EE219: Project 1

Due on Jan. 29, 2018

Zeyu Zhang (505030513)
Yunchu Zhang (805030502)

Introduction

Statistical classification refers to the task of identifying a category, from a predefined set, to which a data point belongs, given a training data set with known category memberships. Classification differs from the task of clustering, which concerns grouping data points with no predefined category memberships, where the objective is to seek inherent structures in data with respect to suitable measures. Classification turns out as an essential element of data analysis, especially when dealing with a large amount of data. In this project, we look into different methods for classifying textual data.

Dataset and Problem Statement

In this project, we work with "20 Newsgroups" dataset. It is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups, each corresponding to a different topic.

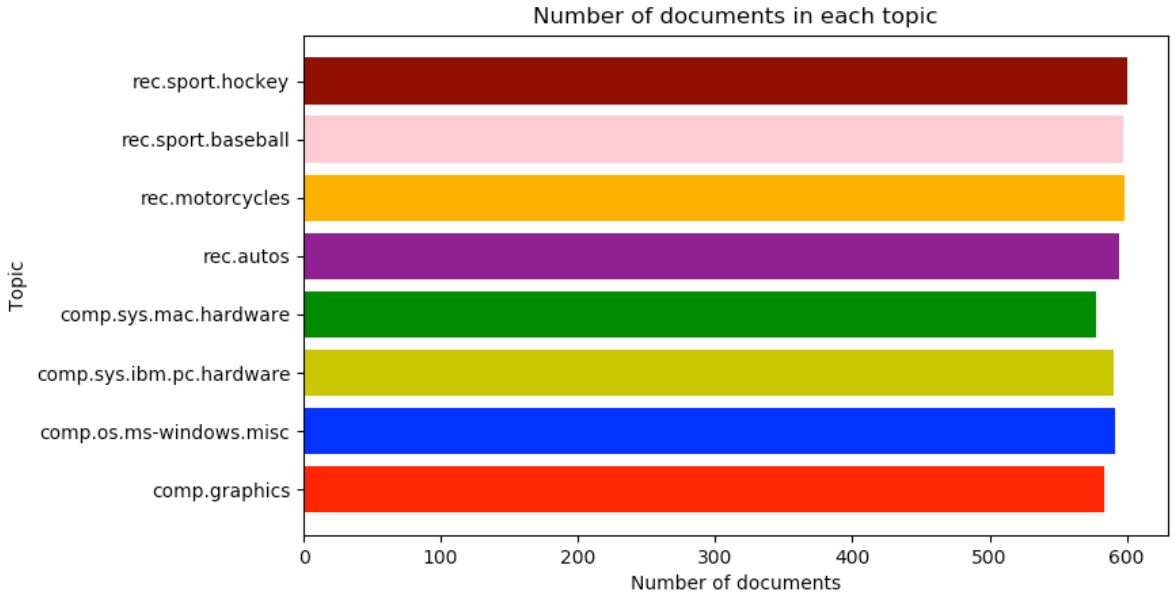
For the purposes of this project we will be working with only 8 of the classes as shown in Table 1. Load the training and testing data for the following 8 subclasses of two major classes 'Computer Technology' and 'Recreational activity'.

Table 1 Subclasses of 'Computer technology' and 'Recreational activity'

Computer Technology	Recreational Activity
comp.graphics	rec.autos
comp.os.ms-windows.misc	rec.motorcycles
comp.sys.ibm.pc.hardware	rec.sport.baseball
comp.sys.mac.hardware	rec.sport.hockey

1. Part (a) plot a histogram of the number of training documents

In this problem, we use the fetch_20newsgroups dataset from sklearn package to plot the number of training documents per class. The histogram is shown as following:



From the figure above, we could see that the data set is already balanced and in every class the number of documents is almost equivalent (balanced dataset).

2. Part (b) Create TFxIDF vector representation

A popular numerical statistic to capture the importance of a word is TFxIDF metric. This measure takes into account the words in the document. The discriminating words will most likely be those that are specialized terms describing different types of accessories and hence will occur in fewer documents.

