School of Computing and Information Systems
The University of Melbourne
COMP30027, Machine Learning, 2020

Project 2: this review sounds positive!

Task:	Build a classifier to predict star ratings for reviews
Due:	**Stage I: Friday 29 May, 11am** UTC+10 (Australian Eastern Standard Time)
	Stage II: Friday 5 June, 11am UTC+10 (Australian Eastern Standard Time)
Submission:	Stage I: Report (PDF) and code to Canvas; test output(s) to Kaggle in-class competition
	Stage II: Peer-Reviews and Reflection to Canvas
Marks:	The Project will be marked out of 20, and will contribute 20% of your total mark.
Groups:	Groups of 1 or 2, with commensurate expectations for each (see below).

1 Overview

The goal of this Project is to build and critically analyse some supervised Machine Learning methods, to predict star ratings (1, 3, or 5) for reviews on restaurants. This problem is a form of sentiment analysis, which is the problem of how to automatically identify and extract polarity (e.g. positive, negative, or neutral) from text[1]. Sentiment analysis has a variety of applications in recommender systems, marketing, economics, and social and political sciences. Although this problem has been well-studied, a general solution remains elusive.

This assignment aims to reinforce the largely theoretical lecture concepts surrounding data representation, classifier construction, and evaluation, by applying them to an open-ended problem. You will also have an opportunity to practice your general problem-solving skills, written communication skills, and creativity.

This project has two stages. The main focus of these stages will be the written report, where you will demonstrate the knowledge that you have gained and the critical analysis you have conducted in a manner that is accessible to a reasonably informed reader.

2 Deliverables

More details about deliverables are given in the Submission (Section 6).

Stage I:

1. **Report**: an anonymous written report, of 1000-1500 words (for a group of one person) or 2000-2500 words (for a group of two people)
2. **Output**: the output(s) of your classifier(s), comprising predictions of labels for the test instances, submitted to the Kaggle[2] in-class competition described below.
3. **Code**: one or more programs, written in Python , which implement machine learning methods that build the model, make predictions, and evaluate the results.

Stage II:

1. **Peer-review**: reviews of two reports written by other students, of 200-400 words each
2. **Reflection**: a written reflection piece of 400-600 words.

[1] `https://en.wikipedia.org/wiki/Sentiment_analysis`
[2] `https://www.kaggle.com/`

3 Terms of Use

The data has kindly been provided to us, under the provision that any resulting work should cite these two resources:

> What Yelp fake review filter might be doing? A. Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance, *ICWSM*, 2013.
>
> Collective Opinion Spam Detection: Bridging Review Networks and Metadata. Shebuti Rayana, Leman Akoglu, *ACM SIGKDD*, Sydney, Australia, August 10-13, 2015

These references must be cited in the bibliography. We reserve the right mark of any submission lacking these references with a 0, due to violation of the Terms of Use.

Please note that the dataset is a sample of actual data posted to the World Wide Web. As such, it may contain information that is in poor taste, or that could be construed as offensive. We would ask you, as much as possible, to look beyond this to the task at hand. For example, it is generally not necessary to read individual reviews.

The opinions expressed within the data are those of the anonymised authors, and in no way express the official views of the University of Melbourne or any of its employees; using the data in an educative capacity does not constitute endorsement of the content contained therein.

If you object to these terms, please contact us (`kris.ehinger@unimelb.edu.au` or `ling.luo@unimelb.edu.au`) as soon as possible.

4 Data

The data files are available via the Canvas, and are described in a corresponding README.

The reviews are collected from Yelp[3], which is a platform that allows the public to publish reviews of businesses. In our dataset, each review contains:

- meta features: `date`, unique `review ID`, `reviewer ID`, `business ID`, votes cast by Yelp users (whether they think the review is `funny`, `cool` or `useful`)
- the original `review text`, provided as a single file with rows corresponding to the meta file
- text features: produced by various text encoding methods for `review text`
- class label: `rating` (3 possible levels, 1, 3 or 5)

You will be provided with training set and a test set. The training set contains the meta features, text features of reviews, and the rating, which is the "class label" of our task. The test set only contains the meta and text features without the rating.

