Independent Project

Assignment 3: Text Analytics for Health [COMP90090 2024]

Mike Conway

5th March 2024

Version: 1.2

1 Introduction

Assignment 3 forms the main summative assessment for this subject. The assignment requires that you *design*, *implement*, and *evaluate* a health NLP project that addresses a *relevant health problem*. The final deliverable consists of a project report equivalent to approximately 2,000 words. This report will constitute 45% of your marks for this subject.

Note that in assessing your project, I am more interested in your critical analysis of methods and results than the raw performance of your models. You may not be able to arrive at a definitive answer to your research question, which is perfectly fine. However, you should analyse and discuss your (possibly negative) results in depth. You will be assessed on the quality of your report, **not** your code. However, your code should be sufficient to replicate the results provided in your report (i.e. in principle, it should be possible to run the code you supply to generate the results provided in your report). You are not required to implement algorithms from scratch. Using existing library implementations of algorithms is encouraged.

The report will consist of:

1. An introduction to the research question you have decided to address and its significance
2. A description of the resources and methodologies adopted, including evaluation strategies
3. A description and discussion of the results obtained, the implications of these results, how well your results align with the research literature, and future steps

You should write the report using the JAMIA (*Journal of the American Medical Informatics Association*) format. A description of the format can be found here, but briefly, reports should consist of:

- **Structured abstract** (up to 250 words containing the headings *Objective*, *Materials and Methods*, *Results*, *Discussion*, and *Conclusion*)
- **Introduction** – Provide a short description of the problem and data set, and the research question addressed. The Introduction section should include a summary of some related literature
- **Methods** – Explain the primary NLP methods that you have used in your project with appropriate references. This section should include a description of your corpus and an ethics

statement [Note that the ethics statement does not count towards the word count for the assignment]
- **Results** – Present your results in terms of evaluation metric(s) and, ideally, illustrative examples. Use of tables and diagrams is highly recommended
- **Discussion & Conclusion** – Contextualize and reflect on your results in the context of your research question. What has been learnt? What are the limitations of your project?
- **References** – a specific referencing style is not required, but consistency in how you format references is important
- **Appendixes** [including all code used]

The Methods section should include an ethics statement [use the heading **Ethics Statement**]. In most circumstances, ethics statements can be short and should typically discuss (a) whether the data is public; (b) any plans to redistribute data generated during the project; and (c) any sensitivity involved with the data. Health data is inherently sensitive, even if it is public, but some data (e.g. regarding stigmatized conditions like, say, sexually transmitted infections) is especially sensitive. Example ethics statements from the *Association of Computational Linguistics* proceedings include very simple statements, like:

> No additional ethics approval was sought for the analysis of data in this study because data were drawn from already published studies.[1]

To more elaborate statements, like:

> The ethical concerns of this work are two-fold. First, readers must be aware that such a deep learning model is prone to make mistakes, as evidenced by the results of the experiments we did (see Section 5). Outputs should be treated as an indication or recommendation, rather than the ground truth. Secondly, our QA-based approach needs to train a single model, by comparison with the summarization-based one that requires 30 models. Having a single model reduces the pressure on computing resources and consequently, on the environment. It also makes the model easier to maintain.[2]

For most projects, ethics statements are likely to consist of only two or three sentences.

For inspiration, here are a few examples of ethics statements from computer science conference proceedings:

- https://ojs.aaai.org/index.php/ICWSM/article/view/22133/21912
- https://ojs.aaai.org/index.php/ICWSM/article/view/22160/21939
- https://aclanthology.org/2022.coling-1.198.pdf
- https://aclanthology.org/2022.coling-1.261.pdf

Note that the Ethics Statement will not count towards the word count

My expectation is that you will cite 15 to 30 references in writing your report. Note that it does not matter which citation style you select, as long as it is consistent throughout the report.

You can write the report using Microsoft Word, LaTeX, or Markdown. Templates are available for LaTeX and Markdown. Note that you can use as many tables and figures as you like, but try

[1] Naseem et al. (2022). Benchmarking for public health surveillance.
tasks on social media with a domain-specific pretrained language model. *Proceedings of the 2022 Workshop on Efficient Benchmarking in NLP*. https://aclanthology.org/2022.nlppower-1.3.pdf

[2] Boissonnet et al. (2022). Explainable assessment of healthcare articles with QA. *Proceedings of the 2022 BIONLP Workshop*. https://aclanthology.org/2022.bionlp-1.1.pdf

and keep the word count around 2,000. However, up to 2,500 is acceptable. Note that there is no word limit on material in the appendixes (please include all the code you use as an appendix). You can find a couple of example JAMIA papers here, but note that these are provided just to help clarify the *structure* of the report in terms of the major headings. You are definitely not expected to produce a report as comprehensive as a published journal paper.

