Jasmine Chen, Yi Tang, Chenxi Yang

February 6, 2020

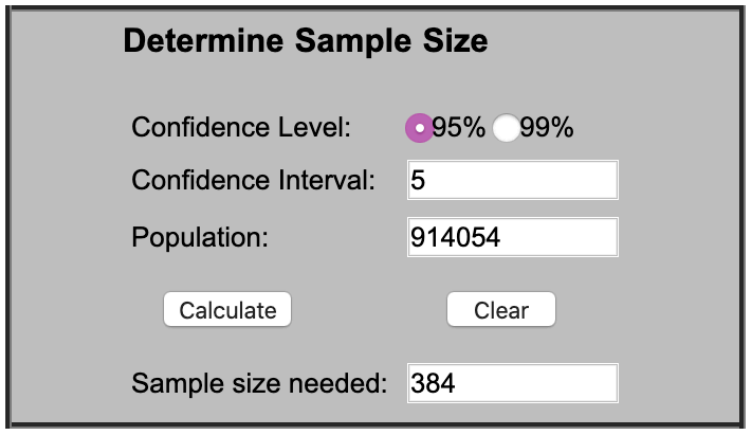
Data 553 Report

Pre-processing data

Firstly, for original data, we do some basic steps, including remove the non-ASCII characters and set all characters to be lower case so that they can be successfully determined by nltk library. Then, we construct a data frame named data_total to store our dataset. Next, adding columns comment, rating, reviewer, id, fee, title, reviewId, dataSource, appId, date, length_words from the original dataset or set to be None. Then, making functions about calculating the total sentiScore, converting from VBD to VBN or from VBP to VBZ, removing stop-words, stemming characters, lemmatizing comments. Lastly, we convert the data frame to json file.

Sampling and labelling

In the original dataset, there are 914054 rows in total. In order to determine the sample size, we use the sample size calculator on the website https://www.surveysystem.com/sscalce.htm. With a confidence level of 95%, confidence interval of 5 and population of 914054, the sample size needed should be 384 rows.



According to the essay, On the automatic classification of app reviews, we can classify app reviews into four types: bug reports, feature requests, user experiences, and text ratings, based on their definition. Bug reports describes problems with the app which should be corrected, such as a crash, an erroneous behavior, or a performance issue. Feature requests means that users ask for missing functionality or missing content and share ideas on how to improve the app in future releases by adding or changing features. User experiences combine ‘‘helpfulness’’ and ‘‘feature information’’ content. Ratings only include praise, dispraise, a distractive critique, or a dissuasion.

The four types are respectively recorded as 1, 2, 3 and 4. At first, Yi and Chenxi label the data individually. Then, Jasmine double checked it. Last, combine the labels of three people to get the final decision. Noticeably, there are 3 comments that are meaningless, so we decide to delete them. Also, there are 5 comments that are not reaching the consensus, so we decide to ignore them as well. Therefore, there are 376 remaining in the dataset. Our final manually classified labels are shown below.

```
Bug 126
Feature 60
UserExperience 140
Rating 336
```

And the testing set is:

label	
Bug	63
Feature	30
Rating	213
UserExperience	70

Result of part k: the precision, recall, and F1-score

Based on the predicted labels, we compute the precision, recall, and F1-score of the sample dataset and reproduce the results in the following tables.

Sample dataset:

