Sentiment Analysis for Hotel Reviews

Presented by:
Vishal Kumar Bimal
A20380261

Abstract

This report consists of the detail implementation of comparison between traditional learning, learning with rationale and TANDEM learning using the sentiment analysis of hotel reviews data from Kaggle. This report consists of the comparison between the above three implementations based on the different experiments conducted using two different classifier Logistic regression with L2 as penalty and multinomial Naïve Baye's and finally a graph is shown comparing the accuracy of the all the three methods on 200 documents.

Contents:

- Purpose
- Dataset
- Approach
- Experiments
- Results
- Furture Scope

Purpose:

- A review can, of course, be negative or positive and with social platforms such as TripAdvisor,
- it’s never been easier for a customer to state their opinion about somewhere they've visited.
- While some hospitality owners and operators perish at the thought of reviews, they've brought great change to the industry and that is a huge need for proactivity.
- 96% of TripAdvisor users consider reading **reviews important** when planning trips and booking hotels.

Dataset

Website: https: //www.kaggle.com/harmanpreet93/hotelreviews/data

Dataset: Hotel Reviews

Number of instances: 38,000

Labeled/unlabeled: Labeled

Data used for generating rationale : 3893

No of documents used for fitting : 200 (Iteratively after every 5 document)

Approach: **Traditional Learning**

Use scikit-learn implementation of Logistic regression and Multinomial naïve Bayes to fit the model (without any rationale) iteratively after every 5 documents from the training set and predict the accuracy of the model on the test score. Repeat this for 200 training documents.

Approach: TANDEM Learning

- Go through each document in the loop.
- Check if the label assigned to document is positive or negative.
- If the label is positive match all the words present in the list of positive rationale and words present in the document. If match is found multiply the value with 1 and if the words does not match with list multiply it with 0.01
- Repeat the same with negative label document but match the words with negative rationale list.
- After every 5 documents fit the model and predict the accuracy of the model with test data and save it.
- Continue this with 200 documents form train.

Approach: Learning with rationale

- Go through each document in the loop.
- Check if the label assigned to document is positive or negative.
- If the label is positive match all the words present in the list of positive rationale and words present in the document. Select one word at random from the matched list. Multiply the occurrence of the matched word with 1 and all the other words in the document with 0.01
- Repeat the same with negative label document but match the words with negative rationale list.
- After every 5 documents fit the model and predict the accuracy of the model with test data and save it
- Continue this with 200 documents form train.

Automated human: Chi sqaure

- Use the chi_square implementation of sklearn to get the words which have highest chi score from the training set.

- Count the occurrence of the word in positive and negative document and depending on the counts divide the words into negative or positive rationale.

- In order to get the better words take the ratio of negative and positive count if the negative-count/positive count > 3 assign it as a negative rationale and if the positive-count/negative-count > 3 assign it has positive rationale

Rationale from automated user:

Positive rationale

['great', 'helpful', 'comfortable', 'perfect', 'excellent', 'friendly', 'wonderful', 'loved', 'definitely', 'highly', 'quiet', 'modern', 'fantastic', 'restaurants', 'recommend']

Negative rationale

['dirty', 'worst', 'poor', 'broken', 'rude', 'stained', 'horrible', 'terrible', 'stains', 'worn', 'worse', 'smelled', 'carpets', 'mold', 'properly', 'disgusting', 'filthy', 'uncomfortable', 'complained', 'clearly', 'joke', 'musty', 'stuck', 'apology', 'speak', 'refund', 'stated', 'apparently', 'outdated']

Negative rationale and its count in training set

```
the negative word is  dirty and count in training set is  151
the negative word is  worst and count in training set is  88
the negative word is  finally and count in training set is  113
the negative word is  carpet and count in training set is  115
the negative word is  broken and count in training set is  60
the negative word is  poor and count in training set is  86
the negative word is  rude and count in training set is  82
the negative word is  stains and count in training set is  53
the negative word is  horrible and count in training set is  61
the negative word is  disgusting and count in training set is  36
the negative word is  terrible and count in training set is  60
the negative word is  filthy and count in training set is  35
the negative word is  worn and count in training set is  68
the negative word is  carpets and count in training set is  46
the negative word is  awful and count in training set is  46
the negative word is  shabby and count in training set is  29
the negative word is  peeling and count in training set is  33
the negative word is  clearly and count in training set is  42
the negative word is  stained and count in training set is  37
the negative word is  worse and count in training set is  39
the negative word is  dirt and count in training set is  23
the negative word is  disappointing and count in training set is  39
the negative word is  smoke and count in training set is  41
the negative word is  hairs and count in training set is  17
the negative word is  poorly and count in training set is  27
the negative word is  shut and count in training set is  38
```

