manner. 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.

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/