Classification techniques	Bug reports			Feature requests			User experiences			Ratings		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Document classification (&NLP)												
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
bow	0.78	0.22	0.35	0.8	0.13	0.23	0.83	0.29	0.43	0.56	0.94	0.7
bigram	0.55	0.78	0.64	0.59	0.77	0.67	0.52	0.89	0.66	0.77	0.52	0.62
bow-bigram	0.72	0.33	0.46	0.89	0.27	0.41	0.79	0.47	0.59	0.61	0.9	0.73
bow-lemmatize	0.79	0.24	0.37	0.8	0.13	0.23	0.84	0.3	0.44	0.56	0.96	0.71
bow-remove_stopwords	0.81	0.21	0.33	1	0.1	0.18	0.83	0.21	0.34	0.56	0.98	0.71
bow-lemmatize-remove_stopwords	0.79	0.17	0.29	0.75	0.1	0.18	0.8	0.23	0.36	0.56	0.97	0.71
bow-bigram-lemmatize-remove_stopwords	0.85	0.27	0.41	0.67	0.13	0.22	0.79	0.37	0.5	0.58	0.94	0.71
Metadata												
rating	1	0.11	0.2	0.5	0.2	0.29	0	0	0	0.63	0.96	0.76
rating-length	0.86	0.1	0.17	0.73	0.27	0.39	0.5	0.03	0.05	0.59	0.96	0.73
rating-length-tense	1	0.03	0.06	0.63	0.4	0.49	0	0	0	0.64	0.87	0.75
rating-length-tense-sentiment1	0	0	0	0.63	0.4	0.49	0	0	0	0.62	0.9	0.73
rating-length-tense-sentiment2	0	0	0	0.65	0.37	0.47	0	0	0	0.62	0.87	0.72
Combined (text and metadata)												
bow-lemmatize-rating	0.75	0.19	0.3	0.75	0.1	0.18	0.85	0.24	0.38	0.57	0.97	0.71
bigram-rating-sentiment1	0.61	0.6	0.61	0.64	0.77	0.7	0.56	0.53	0.54	0.76	0.71	0.73
bow-rating-tense-sentiment1	0.71	0.16	0.26	0.8	0.13	0.23	1	0.13	0.23	0.58	0.96	0.72
bow-rating-sentiment1	0.76	0.21	0.33	1	0.13	0.24	0.93	0.19	0.31	0.56	0.97	0.71
bigram-lemmatize-remove_stopwords-rating-tense-sentiment2	0.79	0.24	0.37	0.76	0.63	0.69	0.67	0.09	0.15	0.69	0.82	0.75
bow-bigram-tense-sentiment1	0.74	0.32	0.44	0.88	0.23	0.37	0.84	0.44	0.58	0.62	0.9	
bow-bigram-lemmatize-rating-tense	0.77	0.32	0.45	0.78	0.23	0.36	0.88	0.43	0.58	0.61	0.9	0.72
bow-bigram-remove_stopwords-rating-tense-sentiment1	0.86	0.19	0.31	0.6	0.1	0.17	0.93	0.19	0.31	0.58	0.95	0.72
bow-lemmatize-remove_stopwords-rating-tense-sentiment1	0.75	0.1	0.17	0.75	0.1	0.18	1	0.09	0.16	0.58	0.98	0.73
bow-lemmatize-remove_stopwords-rating-tense-sentiment2	0.75	0.1	0.17	0.75	0.1	0.18	1	0.03	0.06	0.58	0.97	0.72

Reproduce Data:

