CS 595: PROJECT REPORT

Sentiment Analysis for Product Review

Charu Saxena A20378393
Department of Computer Science,Illinois Institute of Technology, Chicago,IL 60616, USA

Abstract

The main aim of the project is to analyzes people's sentiments, attitudes, or emotions towards a particular product they bought from Amazon.com and posted their views in the form of comments/reviews on the website.

1. Problem Statement

In this project I aim to determine a more efficient way to get the sentiment of a review for a amazon product. Here I have compared rationale learning with traditional learning and have shown the result related.
For the experimental part I have used UCI: Sentiment Labelled Sentences Data set, which had originally three dataset out of which I have used Amazon product dataset, which has product reviews and comments for a cellphone and related accessories. The full dataset consist of 1000 reviews and all are labelled either positive or negative, which are denoted as either 1 (for positive) or 0 (for negative). The reviews were selected in a way that there are “clearly positive or negative connotation”[1], and no neutral. Following is the table from the UCI website[1] showing different aspects of the dataset:

Data Set Characteristics:	Text	**Number of Instances:**	3000
Attribute Characteristics:	N/A	**Number of Attributes:**	N/A
Associated Tasks:	Classification	**Missing Values?**	N/A

Table1: This shows different characteristics of the dataset selected.

2. Proposed Implementation

I used ipython notebook to implement this and python version 3.6.
This project aims at providing a more interactive way to predict the sentiment of the sentences based on the training data. For here the program was set in two parts where in one we considered learning with rationals and other is learning without rationals(traditional approach). This was done in following steps:

I. With rationale

Preparing data

First of all we know that, the reviews do not have any features, so the using CountVectorizer() will provide bag of words as the features for to do modeling. During this, we remove all the stop-words, and we set min_df =5, which means that word should have minimum frequency in 5 documents. We then get a dataframe, which is having bag of words for whole training set as a features and rows as different reviews, also we will define a column which will be having ratings as the score for that review correspondingly, as a sparse matrix.
NOw we will use train_test_split() with different random state value(to keep the documents selected same for every run while experimenting) and test data percentage.

Human-simulator

For learning with rationale approach I have taken simulated human-labeler to provide rationales based on the whole training set using chi-squared. This will give the chi-squared value for each feature and now we add them to a list of feature_score. Here also we are removing the nan values in the training set and setting it to 0. As, a result of which we see that we have 203 features in total. We rank the features and make a list of it. Now to consider a word as a rationale we see that of the word is present in the feature rank and if its occurrence is more in positive reviews we list them into positive rationale list or in negative rationale list.
Thus the the end we get two list: positive_rationale and negative_rationale

Updating weights in training examples

We will now update the values of the features for each review to show the effect of the rationals we have selected. As the matrix we are considering is a sparse matrix, we will work accordingly. In whole training set, for each review first we will see what label is present if the label is positive we will try to find the words present in the documents(value is 1 for it be a sparse matrix) and compare if it is also present in the positive rationales then we will multiply the value by ‘r’ else by ‘o’. And same for the negative labelled review we will compare with the negative rationale if that word is present update as ‘r’ if present and ‘o’ if not.
One case here is that when there is a presence of multiple rationales then we will choose a rationale at random and update the weight for both labels. For here we will consider r=1 and o=0.01

Fit to model

We have considered two classifiers here: Logistic Regression(both L1 and L2 regularization) and Multinomial Naive Bayes.
Here we fitted training dataset iteratively in pieces of 10,20,30.. And so on, and considered whole test set for to check accuracy and auc for it. When train_test_set split is 70%,66% and 80% we considered iterations of size 10 reviews as (10,20,30.. Till 100) and 10 times.
Whereas when we considered smaller training set with 5%, we considered iterations of 2 reviews stating taking 4 reviews as (4,6,8,10...till 22) total 10 times.

All the classifiers were set to default parameters, and fitting was done iteratively. To observe the results discussed in later part of the report.

II. Without rationale approach(Traditional)

For without rationale approach, the preparation of data was done in the similar was as done in rationale approach. After that no weights were updated we just fed the sparse matrix in iterative manner to all the three classifiers as discussed in the (fit the model) above, and the results were observed, which are discussed in later part of this report

3. Results and Discussion

Here we have considered results for all three classifiers with different training data and random state:

I) For Logistic regression(L1 regularization)

Here first of all we see the results with different number of random seed used for selecting the documents and seeing the accuracy curve and AUC curve with number of documents given on as the training data in the different model:

a) So first here we are using

Model: logistic regression L1 regularization
Test data:700 , Train data:300

Top 5 Pos_rationale:
[great, works, price , love, good ,excellent ,best,comfortable]

Top 5 negative_rationale:
[hear, poor, piece, buy ,terrible ,volume,broke, waste]