Negative rationale and its count in test set

```
the negative word is  dirty and count in test set is  1211
the negative word is  worst and count in test set is  771
the negative word is  finally and count in test set is  1008
the negative word is  carpet and count in test set is  1011
the negative word is  broken and count in test set is  644
the negative word is  poor and count in test set is  863
the negative word is  rude and count in test set is  812
the negative word is  stains and count in test set is  367
the negative word is  horrible and count in test set is  571
the negative word is  disgusting and count in test set is  277
the negative word is  terrible and count in test set is  679
the negative word is  filthy and count in test set is  291
the negative word is  worn and count in test set is  580
the negative word is  carpets and count in test set is  318
the negative word is  awful and count in test set is  398
the negative word is  shabby and count in test set is  187
the negative word is  peeling and count in test set is  211
the negative word is  clearly and count in test set is  458
the negative word is  stained and count in test set is  328
the negative word is  worse and count in test set is  418
the negative word is  dirt and count in test set is  123
the negative word is  disappointing and count in test set is  434
the negative word is  smoke and count in test set is  404
the negative word is  hairs and count in test set is  99
the negative word is  poorly and count in test set is  180
the negative word is  shut and count in test set is  315
```

Positive rationale and its count in training set:

```
the negative word is  great and count in test set is  1689
the negative word is  helpful and count in test set is  830
the negative word is  wonderful and count in test set is  372
the negative word is  perfect and count in test set is  347
the negative word is  excellent and count in test set is  509
the negative word is  friendly and count in test set is  981
the negative word is  loved and count in test set is  270
the negative word is  highly and count in test set is  203
the negative word is  comfortable and count in test set is  917
the negative word is  spacious and count in test set is  322
the negative word is  staff and count in test set is  1946
the negative word is  amazing and count in test set is  223
the negative word is  definitely and count in test set is  475
the negative word is  enjoyed and count in test set is  272
```

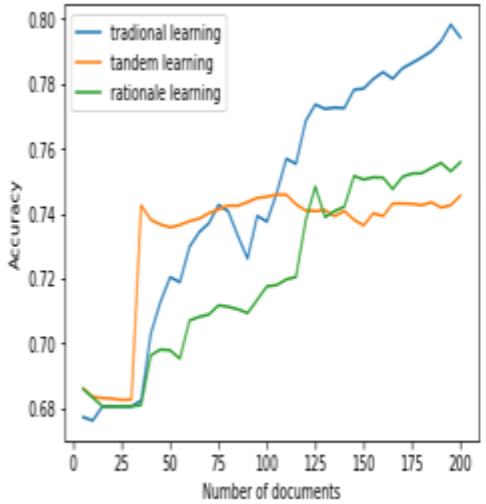
Positive rationale and its count in test set:

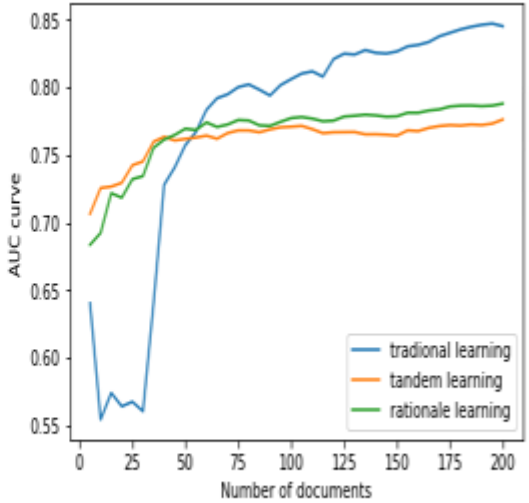
```
the positive word is  great and count in test set is  14915
the positive word is  helpful and count in test set is  7290
the positive word is  wonderful and count in test set is  2967
the positive word is  perfect and count in test set is  3032
the positive word is  excellent and count in test set is  4326
the positive word is  friendly and count in test set is  8893
the positive word is  loved and count in test set is  2228
the positive word is  highly and count in test set is  1750
the positive word is  comfortable and count in test set is  8118
the positive word is  spacious and count in test set is  2814
the positive word is  staff and count in test set is  17154
the positive word is  amazing and count in test set is  1844
the positive word is  definitely and count in test set is  4008
the positive word is  enjoyed and count in test set is  2228
```