Classification techniques	Bug reports			Feature requests			User experiences			Ratings		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Document classification (&NLP)												
bow	0.8	0.65	0.72	0.79	0.56	0.65	0.83	0.6	0.69	0.68	0.87	0.76
bigram	0.59	0.98	0.73	0.62	0.96	0.75	0.63	0.99	0.77	0.83	0.45	0.58
bow-bigram	0.76	0.86	0.81	0.76	0.76	0.76	0.78	0.91	0.84	0.77	0.79	0.78
bow-lemmatize	0.8	0.66	0.72	0.77	0.55	0.64	0.84	0.61	0.7	0.68	0.83	0.74
bow-remove_stopwords	0.8	0.63	0.7	0.79	0.48	0.59	0.85	0.57	0.68	0.66	0.88	0.76
bow-lemmatize-remove_stopwords	0.82	0.66	0.74	0.77	0.53	0.62	0.88	0.58	0.69	0.67	0.85	0.75
bow-bigram-lemmatize-remove_stopwords	0.78	0.83	0.8	0.76	0.71	0.73	0.82	0.89	0.86	0.75	0.81	0.78
Metadata												
rating	0.63	0.83	0.71	0.43	0.45	0.4	0.77	0.89	0.82	0.73	0.35	0.47
rating-length	0.7	0.72	0.71	0.6	0.66	0.62	0.78	0.84	0.81	0.7	0.6	0.65
rating-length-tense	0.72	0.73	0.72	0.63	0.73	0.67	0.78	0.84	0.8	0.73	0.58	0.65
rating-length-tense-sentiment1	0.69	0.77	0.72	0.62	0.74	0.68	0.75	0.89	0.82	0.72	0.58	0.64
rating-length-tense-sentiment2	0.68	0.8	0.74	0.62	0.74	0.67	0.75	0.89	0.81	0.72	0.56	0.63
Combined (text and metadata)												
bow-lemmatize-rating	0.79	0.68	0.73	0.76	0.55	0.63	0.84	0.63	0.72	0.7	0.83	0.76
bigram-rating-sentiment1	0.63	0.95	0.76	0.59	0.96	0.73	0.72	0.98	0.83	0.84	0.46	0.59
bow-rating-tense-sentiment1	0.79	0.66	0.71	0.76	0.56	0.64	0.85	0.64	0.73	0.70	0.84	0.76
bow-rating-sentiment1	0.82	0.69	0.75	0.76	0.53	0.62	0.84	0.65	0.73	0.68	0.87	0.76
bigram-lemmatize-remove_stopwords-rating-tense-sentiment2	0.68	0.87	0.76	0.61	0.90	0.72	0.73	0.91	0.81	0.82	0.48	0.60
bow-bigram-tense-sentiment1	0.76	0.84	0.80	0.74	0.75	0.74	0.81	0.91	0.86	0.77	0.78	0.77
bow-bigram-lemmatize-rating-tense	0.76	0.85	0.8	0.74	0.77	0.75	0.81	0.92	0.86	0.79	0.77	0.78
bow-bigram-remove_stopwords-rating-tense-sentiment1	0.8	0.79	0.79	0.78	0.72	0.75	0.84	0.9	0.87	0.76	0.83	0.79
bow-lemmatize-remove_stopwords-rating-tense-sentiment1	0.83	0.68	0.75	0.75	0.58	0.65	0.85	0.68	0.76	0.72	0.84	0.77
bow-lemmatize-remove_stopwords-rating-tense-sentiment2	0.83	0.7	0.76	0.77	0.57	0.65	0.86	0.72	0.78	0.7	0.85	0.77

The original table in the essay for comparison:

Table 4 Accuracy of the classification techniques using Naive Bayes on app reviews from Apple and Google stores (mean values of the 10 runs, random 70:30 splits for training:evaluation sets)