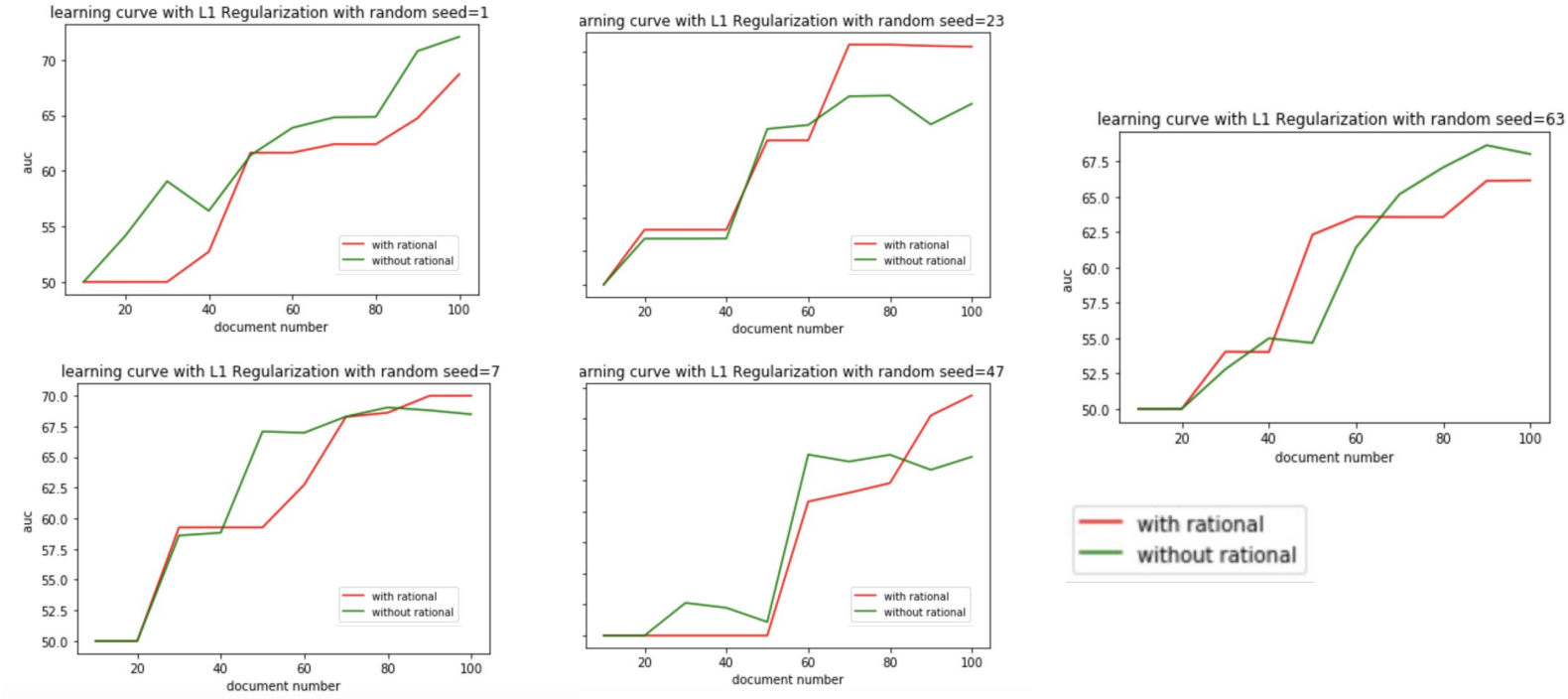


Figure1: This describes AUC learning curve with number of documents used for the training and testing is done on the whole test data with different random seed values as: 1,7,23,47 and 63.

So the mean AUC curve here observed was shown in figure 2:

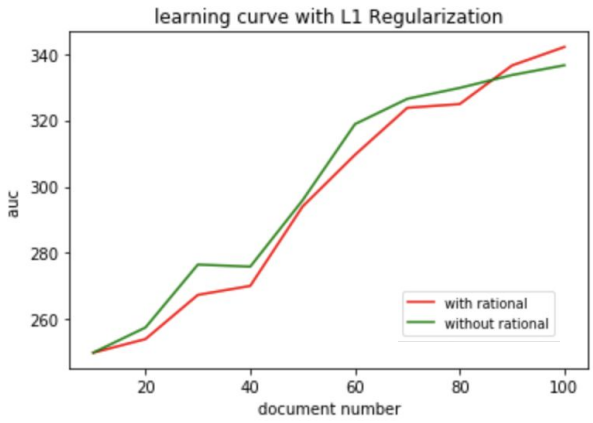


Figure2: This figure shows the mean values of the random seeds to give an approximate value of AUC for different number of the documents given for training the model