Your report will be assessed against the following criteria:

- Clarity of expression and cohesion
- Adequate contextualisation of the health problem in the literature
- Adequate use of visual aids (plots, figures, flowcharts, tables)
- Adequate methodological steps (i.e. sufficient detail to – in principle – replicate the work)
- Adequate discussion of results, implications, and future directions

You can see a more detailed breakdown in Section 3 of this document.

The report should be around 2,000 words in length. Please submit the file as a PDF file with your name, your student ID number, and your report's title clearly indicated on the title page.

Please title your file as follows, replacing {NAME} with your surname:

`NLP4Health_assignment3_{NAME}.pdf`

This assignment is worth 45% of your total marks for the subject.

2 Example Projects from Previous Year

A previous iteration of this subject ran in 2023. These were some of the project names that were submitted:

- Classification of mental health status by analysing online text
- Automated extraction of adverse drug events from medical reports: a comparative analysis of NLP techniques
- Optimisation and deployment challenges of closed-source LLMs for clinical note abbreviation expansion
- Classification of stress-related posts in Social Media
- What really matters? an empirical analysis of best-practice methods for building a binary stress classifier for Reddit posts
- Exploring the shift in food habits after the pandemic and its potential implications for public health in the United States

3 Academic Misconduct

For most students, discussing ideas with peers will form a natural part of the undertaking of this project. However, it is still an individual task, and so reuse of ideas or excessive influence will be considered cheating. I will be checking submissions for originality and will invoke the University's Academic Misconduct policy where inappropriate levels of collusion or plagiarism are deemed to have taken place. Content produced by generative AI (including, but not limited to, ChatGPT) is not in a straightforward sense your own work, and submitting such content may be treated as a case of academic misconduct, in line with the University's policy.

4 Learning Objectives

1. Evaluate a range of health-related text data sources
2. Develop and evaluate NLP pipelines using the Python programming language
3. Employ appropriate health-related NLP libraries and knowledge resources
4. Critically appraise ethical issues in health NLP
5. Evaluate health NLP publications

5 Detailed Rubric

Please find below a guide to how marks will be allocated (45 points in total are available for this project).

5.1 Style (total 5 pt)

[**4 pt - full marks**] This document contains ALL the required components: title page, structured abstract, introduction, methods, results, discussion & conclusion, references, appendixes

[**0 pt**] Document does not contain required components

5.2 Motivation and Background (total 6 pt)

[**6 pt - full marks**] The document clearly defines the health problem being addressed, including the existing knowledge and approaches to address it. The document provides a well-supported description of the context and its relevance.

[**0 pt**] The health problem is not clearly stated, or there is not enough context provided to understand the relevance of the problem.

5.3 Methodology (total 6 pt)

[**6 pt - full marks**] The methods section provides an exceptionally clear and accurate description of the proposed methods, with a clear and well-argued connection between the research question/task, the data available and the methods proposed. A short ethics statement is included.

[**0 pt**] The methods are not clearly presented or argued; ethics section is missing

5.4 Results (total 6 pt)

[**6 pt - full marks**] The result section does an outstanding job at presenting the results in a consistent and coherent manner. The results are clearly understood and it uses a well-balanced combination of text, tables, and figures to explain the results. The results are consistent with the health problem, the task and the methods used.

[**0 pt**] The results section is incoherent and not clearly presented

5.5 Discussion (total 6 pt)

[**6 pt full marks**] The discussion section provides an outstanding reflection on the results and how they connect with the original health problem and task proposed. The section includes a reflection on the implications of the results and a discussion of limitations and future work.

[**3 pt - 0 pt**] The discussion section is poorly written, with little/no reflection on the results or is a mere reiteration of the results presented in the results section. The discussion presents the results in an excessively positive/negative tone.

5.6 Communication & Visual Aids (total 6 pt)

[**6 pt full marks**] Communication aids enhance the presentation of complex concepts or results. The diagrams/figures are visually pleasant and well-designed.

[**4 pt - 2 pt**] Visual aids do not improve the communication of complex concepts/results but also do not distract or confuse.

[**2 pt - 0 pt**] Communication aids are poorly designed and do not help clarify concepts. The diagrams/figures are not visually pleasant, contain too much or too little information or present information in confusing ways.

5.7 Overall Document Quality (total 10 pt)

[**8 pt full marks**] The document is formatted according to the template provided. The font on the document is readable, uses margins that facilitate reading, information is represented and organised to maximize comprehension. The document contains no orthographic or grammatical errors. Citation style is used consistently throughout the document

[**6 pt - 2 pt**] The document follows the required template but is not well-organised or is hard to read. The document contains a few orthographic or grammatical errors.

[**2pt - 0 pt**] The document is not formatted correctly, font size is too small to read. Multiple orthographic and grammatical errors. Citation style is not followed consistently.