School of Computing & Information Systems
The University of Melbourne

COMP90049 Introduction to Machine Learning
Semester 2, 2023

Assignment 3[1]

Sentiment Analysis of Patient Satisfaction Comments

Mike Conway

10th August 2023 [Version: 1.19]

Released: Friday 8th September (end of Week 7)

Due: Stage I (main assignment deadline): Friday 6th October at 5pm (end of Week 10)
Stage II (peer review deadline): Wednesday 11th October at 5pm

Marks: This project will be marked out of 30 and consist of 30% of your total marks for the subject

1 Introduction

In this assignment you will develop, evaluate, and critically assess machine learning models for identifying patient sentiment towards clinicians. You will do this using reviews derived from the `ratemds.com` United States clinician review website. You will be provided with a dataset of clinician reviews that have been labeled with a sentiment classification (i.e. whether a patient was *satisfied* or *dissatisfied* regarding an interaction with a clinician). In addition, each review is labelled with the gender of the clinician (*male*, *female*, or *unknown*). You may use this information to determine whether your model works equally well in predicting sentiment for male and female genders.[2]. This assessment provides you with an opportunity to reflect on concepts in machine learning in the context of an open-ended research problem, and to strengthen your skills in data analysis and problem solving.

Online clinician review sites (i.e. websites like `ratemds.com`) that publish patient reviews are a useful resource for individuals seeking to make informed healthcare choices, for health systems seeking to improve patient safety and healthcare quality, and for researchers seeking to identify systematic and emerging problems in healthcare provision. Given that commercial websites like `ratemds.com` and non-commercial entities like the UK National Health Service generate thousands of user feedback comments per week, in the past decade considerable effort has been expended on designing and evaluating appropriate machine learning/natural language processing tools to identify trends in patient comments. Examples of this kind of work include: Greaves et al. (2013), Doing-Harris et al. (2016), Cammel et al. (2020), Zhang et al. (2018), and Chekijian et al. (2021). Data provided for this project are derive from the following two papers: Wallace et al. (2014) and López et al. (2012)

The goal of the assignment is to ***critically assess and evaluate*** the effectiveness and appropriateness of various machine learning approaches applied to the problem of determining the sentiment (positive or negative) of patient-generated clinician reviews, and to ***articulate the knowledge that you have gained in a technical report***. The technical aspect of this project will involve applying appropriate machine learning algorithms to the data to address the task. There will be a Kaggle in-class competition where you can compare the performance of your algorithm against your classmates.

The primary output of this project will be the report, which will be formatted as a short research

[1] Note that this assignment is largely based on a format developed by Dr Lea Frermann
[2] Note that there was insufficient data and evidence to identify non-binary gender classes

paper. In the report, you will demonstrate the knowlege that you have gained over the duration of the subject in a manner that is accessible to an informed reader.

Note that you do not need to implement algorithms "from-scratch" from this assignments. It is expected that you will use algorithms implemented in existing libraries (e.g. scikit-learn). Assessment will be based on the quality of your report.

2 Deliverables

Stage I: Model development and testing and report writing (due Friday 6th October at 5pm — end of Week 10)

1. One or more programs (where "programs" can include stand-alone scripts or Jupyter Notebooks) written in Python 3, including all the necessary code to reproduce the results of your report (including model implementation, label prediction, and evaluation). You should also include a README file that briefly details your implementation. This component of the assessment should be submitted through the Canvas LMS. All your code files (and README) should be contained in a single zip file.
2. An anonymous report of approximately 2,000 words (±10%) **excluding** bibliographic references and the ethics statement (described below), but **including** material in tables and captions. Your name and student ID should **not** appear anywhere in the report, including the metadata (e.g. filename). This component of the evaluation should be submitted through Canvas. You must upload the report as a separate PDF file. Do not upload it as part of a compressed archive file (e.g. zip, tar) or in a different format (e.g. Microsoft Word). Anonymity is required in order to enhance the fairness of the peer review process (i.e. your reviewer should not know your name and you should not know your reviewer's name)
3. Predictions for the test set of clinician sentiment predictions submitted to Kaggle[3] are described in Section 7

Stage II: Peer reviews (due Wednesday 11th October at 5pm)

1. Reviews of two reports written by your classmates. These reports will be approximately 200-400 words each. This component of the assessment should be submitted through Canvas.