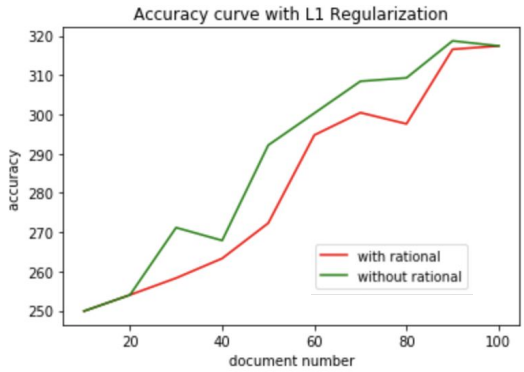


Figure3: This figure shows the mean values of the random seeds to give an approximate value of Accuracy for different number of the reviews given for training the model

Discussion

Here we observe that when number of documents increasing model is learning better in the case of rationales rather than traditional learning(AUC curve). This is happening because the rationales are providing weights to only to important words(rationales) as suggested by our simulated labeler. As the words represent positive rationales make sense and some of the negative rationales too, and these were given weights for the reviews in training set

b) So now we try reducing the training dataset and observing the results.

Model: logistic regression L1 regularization

Test data:800 , Train data:200

Top 5 Pos_rationale:

[great, love,works, best,price, excellent]

Top 5 negative_rationale:

[poor, buy, piece, waste money]

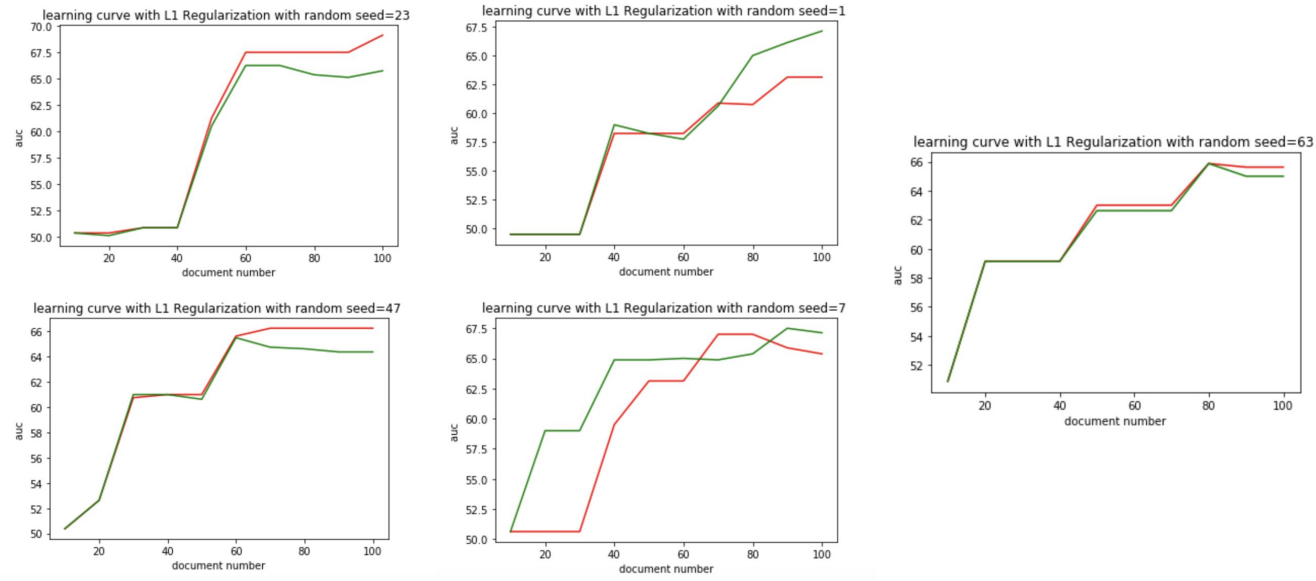


Figure4: This describes AUC learning curve with number of documents used for the training and testing is done on the whole test data with different random seed values as: 1,7,23,47 and 63.

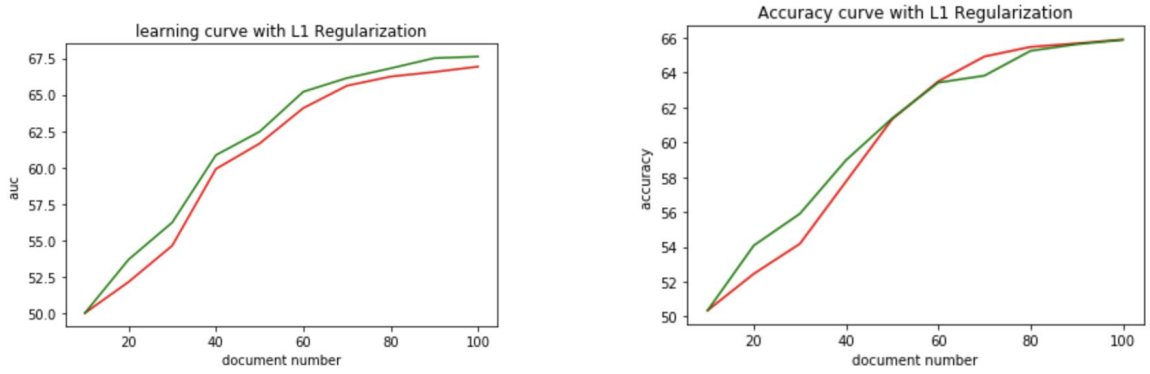


Figure5: This figure shows the mean values of the random seeds to give an approximate value of AUC and Accuracy for different number of the reviews given for training the model.