The files provided are:

- *review_meta_train.csv* (rating label in this file)
- *review_text_train.csv*: original review text.
- *review_meta_test.csv*
- *review_text_test.csv*: original review text.
- *review_text_features_*.zip*: preprocessed text features for training and test sets, 1 zipped file for each text encoding method. Details about using these text features are provided in README.

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

5 Task

You are expected to develop Machine Learning system(s) to predict the star rating of a review based on its meta features (e.g. reviewer ID, business ID etc.) and/or review text. You will implement and compare different machine learning models and explore the effective features for this sentiment prediction task.

- **The training-evaluation phase**: The holdout or cross-validation approaches can be applied on the training data provided.
- **The test phase**: the trained classifiers will be evaluated on the unlabelled test data. The predicted labels of test reviews should be submitted as part of the Stage I deliverable.

Various machine learning techniques have been (or will be) discussed in this subject (0R, Naive Bayes, Decision Trees, kNN, SVM, neural netowrk, etc.); many more exist. You may use any machine learning method you consider suitable for this problem. *You are strongly encouraged to make use of machine learning software and/or existing libraries (such as* `sklearn`*) in your attempts at this project.*

In addition to different learning algorithms, there are many different ways to encode text for these algorithms. The files in *review_text_features_*.zip* are some possible representations of the content of the review text we have provided. For example, one of the encoding method is `CountVectorizer` in `sklearn`, which converts text documents into "Bag of Words" – the documents are described by word occurrences while ignoring the relative position information of the words. You can use these representations to develop your classifiers, but you should also feel free to extract your own features from the raw review text, according to your needs. Just keep in mind that any data representation you use for the text in the training set will need to be able to generalise to the test set.

6 Submission

The report, code, peer-review and reflections should be submitted via the Canvas; the predictions on test data should be submitted to Kaggle.

Stage I submissions will open one week before the due date. Stage II submissions will be open as soon as the reports are available (24 hours following the Stage I submission deadline).

6.1 Individual vs. Two-Person Participation

You have the option of participating as a "group" of one individual, or in a group of two. In the case that you opt to participate individually, you will be required to enter at least 1 and up to 4 distinct systems, while groups of two will be required to enter **at least** 3 and up to 4 distinct systems, of which *one is to be an ensemble system (stacking) based on the other systems*. The report length requirement also differs, as detailed below:

Group size	Distinct system submissions required	Report length
1	1–4	1,000–1,500 words
2	3–4	2,000–2,500 words

If you wish to form a two-person group, **only one** of the members need to register on Canvas by Friday 15 May, via the survey "**Assignment 2 Group Registration**".

Note that once you have signed up for a given group, you will not be allowed to change groups. If you do not register before the deadline above, we will assume that you will be completing the assignment as an individual, even if you were in a two-person group for Assignment 1.

6.2 Stage I: Report

The report should be 1,000-1,500 words (groups of one person) or 2,000-2,500 words (groups of two people) in length and provide a basic description of:

1. the task, and a short summary of some related literature
2. what you have done, including any learners that you have used, or features that you have engineered. *This should be at a conceptual level; a detailed description of the code is not appropriate for the report. The description should be similar to what you would see in a machine learning conference paper.*
3. evaluation of your classifier(s).

You should also aim to have a more detailed discussion, which:

4. Contextualises the behaviour of the method(s), in terms of the theoretical properties we have identified in the lectures
5. Attempts some error analysis of the method(s)

And don't forget:

6. A bibliography, which includes Mukherjee et al. (2013), Rayana and Akoglu (2015) listed in **Terms of Use**, and other related work

Note that we are more interested in seeing evidence that you have thought about the task and investigated the reasons for the relative performance of different methods, rather than in the raw scores of the different methods. This is not to say that you should ignore the relative performance of different runs over the data, but rather that you should think beyond simple numbers to the reasons that underlie them, and connect these to the theory that we have discussed in this class.