Classification techniques	Bug reports			Feature requests			User experiences			Ratings		
	Precision	Recall	*F*1	Precision	Recall	*F*1	Precision	Recall	*F*1	Precision	Recall	*F*1
Basic (string matching)	0.58	0.24	0.33	0.39	0.55	0.46	0.27	0.12	0.17	0.74	0.56	0.64
Document classification (&NLP)												
Bag of words (BOW)	0.79	0.65	0.71	0.76	0.54	0.63	0.82	0.59	0.68	0.67	0.85	0.75
Bigram	0.68	**0.98**	0.80	0.68	0.97	0.80	0.70	**0.99**	0.82	0.91	0.62	0.73
BOW + bigram	0.85	0.90	0.87	0.86	0.85	0.85	0.87	0.91	0.89	0.85	0.89	0.87
BOW + lemmatization	0.88	0.74	0.80	0.86	0.65	0.74	0.90	0.67	0.77	0.73	0.91	0.81
BOW − stopwords	0.86	0.69	0.76	0.86	0.65	0.74	0.91	0.67	0.77	0.74	0.91	0.81
BOW + lemmatization − stopwords	0.85	0.71	0.77	0.87	0.67	0.76	0.91	0.67	0.77	0.75	0.90	0.82
BOW + bigrams − stopwords + lemmatization	0.85	0.91	**0.88**	0.86	0.83	**0.85**	0.89	0.94	0.91	0.85	0.90	**0.87**
Metadata												
Rating	0.64	0.82	0.72	0.31	0.35	0.31	0.74	0.89	0.81	0.72	0.34	0.46
Rating + length	0.76	0.75	0.75	0.68	0.67	0.67	0.72	0.82	0.77	0.70	0.68	0.69
Rating + length + tense	0.74	0.73	0.74	0.64	0.71	0.67	0.74	0.80	0.77	0.70	0.68	0.69
Rating + length + tense + 1× sentiment	0.69	0.76	0.72	0.66	0.66	0.66	0.71	0.85	0.77	0.71	0.66	0.68
Rating + length + tense + 2× sentiments	0.66	0.78	0.71	0.65	0.72	0.68	0.67	0.88	0.76	0.69	0.67	0.68
Combined (text and metadata)												
BOW + rating + lemmatize	0.85	0.73	0.78	**0.89**	0.64	0.74	0.90	0.67	0.77	0.73	0.89	0.80
BOW + rating + 1× sentiment	**0.89**	0.72	0.79	**0.89**	0.60	0.71	**0.92**	0.73	0.81	0.75	**0.93**	0.83
BOW + rating + tense + 1 sentiment	0.87	0.71	0.78	0.87	0.60	0.70	**0.92**	0.69	0.79	0.74	0.90	0.81
Bigram + rating + 1× sentiment	0.73	**0.98**	0.83	0.71	0.96	0.81	0.75	**0.99**	0.85	0.92	0.69	0.79
Bigram − stopwords + lemmatization + rating + tense + 2× sentiment	0.72	0.97	0.82	0.70	**0.94**	0.80	0.75	0.98	0.85	**0.92**	0.72	0.81
BOW + bigram + tense + 1× sentiment	0.87	0.88	0.87	0.85	0.83	0.83	0.88	0.94	0.91	0.83	0.87	0.85
BOW + lemmatize + bigram + rating + tense	0.88	0.88	**0.88**	0.87	0.84	**0.85**	0.89	0.94	**0.92**	0.84	0.90	**0.87**
BOW − stopwords + bigram + rating + tense + 1× sentiment	0.88	0.89	**0.88**	0.86	0.84	**0.85**	0.87	0.93	0.90	0.83	0.89	0.86
BOW − stopwords + lemmatization + rating + 1× sentiment + tense	0.88	0.71	0.79	0.87	0.64	0.74	0.91	0.72	0.80	0.73	0.90	0.80
BOW − stopwords + lemmatization + rating + 2× sentiments + tense	0.87	0.71	0.78	0.86	0.68	0.76	0.91	0.73	0.81	0.75	0.90	0.82

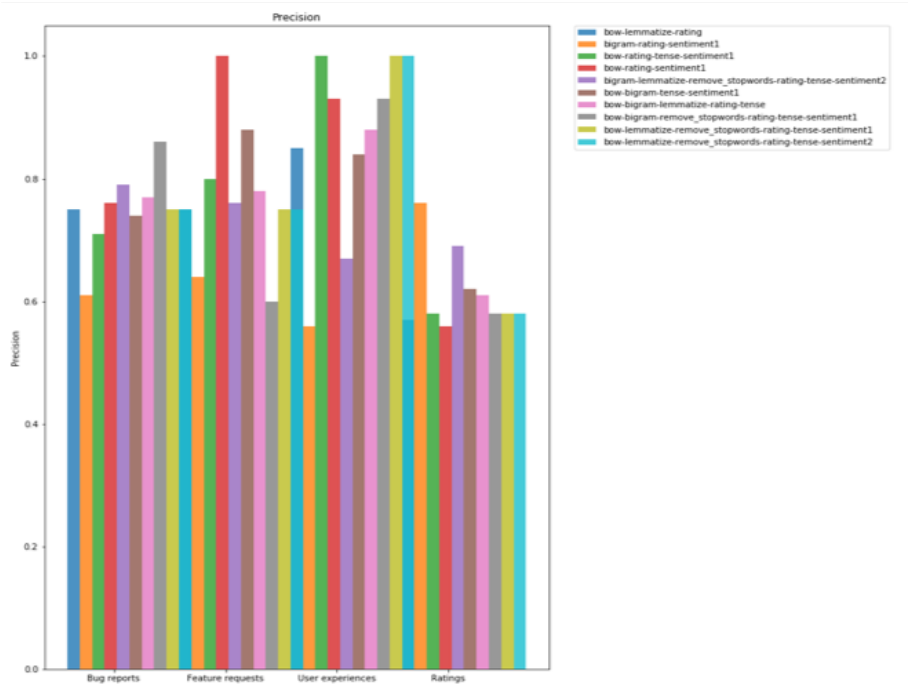
Bold values represent the highest score for the corresponding accuracy metric per review type