Discussion

Here we observe that when number of documents increasing model is learning better in the case of without rationales rather than rational learning this is opposite to what we observed in when the training set was larger by 100 reviews. This might be happening because the rationales provided might be not much as in the previous case and not good words observed as listed in the list of positive and negative rationals.

As the words represent positive rationales make sense to a level like(positive words included great love and best) and some of the negative rationales (like poor and waste)made sense but

not buy or piece but these must be related with negative phrases like , and these were given weights for the reviews in training set

c) So now we try increase the training dataset and observing the results.

Model: logistic regression L1 regularization

Test data:660 , Train data:340

Top 5 Pos_rationale:

[great, good, love, works ,nice, price ,best]

Top 5 negative_rationale:

[money, calls ,big ,bad ,disappointed ,stay,problem]

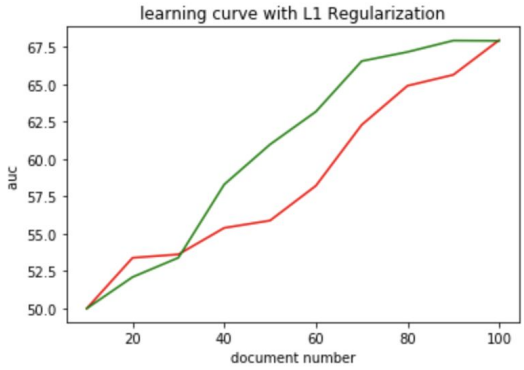


Figure6: This figure shows the mean values of the random seeds to give an approximate value of AUC for different number of the reviews given for training the model.

Discussion

Here we observe that when number of documents increasing model is learning better in the case of without rationales rather than rational learning this is opposite to what we observed in when the training set was smaller by 40 reviews. Here even though rationals make more sense than above cases like pos rationales(great, good, love) but negative rationales not much ([money, calls ,big where the first three don't make sense). This might be happening because the rationales provided might be not be good enough to determine a good result much as in the previous case and not good words observed as listed in the list of positive and negative rationals.so the auc curve for rationals is higher than without rationales.

II) for logistic regression(L2 regularization)