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.

To facilitate anonymous peer-review, your name and student ID **should not appear** anywhere in the report, including the metadata (filename, etc.).

6.3 Stage I: Predictions of test data

To give you the possibility of evaluating your models on the test set, we will be setting up this project on Kaggle InClass competition. You can submit results on 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 as compared to other classmates participating on-line. The Kaggle in-class competition URL will be announced on Canvas shortly.

You will receive marks for submitting (at least) one set of predictions for the unlabelled test dataset into the competition; and get basically reasonable accuracy, e.g. better than Zero-R. The focus of this assignment is on the quality of your critical analysis and your report, rather than the performance of your Machine Learning models.

6.4 Stage II: Reviews

During the reviewing process, you will read two submissions by other students. This is to help you contemplate some other ways of approaching the project, and to ensure that students get some extra feedback. For each report, you should aim to write 200-400 words total, responding to three "questions":

- Briefly summarise what the author has done
- Indicate what you think that the author has done well, and why
- Indicate what you think could have been improved, and why

Please be courteous and professional in the reviewing process. A brief guideline for reviewers published by IEEE can be found `https://www.ieee.org/content/dam/ieee-org/ieee/web/org/members/students/reviewer_guidelines_final.pdf`

6.5 Stage II: Reflections

A comprehensive written reflection piece summarising your critical reflection on this project within 400-600 words. This report is not anonymous.

7 Assessment Criteria

The Project will be marked out of 20, and is worth 20% of your overall mark for the subject. The mark breakdown will be:

Report	12 marks
Performance of classifier	2 mark
Reviews	4 marks
Reflection	2 marks
TOTAL	20 marks

The report, reviews and reflection will be marked according to the rubric, which will be announced via the Canvas.

The performance of classifier (2 marks) is for submitting (at least) one set of model predictions to the Kaggle competition; and get basically reasonable accuracy, e.g. better than Zero-R.

Since all of the documents exist on the World Wide Web, it is inconvenient but possible to "cheat" and identify some of the ratings from the test data using non-Machine Learning methods. If there is any evidence of this, the performance of classifier will be ignored, and you will instead receive a mark of 0 for this component. The code will not be directly assessed, but if you do not submit it, it will be assumed that you are attempting to circumvent the Machine Learning requirement, and you will receive a mark of 0 for the performance of classifier.

8 Using Kaggle

The Kaggle in-class competition URL will be announced on Canvas shortly. To participate do the following:

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

9 Changes/Updates to the Assignment Specifications

We will use the Canvas to advertise any (hopefully small-scale) changes or clarifications in the assignment specifications. Any addendums made to the assignment specifications via the Canvas will supersede information contained in this version of the specifications.

10 Late Submissions

Late submissions will bring disruption to the reviewing process. You are strongly encouraged to submit by the date and time specified above. If circumstances do not permit this, then the marks will be adjusted as follows:

- Each day (or part of the day) that the report is submitted after the specified due date and time for Stage I, 10% will be deducted from the marks available, up until 7 days (1 week) has passed, after which regular submissions will no longer be accepted. A late report submission will mean that your report might not participate in the reviewing process, and so you will probably receive less feedback.
- Any late submission of the reviews will incur a 50% penalty (i.e. 2 of the 4 marks), and will not be accepted more than 7 days (1 week) after the Stage II deadline.
- Any late submission of the reflection will incur a 50% penalty (i.e. 1 of the 2 marks), and will not be accepted more than 7 days (1 week) after the Stage II deadline.

11 Academic Honesty

While it is acceptable to discuss the assignment with others in general terms, excessive collaboration with students outside of your group is considered cheating. We will be vetting system submissions for originality and will invoke the University's Academic Misconduct policy (`http://academichonesty.unimelb.edu.au/policy.html`) where either inappropriate levels of collaboration or plagiarism are deemed to have taken place.