3 Data sets

The data in its entirety (i.e. before being divided into *training*, *validation*, and *test* sets) consists of:

- 5 columns (see **Table 1** for a detailed description of the column names)
- 54,107 rows (i.e. individual comments on clinicians by patients with associated binary sentiment labels of "-1" [*negative sentiment*] and "1" [*positive sentiment*])
- 19,097 distinct clinicians are reviewed in the dataset (i.e. many clinicians are the target of more than one patient review). For this project, the primary task is comment-based classification (i.e. classifying individual comments for sentiment), but you may experiment with clinician-level classification, too.
- 72% of the comments are labelled as *positive sentiment* (38,847) and 28% are labelled as *negative sentiment* (15,170)

You will be provided with:

1. A *training set* of 43,003 clinician reviews [`TRAIN.csv`]
2. A *development (or validation) set* of 5,500 clinician [`VALIDATION.csv`] reviews

[3] `https://www.kaggle.com`

Col #	Name	Description
1	`index`	Row index value [0 to 54,106]
2	`dr-id-adjusted`	Clinician ID. Note that there are 19,097 distinct clinicians in the dataset
3	`dr_id_gender`	Provides clinician ID level gender information. **This should be the gender demographic label you use in your project** [0: female; 1: male; 2: unknown]
4	`review-text-cleaned`	Main comment field (e.g. *Very professional and concerned with your health*, and *worst doctor ever. Awful!*)
5	`rating`	Target sentiment label. **This is the target label** [-1: negative sentiment; 1: positive sentiment]. **This is a binary label**

Table 1: Data description

3. A *test set* of 5,514 with no target (sentiment) labels which will be used for final evaluation of the Kaggle in-class competition [`TEST_NO_LABELS.csv`]

Note that your classified test corpus will be submitted to Kaggle for evaluation. In your report for this assignment, you should present the results of your classifier on the provided validation corpus.

The format of the data is described in **Table 1**.

3.1 Target Label

This is the label that your model should predict (y). The label refers to the sentiment of the clinician review, with a value of "-1" being negative (i.e. *negative sentiment*) and "1" being positive (i.e. *positive sentiment*). Despite the fact that it is represented as a number, **this is a categorical variable**. Each row has an associated label. This is the primary classification task that will be evaluated as part of the Kaggle contest.

3.2 Gender Labels

Gender labels (`dr_id_gender`) in the associated CSV files) provide information regarding the gender of the clinician. They should only be used to evaluate models on specific subgroups of employees (e.g. train a model on the entire training set, then test it on male gender in the test set), but not be predicted or used as a feature. In the provided dataset, each row is labelled with one of three possible gender labels:

- 0: female
- 1: male
- 2: unknown

3.3 Features

To aid in your initial experiments, we have created different feature representations based on the review data (raw text, TFIDF, word embeddings). You may use any subset of the provided representations described below in your experiments as well as any feature representations you provide yourself. Note that each row in the feature set files provided maps to a corresponding row in the "raw text" file.

3.3.1 Raw text

We have provided the raw text of the comment in the field `review-text-cleaned`. For example:

- *Was not very pleased at all. Had to pay to get a copy of my labwork to take to another doctor I could feel more confident with. Nice lady but not someone I'd see for anything other than a very minor issue if no one else was available*

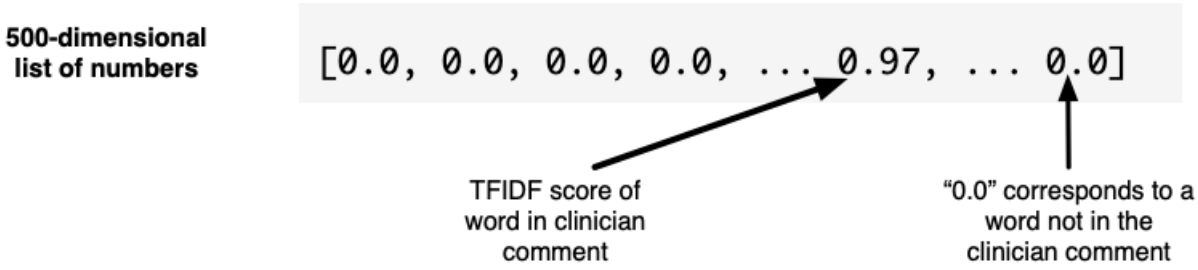
- *Very caring, always hopeful, currently treating us. We very happy with his knowledge and success with our leukemia treatment*

You can see that we have not pre-processed these texts to remove punctuation, correct spelling, or normalize case. It is up to you to determine the extent to which pre-processing the text will affect classification performance.