a) So first here we are using

Model: logistic regression L2 regularization

Test data:800 , Train data:200

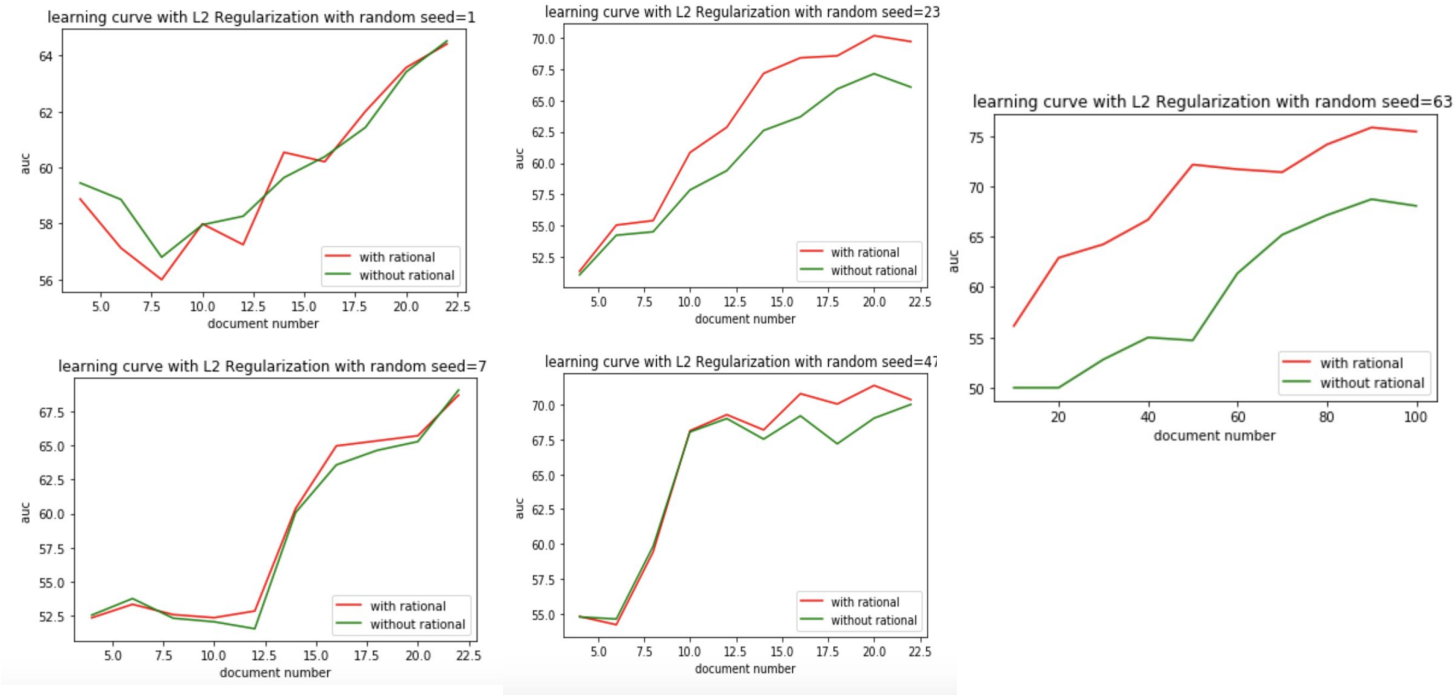


Figure 7: This describes AUC learning curve with number of documents used for the training and testing is done on the whole test data with different random seed values as: 1,7,23,47 and 52 for L2 regularization.

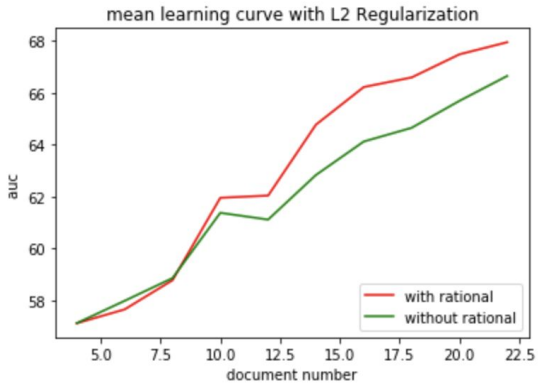


Figure8: This figure shows the mean values of the random seeds to give an approximate value of AUC and Accuracy for different number of the reviews given for training the model L2 logistic regression

<u>Discussion</u>

Here we observe that accuracy and AUC curve is changing from the last L1 approach here the learning curve performs better in case of L2 than in case L1. here the curve(fig8) shows that the model is learning better and performing better for Learning with rationals than traditional and

also the learning with rationals performs better with L2 better than with L1 for when test set split is 80%.