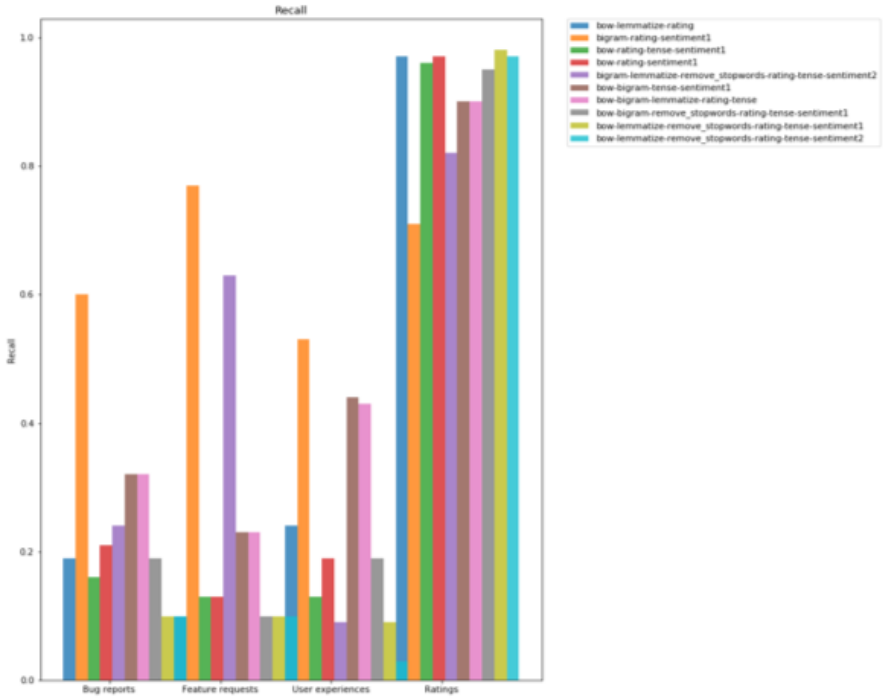
By comparing two tables, we notice that most of the results of the precision, recall and F1 score of our table are smaller than those of the original table in the essay, which indicates that our results are underestimated while the results in the original table are overestimated. To be specific, in the order of bug reports, feature requests, user experiences and ratings, the biggest difference in precision are 0.26, 0.3, 0.2 and 0.24; in recall are 0.29, 0.43, 0.34 and 0.47; in F1 score are 0.08, 0.18, 0.12 and 0.24 respectively.

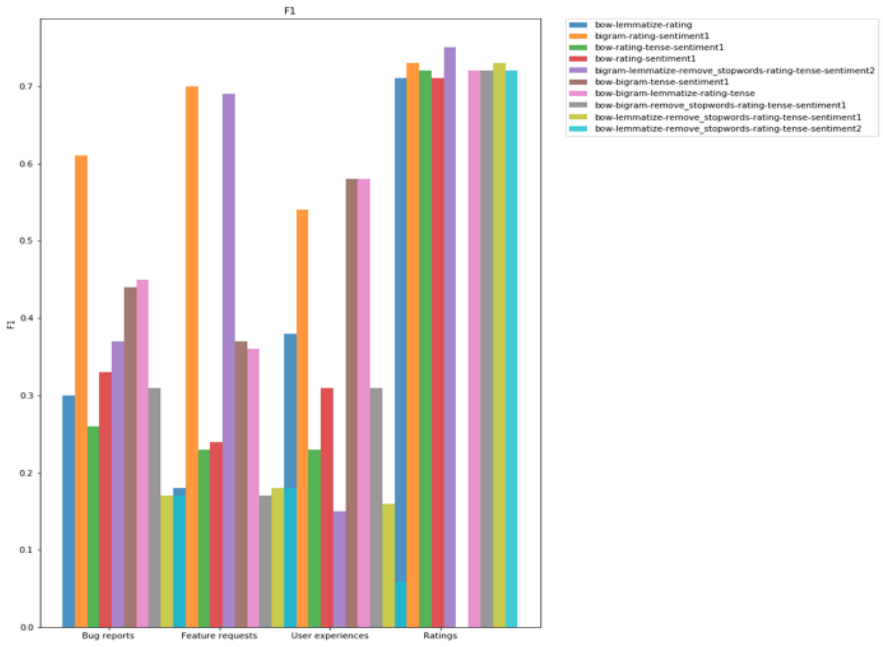
However, there are several results of the precision, recall and F1 score of our table are smaller than those of the original table in the essay. In the same order as above, in precision the biggest difference are 0.09, 0.12, 0.08 and 0.09; in recall are 0.23, 0.36, 0.25 and 0.18; in F1 score are 0.09, 0.09, 0.05 and 0.01 respectively.