Data used for Experiment1

- Size of dataset : 38932
- Size of training set : 3893
- Number of features : 18887
- Random seed = 0
- Number of positive instance :2670
- Number of negative instance :1233
- Number of rationale : 44 (Positive rationale : 19 , Negative rationale:26)
- Classifier : Logistic Regression(Penalty= L2)
- Evaluation measure: Accuracy and AUC

Results of experiment1



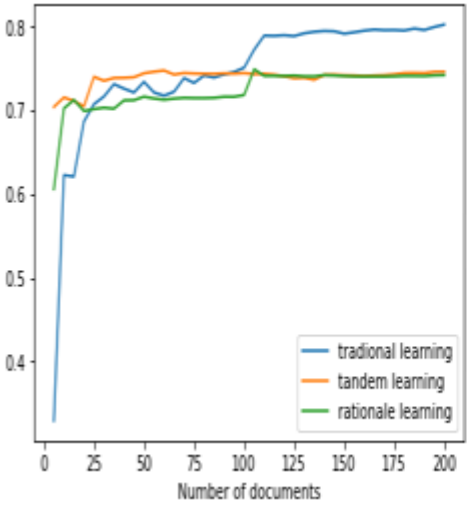


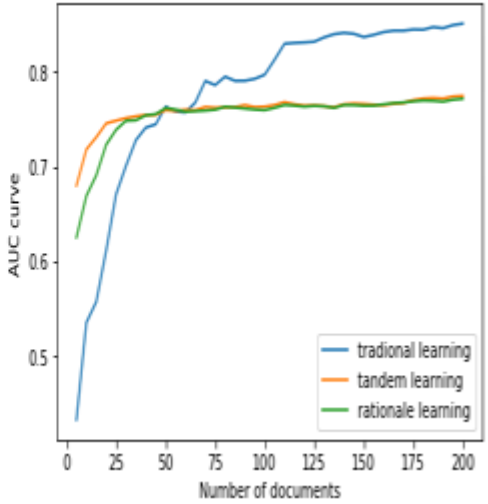
●It can be seen that that in AUC curve TANDEM learning and Learning with rationale dine well for the frirst 50 documents but after that tradional learning pvertook them .

Data used for Experiment2

- Size of dataset : 38932
- Size of training set : 3893
- Number of features : 18887
- Random seed = 6
- Number of positive instance :2645
- Number of negative instance :1248
- Number of rationale : 40 (Positive rationale : 14 , Negative rationale:26)
- Classifier : Logistic Regression(Penalty= L2)
- Evaluation measure: Accuracy and AUC

Results of experiment2

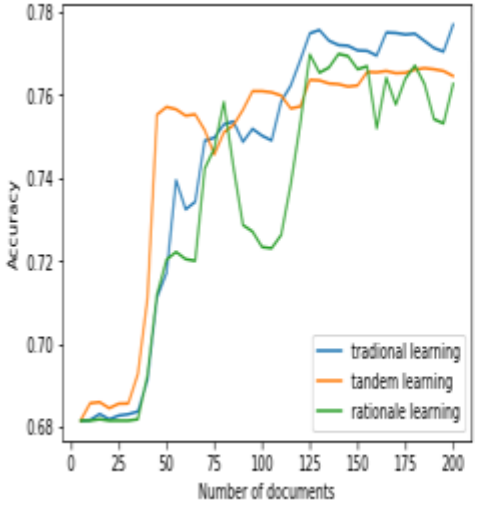


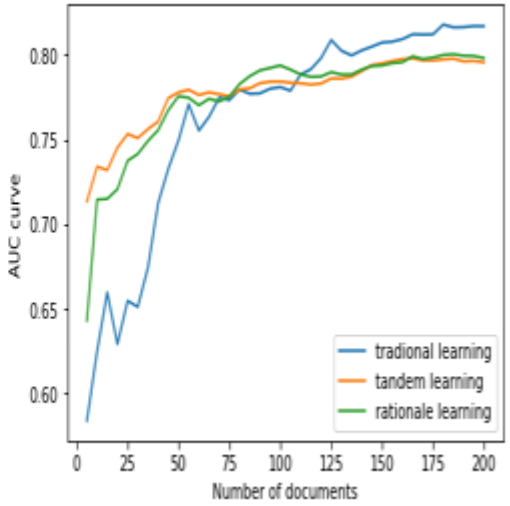


Data used for Experiment3

- Size of dataset : 38932
- Size of training set : 3893
- Number of features : 18887
- Random seed = 11
- Number of positive instance :2641
- Number of negative instance :1252
- Number of rationale : 36 (Positive rationale : 16 , Negative rationale:20)
- Classifier : Logistic Regression(Penalty= L2)
- Evaluation measure: Accuracy and AUC