b) So first here we are using

Model: logistic regression L2 regularization

Test data:700 , Train data:300

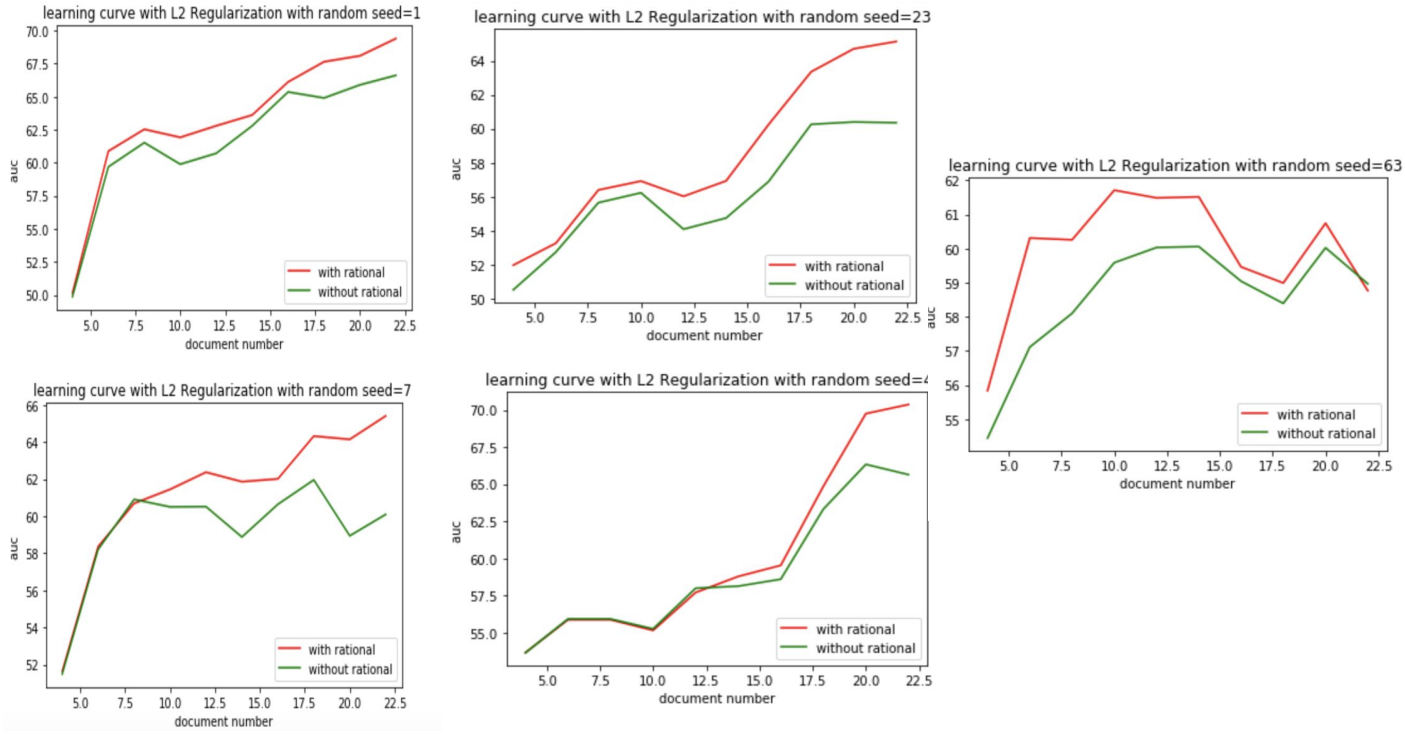


Figure 9: This describes AUC learning curve with number of documents used for the training and testing is done on the whole test data with different random seed values as: 1,7,23,47 and 63 for L2 regularization.

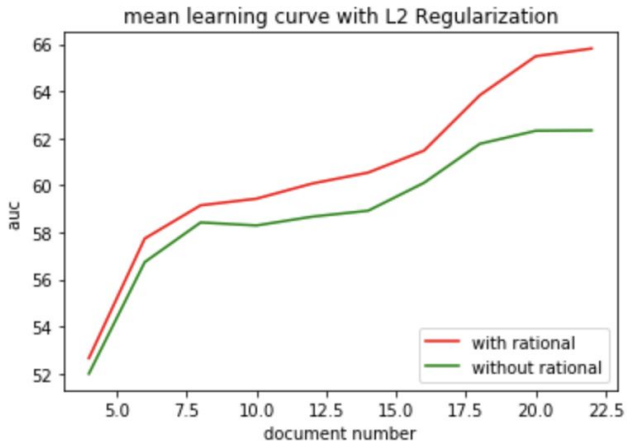


Figure 10: This figure shows the mean values of the random seeds to give an approximate value of AUC for different number of the reviews given for training the model L2 logistic regression

Discussion

Here we observe that AUC curve is changing from the last L1 approach here the learning curve performs better in case of L2 than in case L1 when train-test split is 70%. Previously in case of learning with rationales curve was minutely increasing but in case of L2 regularization we see a good difference and in fact rationals make the model perform better than traditional learning

c) So first here we are using

Model: logistic regression L2 regularization

Test data:660 , Train data:340

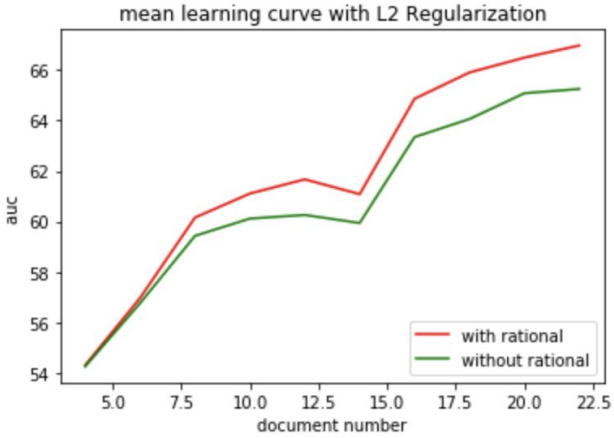


Figure 11:This figure shows the mean values of the random seeds to give an approximate value of AUC for different number of the reviews given for training the model L2 logistic regression .