Raw text is provided as the field "review-text-cleaned" in the training, validation, and test CSV files. The target label for classification is the field "rating" with two values **-1** (negative sentiment) and **1** (positive sentiment). All files are provided in a data directory available from Canvas.

3.3.2 TFIDF

We applied Term Frequency - Inverse Document Frequency (TFIDF) to generate a feature set that you can use in your report. Specifically, we (1) removed all stopwords; (2) only retained the 500 words with the highest TFIDF values. As a result, each comment is now represented as a 500 dimensional feature vector, each dimension corresponds to one of 500 words in the review text. Note that most values will be 0.0 as most of the reviews provided are short. For example:



The feature selection and associated code in Week 6 provides more information on TFIDF. For a full description, please see Manning et al. (2008).

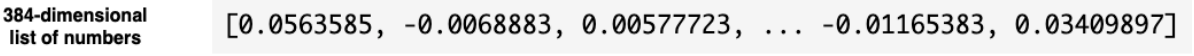
The file `tfidf_words.txt` contains the 500 words with the highest TFIDF value, as well as their index in the vector representation. You may use this information in model/error analysis.

TFIDF representations are provided in the files:

- `TFIDF_TRAIN.csv`
- `TFIDF_VALIDATION.csv`
- `TFIDF_TRAIN.csv`

3.3.3 Word Embeddings

Each comment has been mapped to a 384-dimensional embedding computed with a pre-trained language model. These are intended to capture the semantics of each comment in order that similar comments will be located closely together in the 384-dimensional space. Embeddings were generated with Sentence Transformer[4] embeddings (Reimers and Gurevych, 2019). For example:



Word embeddings are provided in the files:

- `384EMBEDDINGS_TRAIN.csv`
- `384EMBEDDINGS_VALIDATION.csv`
- `384EMBEDDINGS_TEST.csv`

[4] `https://www.sbert.net/`

4 Project Stage I

You should formulate a research question (two example research questions are provided below), and develop machine learning algorithms and appropriate evaluation strategies to address the research question.

You should minimally implement and analyse in your report **one baseline**, and **at least two different machine learning models**. Note that we are 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.

4.1 Research Question

You should address **one** research question in your project. We propose two research question below for inspiration, but you are not constrained to these specific research questions. However, addressing more than one research question in your report will not lead to higher marks. We are more interested in your critical analysis of methods and results than covering more content or material.

4.1.1 Example Research Questions

Q: Does the classifier demonstrate gender bias?

First, identify models and feature sets that perform well when applied to *all* the test data. Then, compare different models and feature representations when applied to male or female genders in the dataset. Do the models that yield the best performance when applied to *all* the data also yield the best performance on different demographic groups in the test data (e.g. the two genders identified in the test data)? Attempt to account for any differences that you might observe in terms of the concepts covered in this subject. Note that it is your **analysis** that is of most importance for this assignment, rather than the performance of your trained models.

Q: Does feature engineering improve classifier performance

How (and why) does using different feature sets affect classifier performance? You are free to use your own feature representations to augment (or replace) the feature representations provided as part of the subject (i.e. TFIDF and word embeddings). Examples of possible areas of experimentation include the use of features derived from sentiment dictionaries, using some of the feature selection methods covered in this subject (i.e. different feature selection methods and thresholds), and stemming or lemmatising the raw text.

4.2 Unsupervised Learning

It is *optional* for you to utilize unsupervised learning approaches (e.g. k-means clustering) to investigate your chosen research question. Examples of possible areas of experimentation include the use of automatically derived clusters to help identify systematic differences between comments related to different genders that may affect classification accuracy. Use of unsupervised learning approaches may provide insights for the discussion and analysis section of your report.

4.3 Evaluation

The objectives of your models will be to predict the labels of unseen data. We will use a **holdout strategy**. The data collection has been split into three parts: a training set, a validation set, and a test set. This data is available in Canvas.

To give you the opportunity of testing the performance of your trained models against other students in the subject, we will be setting up a Kaggle in-class competition. You can submit results on the the test set there, and get immediate feedback on your model's performance. There is a Leaderboard, that will allow you to see how well you are doing compared to other students participating in the subject.