Result for Experiment:3

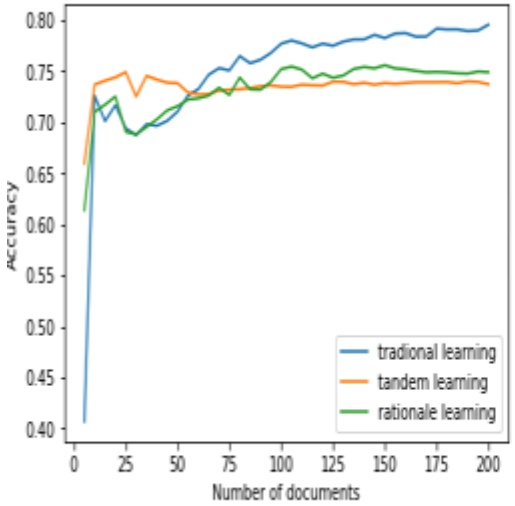


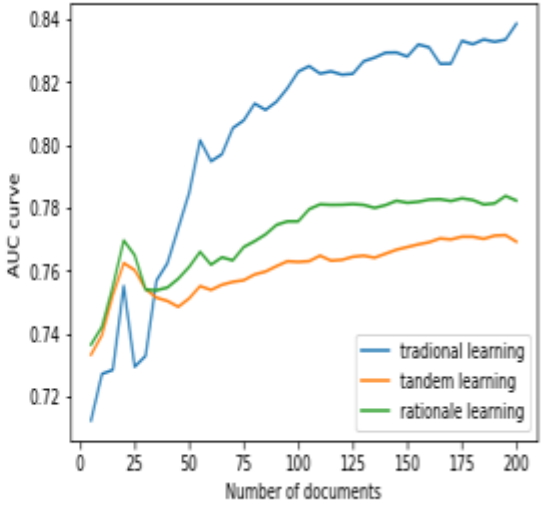


Data used for Experiment4

- Size of dataset : 38932
- Size of training set : 3893
- Number of features : 18887
- Random seed = 19
- Number of positive instance :2623
- Number of negative instance :1270
- Number of rationale : 41 (Positive rationale : 17 , Negative rationale:24)
- Classifier : Logistic Regression(Penalty= L2)
- Evaluation measure: Accuracy and AUC

Results for experiment4:

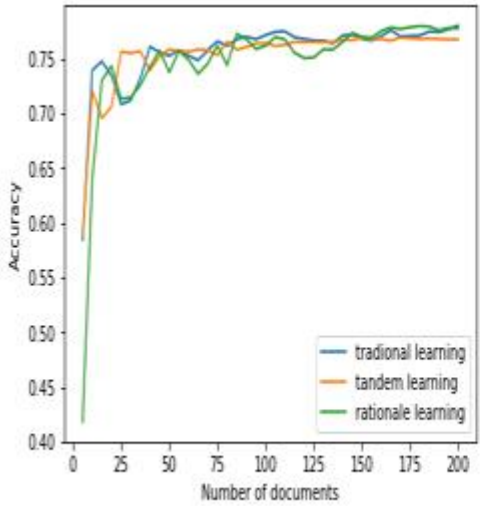


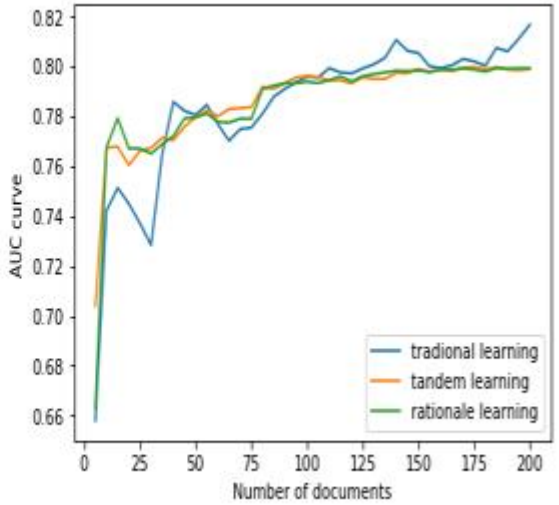


Data used for Experiment5

- Size of dataset : 38932
- Size of training set : 3893
- Number of features : 18887
- Random seed = 67
- Number of positive instance :2649
- Number of negative instance :1244
- Number of rationale : 37 (Positive rationale : 16 , Negative rationale:21)
- Classifier : Logistic Regression(Penalty= L2)
- Evaluation measure: Accuracy and AUC