Discussion:

Also for L2 we see with rationale approach works well than traditional approach for when we take just 50 documents and it make sense that it is increasing with number of documents as the rationales we are getting are better and better. Previously for L1 traditional method was performing well but here we see that with L2 regularization rationals is performing better

III) for Multinomial NAive bayes

a) So first here we are using

Model: MultinomialNB

Test data:700 , Train data:300

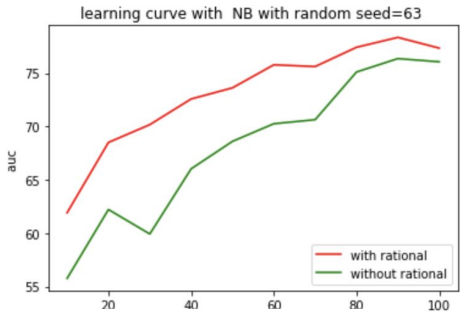
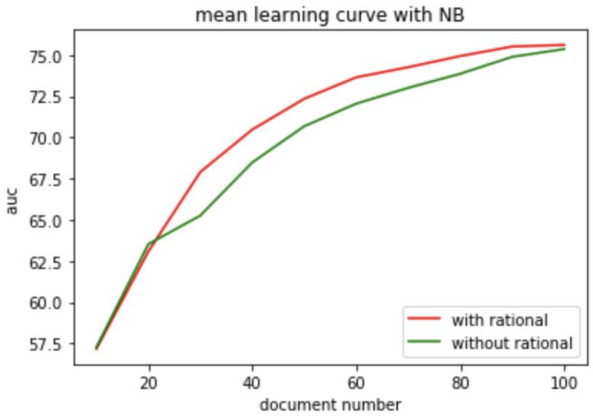


Figure 12: This describes AUC learning curve with number of documents used for the training and testing is done on the whole test data with different random seed values as: 1,7,23,47 and 63 for naive bayes

So mean aUC and accuracy:



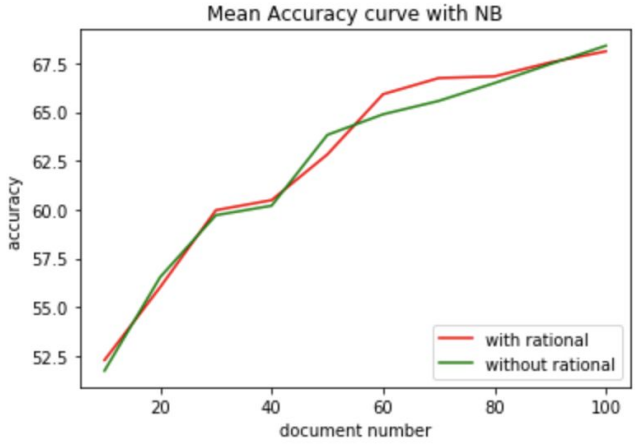


Figure 13:This figure shows the mean values of the random seeds to give an approximate value of AUC and Accuracy for different number of the reviews given for training the model Naive bayes.

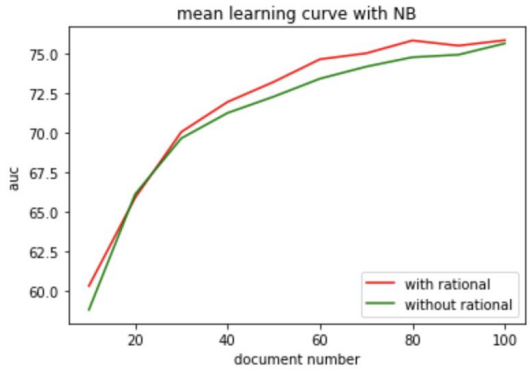
Discussion:

Here we see that first of all, learning with rationals are performing better than traditional method which adds on to the result of L2 regularization. But there is less significant difference between the two.

b) So first here we are using

Model: MultinomialNB

Test data:800 , Train data:200



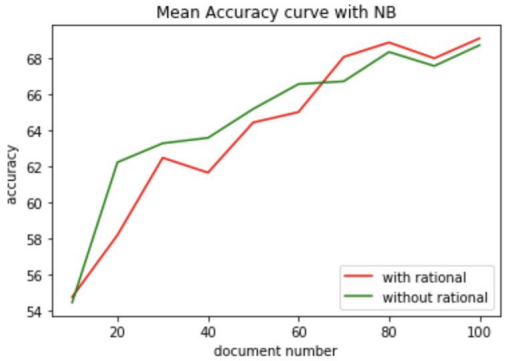
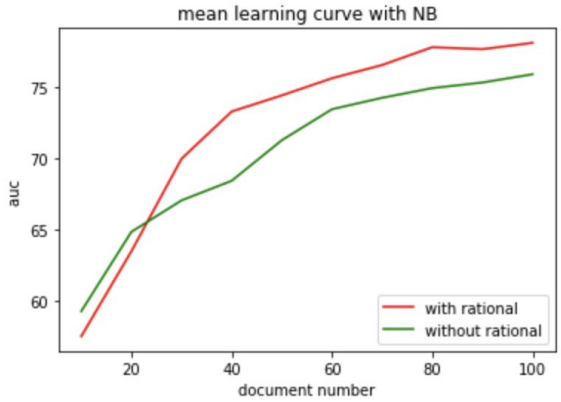


Figure 14:This figure shows the mean values of the random seeds to give an approximate value of AUC and Accuracy for different number of the reviews given for training the model Naive bayes.

