Grade Rubric

Final Project: Text Classification

Parts of the report submission:

Step 1: Text Processing
For your choice of dataset, you will first process the text, tokenize it and choose whether to do further pre-processing or filtering. If you do some pre-processing or filtering, then using the text with and without it can be one of your experiments.

Step 2: Feature Engineering
Produce the features in the notation of the NLTK, writing feature functions in Python as done in the Labs and starting with the "bag-of-words" features. Use the NLTK Naïve Bayes classifier to train and test a classifier on feature sets. Also use cross-validation to obtain precision, recall and F-measure scores.

Step 3: Experiments
A. For a base level completion of experiments, carry out at least several experiments using two different sets of features and comparing the results. One of the feature functions must combine features, e.g. combining those used in labs. Another of the feature functions must be a new one not demonstrated in lab. This may be one of the ones suggested in the advanced list.

B. Choose an additional, more advanced type of task from this list, or propose your own

The report and submission will be graded according to the following rubric:

	Level of Achievement		
	1 (50%)	**2 (75%)**	**3 (100%)**
reproducibility 20%	The code submitted does not reproduce results reported in the Final Project document. The development process is not clear with critical code for processing missing. The code does not show output for features beyond the baseline. Cross validation and evaluation measures code is absent or fails to report scores.	The code submitted partially reproduces the results reported in the Final Project document. The development process is clear in general with some details missing. The output shows results for one feature only over the baseline. Cross validation code works properly. Evaluation measures outputs are offered but not for all measures.	The code submitted reproduce exactly the same results discussed in the Final Project report. Development process is clear. The output shows results for two or more features on top or the baseline. Cross validation code is optimized for this task. Evaluation measures outputs are offered in terms of accuracy, precision, recall, F1, and confusion matrices.
correct analysis 50%	Text processing steps are absent or unclear. No features are offered beyond baseline. Cross validation is not performed or done at a minimal level. Evaluation	Text processing is incomplete or irrelevant to the task at hand Features are included but they relevancy to the task is unclear and/or they are prone to overfitting or underfitting.	Each step or technique performed for data pre-processing is described with enough detail. Feature production is discussed with details of reasons for

	measures are absent or incomplete (i.e. offering accuracy only)	Cross- validation decisions are explained. Evaluation measures are offered but not discussed; or overemphasis on accuracy undermines precision, recall and F1 results.	choosing particular features. Cross- validation decisions are explained and justified in terms of reliability. Evaluation measures are offered with emphasis on precision, recall and F1
writing clarity and convincing conclusions 30%	Conclusions are absent or minimal. No explanation is offered regarding decision making processes for each step.	Conclusions are limited to output description without contextualizing the data as either movie reviews, spam detection, or sentiment analysis in Twitter. Conclusions also overemphasize accuracy scores over analyzing precision, recall or F1.	The conclusions show a clear understanding of feature engineering, evaluation measures (with emphasis on precision, recall, and F1 versus accuracy), and the effect of cross-validation in reliability of results. Conclusions are also contextualized in the type of subject or business the date comes from (i.e. movie reviews, spam detection, or sentiment in social media data)

Interpretation of numeric grades as letter grades:

90 – 100 A
85 – 89.9 A-
80 – 84.9 B+
75 – 79.9 B
70 – 74.9 B-

below 70 has similar interpretation in the C and lower range.

Late assignment submissions will be accepted, but will be penalized:

1 Week late- 10 points taken off (2/3 letter grade)

2 Weeks or more late- 20 points taken off (1 1/3 letter grade)

Late assignments may possibly be excused by emailing the instructor with the appropriate excuse.