Assignment / FCP

The final course project (FCP) consists of a text classification task: you're supposed to adjudicate between "okay", "good", and "excellent" reviews'ratings from the BeerAdvocate community/website.

For the sake of redundancy, your goal is to classify each beer review "i" into one of the three categories:

0/okay: rating is in [3.5 - 4)

1/good: rating is in [4 - 4.5)

2/excellent: rating is in [4.5 - 5]

Data and file description

The training data (train.csv, see the download area below) contains 21,057 labelled reviews. The test data contains 8,943 reviews that are unlabeled (test.csv, see below).

File description:

- **train.csv:** your training data (N = 21,057). The first column, ''label", is the label of each training beer review; the second column, "text", contains the text of the beer review.
- **test.csv: y**our test data (N = 8,943). The first column, "_id", is a beer review-level identifier; the second column, "text", is the unlabelled beer review to

classify.

- **submission.csv:** this is the file you're asked to populate and submit. The first column, "_id", is the review-level identifier reported in test.csv; the second column, ''pred_label", will contain your predicted class label (0, 1, or 2) for the corresponding line in text data. Note: i) please include the exact headers "_id, pred_label" in your submission — or just used the file I make available in the download area below; ii) stick with the beer identifiers included in the column "_id" of test.csv.

Download the files ⬇️ 🗂️

_2023_fcp

Shared with Dropbox

https://www.dropbox.com/scl/fo/dl1k4ljpbb5n5d1x58brz/h?dl=0&rlkey=ocmk9t1hxnqs7fhxldo6l2p2x

Guidelines

To carry out the classification task, you can use any classifier or combination of classifiers, any combination or selection of features, and either supervised, semi-supervised, or even transfer learning approaches. For example, the features may include (but are not limited to) BoW vectors, TFIDF vectors, vectors achieved with embedding algorithms, topic-to-document probabilities. In terms of estimators, you can use logistic regression, Naive Bayes Classifier, or any other estimator you think is appropriate given the nature of the task.

Submission

When / deadline

Submit your FCP via Moodle by Wednesday 21st July 2023 (4:00 PM)

What / submission package

The submission should comprise:

- **submssion.csv:** the first column contains beer review-identifiers, "_id", as per test.csv (e.g., "11434"); the second column, "pred_label", is the category (or label) predicted by your classifier, one of {0/okay, 1/good, 2/excellent}

- **a 2,000-word companion report.** This document should report:
  - the various logical steps behind your classifier
  - the justification of each step — e.g., why you use a certain estimator rather than others
  - how the classifier could be improved — this is not necessary if your F1 Score = 1. 😎🆒
- **the companion Python, Julia, or R code:** I leave the choice of one of these three languages with you. Irrespective of the language you go with, the code must be reproducible.

Assessment

Your mark is calculated as follows:

FCP mark = 0.5 X solution effectiveness + 0.5 X clarity and organization of materials

By **solution effectiveness**, I mean a very simple thing: the F1 Score of your classifier, that is, the harmonic mean of precision and recall. Don't try to overfit your model. If you do that, I'll spot the problem and your submission will go straight to the ethics committee. 🚔👮‍♀️👮‍♂️

Regarding the 'clarity and organization of materials' dimension, I value it positively:

- clear and simple writing, e.g., short sentences
- concise writing, i.e., added value per word
- the use of academic references supporting the design choice of your classifier
- the use of compelling logical arguments supporting the design choice of your classifier
- evidence of critical thinking at large
- cross-references connecting the computer code with the companion report
- formatted computer code
- commented computer code