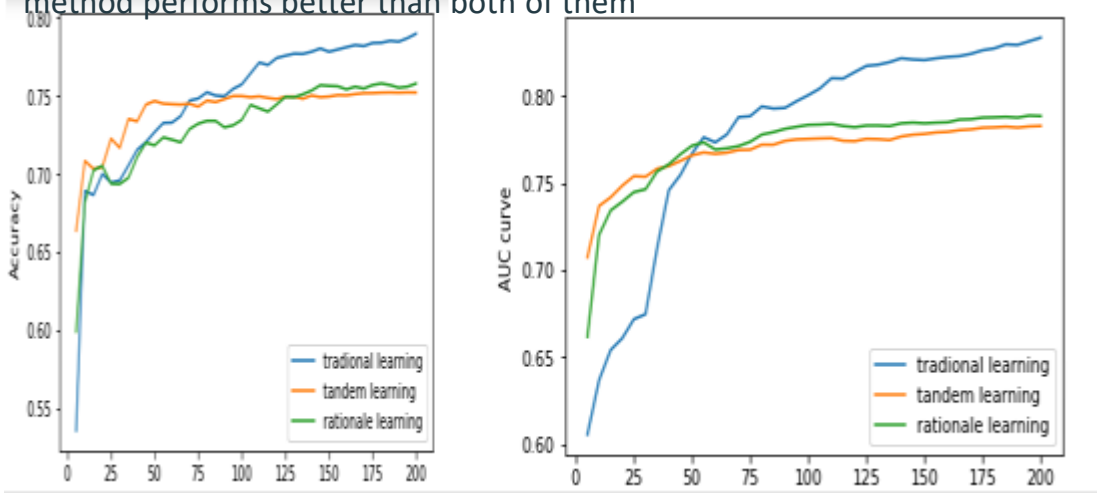
Result for experiment5:





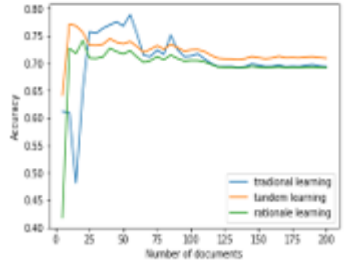
Average Result of 5 experiment:

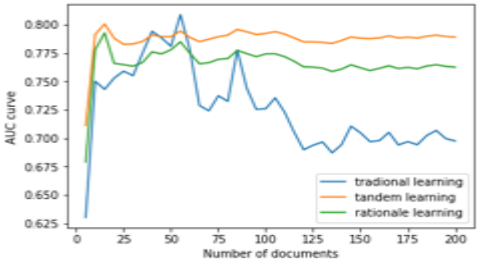
●We can see from the average result that till first 50 documents the learning rationale and TANDEM learning performs better that traditional learning but as the number of documents increase traditional method performs better than both of them



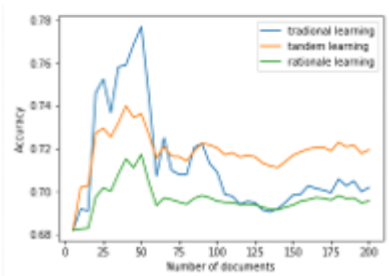
Results with Naïve Bayes(2 experiments)