c) So first here we are using

Model: MultinomialNB

Test data:660 , Train data:340



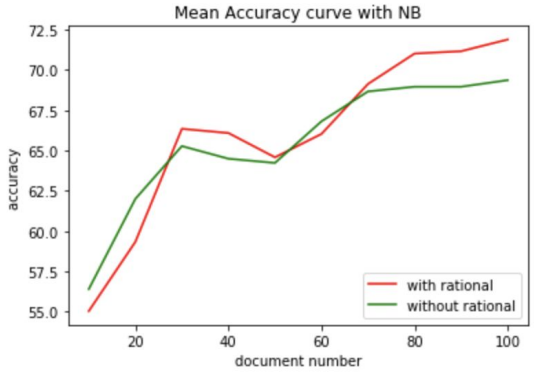


Figure 15:This figure shows the mean values of the random seeds to give an approximate value of AUC and Accuracy for different number of the reviews given for training the model Naive bayes.

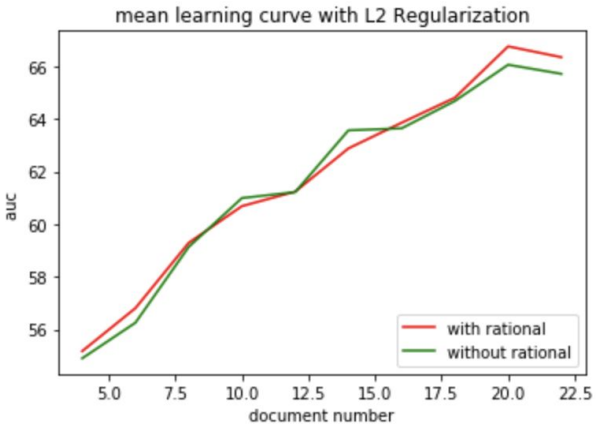
Discussion: Here for both 68% and 80% we observe that the results are similar to 70% only, where learning with rationals are performing better than traditional method with lesser difference.

III) **This is the part which was done to see if we give very less documents:**
Second part of testing was done taking very small set for training, here we considered:
Training dataset=50, and test dataset as 950.

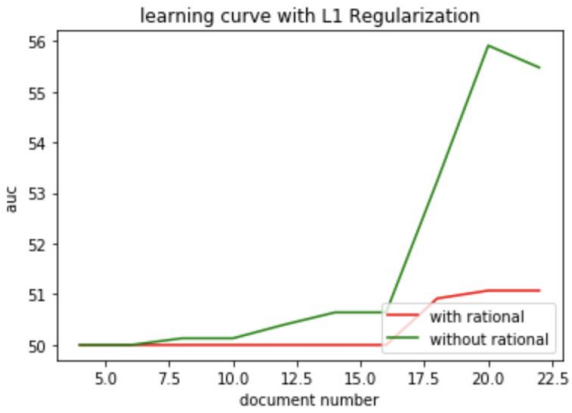
Top 5 Pos_rationale:
[great ,works, cell,price,fits ,time,recommend]

Top 5 negative_rationale:
[buttons,calls,life,year,terrible,small]

For L2 regularization following was the result:



For L1 Regularization following was the result:



For NB following was the result:

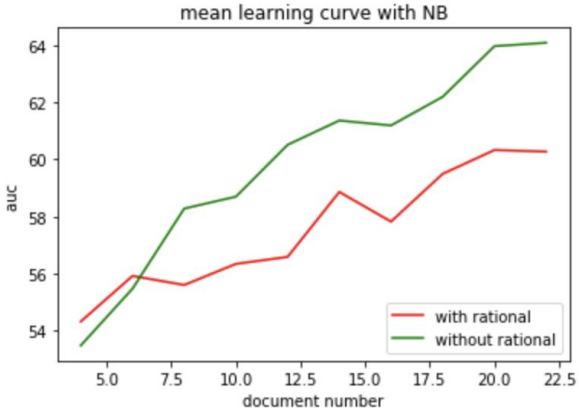


Figure 16:This figure shows the mean values of the random seeds to give an approximate value of AUC and Accuracy for different number of the reviews given for training the model Naive bayes.

Discussion:

We see that when very less documents are given then learning with rationals work better only with L2 regularization and not in other classifiers. One of the plausible reason might be that, here there are very less reviews thus, simulator has very less documents to see and get rationales that might make sense also, as it will pick up rationales based on the ratio of positive and negative documents

4. Conclusion

Conclusively, we see that with l2 regularization of logistic regression and multinomial naive bayes learning with rationals works better than traditional learning, whereas logistic regression with L1 regularization was an exception for this.

For number of train set, we see that more the number of reviews we take as training ,better the simulator will choose the rationals and better will be the learning.

References:

[1]- https://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences
[2]- http://www.cs.iit.edu/~ml/pdfs/sharma-naaclhlt15.pdf