Regarding observations above, we conclude that it is hard to achieve reproductivity as of some results, the differences are pretty large while largest difference occurs in bigram-rating-sentiment1 in recall value of ratings.

Histograms of the precision, recall and F1 score of 10 categories are shown below:







Answers to Part m

We cannot achieve the same accuracy or F1 score for each class, but the results are generally very close. For “Document classification (&NLP)”, all the classification methods that our team generated close but slightly lower precision, recall and F1 scores from the paper, except BOW has F1 score of 0.72 which is higher than 0.71 in the paper in bug report,

0.65 versus 0.63 in feature requests, 0.69 over 0.68 in user experience, and 0.76 over 0.75 in rating. For “Metadata”, all the classification methods generated very close precision, recall and F1 scores from the paper. Some features even have higher F1 scores, for instance, Rating-length-tense-sentiment2 has slightly higher F1 score in bug report. For “Combined (text and metadata)”, precision, recall and F1 scores generated from the classification methods generated slightly lower than the ones from the paper.

By comparing the results with the paper, this library is not completely reproducible. According to the The Machine Learning Reproducibility Checklist from Dr. Joelle Pineau at McGill University, the library has a complete description of the data collection process including sample size, a link to a downloadable version of the dataset, an explanation of any data that were excluded, description of any pre-processing step, and a clear definition of the specific measure or statistics used to report results.

However, it lacks an explanation of how samples were split for training/validation/testing, which is the major reason why the results on the paper are different from the one our team generated, because of random split for cross validation. Additionally, it lacks the range of hyper-parameters considered, method to select the best hyper-parameter configuration, and specification of all hyper-parameters used to generate results, a description of how experiments were run, the exact number of evaluation runs.

Group Members and Roles

Yi Tang was responsible for writing the code for pre processing.

Chenxi Yang was responsible for report writing and modifying the code.

Jasmine Chen was responsible for report writing and organizing data.

The whole team was involved in discussing problems, reading literature and report editing.

Extra thoughts, challenges, learnings, ideas

One of extra thoughts about this project is that we can amplify the sample size. For larger sample size, cross validation can be more stable, so that the results may be more ideal.

We have learned that, for app user reviews, particularly, the main text part and metadata such as length, tense, star rating, sentiments and submission time, are both important for classification. We have also learned different NLP methods including stop-words, lemmatization and n-gram.

We have encountered a few challenges along the project. The code for lemmatize_sentence part is complicated and requires a good understanding of the nltk package, map function, and lambda function. sentiScore requires to run a software then runs in Python. Also, labelling requires careful reading and comparing the four class types. Sometimes, one comment belongs to a few types, so it's challenging to label it with the most fitted type.

Final thoughts for reproducibility and robustness issue: in order to solve the machine learning reproducibility crisis, we should keep in mind to include an explanation of how samples split into training / validation / testing, as well as the range of hyper-parameters considered, method to select the best hyper-parameter configuration, and specification of all hyper-parameters used to generate results. These are the common factors that are always overlooked by data scientists and researchers.