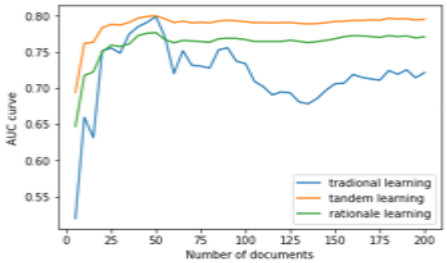
random seed = 67



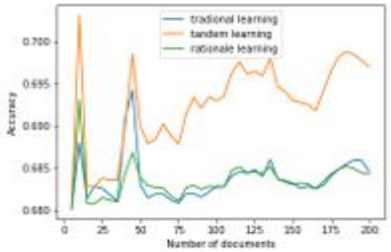


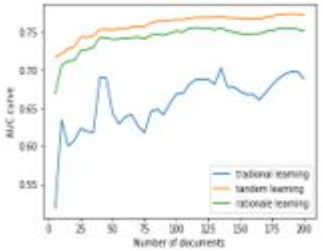
Random seed = 109





Results with Naïve Bayes(2 experiments)



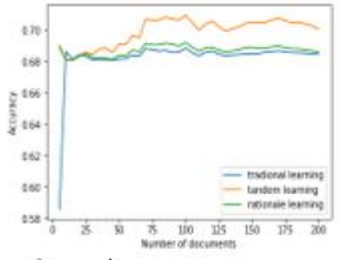


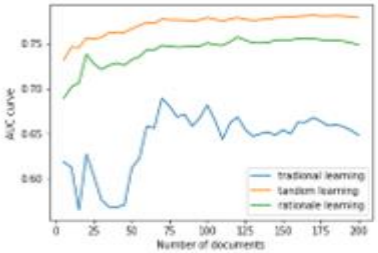
Random seed =319

Total number of documents = 3893

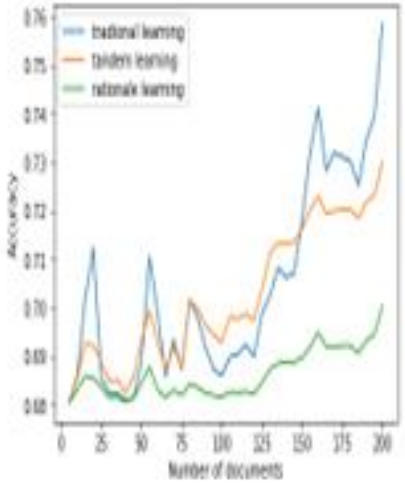
positive documents = 2680

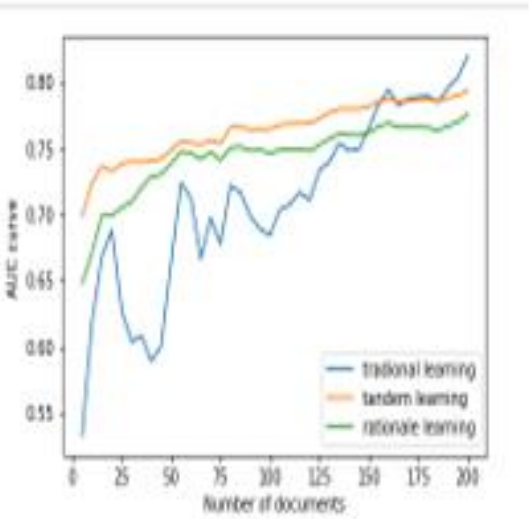
negative documents = 1213



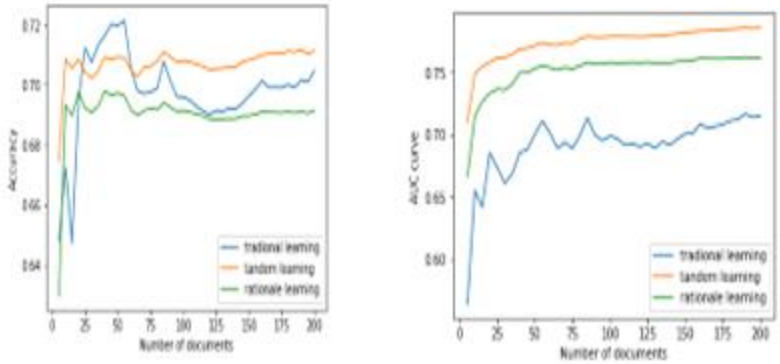


Result with Naïve Bayes(5th experiment)





Final result of Naïve Bayes



●In the above result we can see that TANDEM learning and Learning with rationale performed far better than traditional learning. Even if the accuracy is not good for learning with rationale but that may be because of the un-balanced dataset, because AUC is showing better results for it.

Conclusion:

- TANDEM and Learning with rationale can be used when the number of labelled document is less but if we have large set of labelled document its better to use traditonal learning.

Future Scope:

- Active Learning
- Increase or decrease Rationale
- Different datasets
- More classifier