4.4 Report

You will submit an anonymised report of 2,000 words in length (±10%), **excluding** your bibliographical reference list and ethics statement, but **including** text in tables and captions. The report should follow the structure of a short research paper, as discussed in the guest lecture on Academic Writing. It should describe your approach and observations in the context of your chosen research question, both in engineering (optional) features, and the machine learning algorithms you tried. The main aim of the report is to provide the reader with knowledge about the problem, in particular, critical analysis of your results and discoveries. The internal structure of well-known machine learning models should only be discussed if it is important for connecting the theory to your practical observations.

- **Introduction:** Provide a short description of the problem and data set, and the research question addressed
- **Literature review:** Generate a short summary of some related literature, including the data set reference and at least two additional relevant research papers of your choice. You might find inspiration in this document's bibliography. You are encouraged to search for other references, for example among the articles cited within the papers referenced in this document.
- **Method:** Identify the newly engineered feature(s), and the rationale behind including them (optional). Explain the machine learning models and evaluation metric(s) you have used (and why you have used them)
- **Results:** Present the results, in terms of evaluation metric(s) and, ideally, illustrative examples. Use of tables and diagrams is highly recommended
- **Discussion:** Contextualise the system's behavior, based on the understanding from the subject materials as well as in the context of the research question.
- **Conclusion:** Clearly demonstrate your identified knowledge about the problem
- **Ethics Statement:** See below for details
- **Bibliography:** List references, including references for the publications that describe the dataset (i.e. Wallace et al. (2014) and López et al. (2012)), as well as references to any other related work you used in your project. You are encouraged to use the APA 7 citation style, but may use different styles as long as you are consistent throughout your report.

Note that when we say "contextualize" above, we we are more interested in seeing evidence of you having thought about the task, and determined reasons for the relative performance of different methods, rather than the raw scores of the different methods you select. This is not to say that you should ignore the relative performance of different models and feature sets, but rather that you should think beyond simple numbers to the reasons that underlie them.

After the conclusion, but before the reference list, please include a brief **ethics statement** (use the heading **Ethics Statement**). This statement can range from two sentences to a short paragraph and should include any ethical issues that you believe may be associated with the work (e.g. intended use, misuse potential, issues regarding redistributing data). 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

We will provide LaTeX and RTF style files that we would prefer that you use in writing the report. Reports are to be submitted in the form of a single PDF file. If a report is submitted in any format other than PDF, we reserve the right to return the report with a mark of 0.

Your report should be **anonymous**. Your name and student ID should not appear anywhere in the report, including any metadata (e.g. filename). If we find any such information, we reserve the right to return the report with a mark of 0. Anonymity is required as revealing your name is likely to compromise the peer-review process.

Given that the test set is held-out for Kaggle evaluation, in your report, please present the results of your classification experiments on the **validation** set (i.e. you should not use the validation set for training or parameter tuning). You can optionally include results on the training set and potentially explore any performance differences that occur between the test and validation sets.

5 Project Stage 2

During the reviewing process, you will read two anonymous submissions by your classmates. This is to help you contemplate different ways of approaching the project and to ensure that every student receives additional feedback. You should aim to write 150-300 words total per review, responding to three prompts:

- Briefly summarise what the author has done in one paragraph (50-100 words)
- Indicate what you think that the author has done well, and why in one paragraph (50-100 words)
- Indicate what you think could have been improved, and why in one paragraph (50-100 words)

6 Assessment Criteria

The Project will be marked out of 30, and is worth 30% of your overall mark for the subject. The mark breakdown will be:

- **Report Quality: (26/30 marks).** You can consult the marking rubric on the Canvas/Assignment 3 page which indicates in detailed categories what we will be looking for in the report.
- **Kaggle: (2/30 marks).** For submitting (at least) one set of model predictions to the Kaggle competition.
- **Reviews: (2/30 marks).** You will write a review for each of two reports written by other students; you will follow the guidelines stated above.

7 Using Kaggle

The Kaggle competition will be on predicting sentiment. To participate in the Kaggle competition:

- Each student should create a Kaggle account (unless they have one already) using your Student-ID
- You may make up to 8 submissions per day. An example submission file can be found on the Kaggle site.
- Submissions will be evaluated by Kaggle for accuracy, against just 30% of the test data, forming the public leaderboard.
- Prior to the closing of the competition, you may select a final submission out of the ones submitted previously – by default the submission with highest public leaderboard score is selected

by Kaggle.

- After the competition closes, public 30% test scores will be replaced with the private leaderboard 100% test scores.

8 Assignment Policies

8.1 Introduction

The dataset used in this assignment is originally derived from qualitative work conducted by López et al. (2012)):

> Lopez, A., Detz, A., Ratanawongsa, N., and Sarkar, U. (2012). What patients say about their doctors online: a qualitative content analysis. J Gen Intern Med, 27(6):685–92.

And further developed into a dataset suitable for computational analysis by Wallace et al. (2014):

> Wallace, B. C., Paul, M. J., Sarkar, U., Trikalinos, T. A., and Dredze, M. (2014). A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. J Am Med Inform Assoc, 21(6):1098–103.

These two references ***must*** **be cited in the bibliography**. The dataset is not provided with an explicit terms of use, but good academic practice means that any report that uses data derived from these publications should cite them.

Please note that the dataset is a sample of actual data posted to the World Wide Web by real people. As such, it may contain content that a reasonable person could construe as in poor taste or offensive. We would ask you, as much as possible, to look beyond this to focus on the task at hand. If you object to these terms, please contact us (`mike.conway@unimelb.edu.au`) as soon as possible.

8.2 Changes/Updates to the Project Specifications

We will use Canvas announcements for any large-scale changes (hopefully this will not be necessary) and Ed for small clarifications. Any addendums made to the Project specifications via the Canvas LMS system will supersede information contained in this version of the specifications.

8.3 Late Submission Policy

There will be no late submissions allowed to ensure a smooth peer review process. Submission will close Friday 6th October at 5pm 2023. For students who are demonstrably unable to submit a full solution in time, we may offer an extension, but note that you may be unable to participate in and benefit from the peer review process in that case. A solution will be sought on a case-by-case basis. Please email Hasti Samadi (`hasti.samadi@unimelb.edu.au`) with documentation of the reasons for the delay.

8.4 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 in algorithm choice and development will be considered cheating. We highly recommend to (re)take the academic honesty training module in this subject's Canvas site. We 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 your own work, and submitting such content will be treated as a case of academic misconduct, in line with the University's academic integrity policy and specific recent guidance on the use of ChatGPT and other Large Language Models in student work.

9 Final Comments & Reminders

Some final comments for you to bear in mind when preparing your assignment:

- Note that it is the **analysis** contained in your project that is of most importance for the assessment of this assignment, rather than the performance of your trained models. We want to see evidence of you having thought about the task, and have determined reasons for the relative performance of different methods, rather than just the raw scores of the different methods you select
- Use at least one baseline algorithm/feature representation in addition to two other classifiers that you hypothesise will perform better on the task
- Please make use of existing Python machine learning libraries for this project. It is not necessary to implement algorithms "from scratch" and doing so will not earn you additional marks
- In your report, you will report the performance of your validation set. The test set will be held out for final evaluation on Kaggle.

References

Cammel, S. A., De Vos, M. S., van Soest, D., Hettne, K. M., Boer, F., Steyerberg, E. W., and Boosman, H. (2020). How to automatically turn patient experience free-text responses into actionable insights: a natural language programming (NLP) approach. *BMC Med Inform Decis Mak*, 20(1):97.

Chekijian, S., Li, H., and Fodeh, S. (2021). Emergency care and the patient experience: Using sentiment analysis and topic modeling to understand the impact of the covid-19 pandemic. *Health Technol (Berl)*, 11(5):1073–1082.

Doing-Harris, K., Mowery, D. L., Daniels, C., Chapman, W. W., and Conway, M. (2016). Understanding patient satisfaction with received healthcare services: A natural language processing approach. *AMIA Annu Symp Proc*, 2016:524–533.

Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., and Donaldson, L. (2013). Use of sentiment analysis for capturing patient experience from free-text comments posted online. *J Med Internet Res*, 15(11):e239.

López, A., Detz, A., Ratanawongsa, N., and Sarkar, U. (2012). What patients say about their doctors online: a qualitative content analysis. *J Gen Intern Med*, 27(6):685–92.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Wallace, B. C., Paul, M. J., Sarkar, U., Trikalinos, T. A., and Dredze, M. (2014). A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. *J Am Med Inform Assoc*, 21(6):1098–103.

Zhang, W., Deng, Z., Hong, Z., Evans, R., Ma, J., and Zhang, H. (2018). Unhappy patients are not alike: Content analysis of the negative comments from China's good doctor website. *J Med Internet Res*, 20(1):e35.