In this problem, we need to convert the documents of 8 classes into numerical feature vectors. Firstly, we use ‘snowball’ to remove morphological affixes from words. After that, we remove all punctuation, all symbols and non-ascii characters. We used TfidfTransformer to calculate TFIDF, where tf(t, d) represents the frequency of term t in document d, and inverse document frequency is defined as: idf(t) = log [n / df(t)] + 1 and df(t) is the document frequency; the document frequency is the number of documents that contain the term t.

When we set min_df = 2, the shape of the TFxIDF vector is (4732, 21795), which means the final number of terms is 21795. When we set min_df = 5, the shape of the TFxIDF vector is (4732, 8818), which means the final number of terms is 8818.

3. Part(c) Calculate the TFxICF

In order to quantify how significant a word is to a class, we calculate TFxICF. It is the same as TFxIDF, except that a class sits in place of a document; that is for a term t and a class c. And we need to find the 10 most significant terms in each of the 4 classes with respect to TFxICF measure.

The way we calculate the TFxICF is exactly as same as TFxIDF. This this part we treat each category as a huge document. We concatenate all the documents in the same category together to form a 'large' document, and this document represents the category. Here we have 20 categories, then we have 20 large documents. We then feed these 20 documents into the calcTFxIDF function, the output will be the TFxICF. The 10 most significant terms for each category is as following.

Table 2 Top 10 Significant Terms

comp.sys.ibm.pc.hardware	comp.sys.mac.hardware	misc.forsale	soc.religion.christian
scsi	edu	edu	god
drive	line	line	edu
edu	mac	sale	christian
ide	subject	subject	jesus
line	organization	organization	say
use	apple	new	church
com	use	post	subject
subject	quadra	com	people
organization	scsi	university	line
controller	problem	offer	know

4. dimensionality Reduction (LSI & NMF)

In this problem, we need to transform the data's dimensional into a lower dimensional space. We use Latent Semantic Indexing (LSI) and NMF to lower the space into 50. As to LSI, we used the SVD decomposition to obtain left and right singular vectors and then get a lower space. We use Latent Semantic Indexing (LSI) to find the optimal representation of the data in a lower dimensional space. We use TruncatedSVD from sklearn's decomposition package to decompose the vectors with 50 as the number of elements. Therefore, we get the selected features for our learning algorithms.

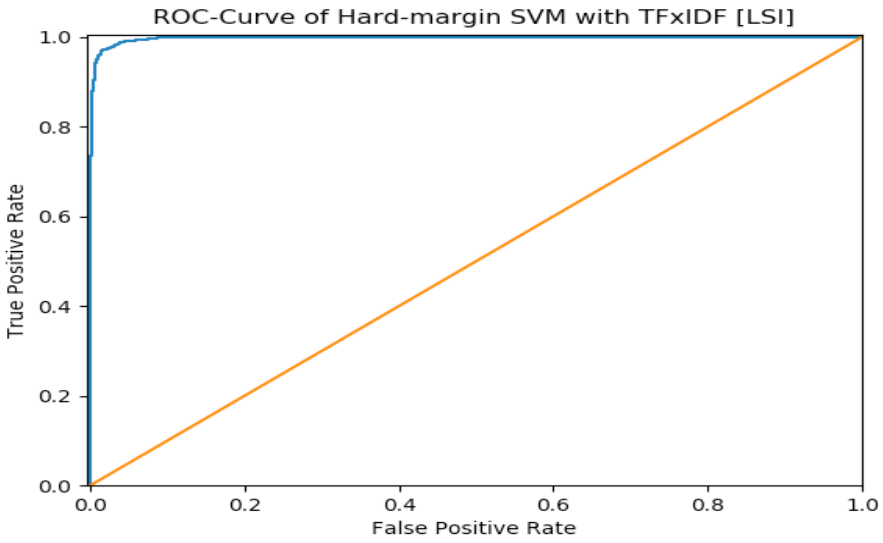
- Shape of feature vector after LSI: (4732, 50)
- Shape of feature vector after NMF: (4732, 50)

5. Linear Support Vector Machine

In next parts, we need to classify the documents into two categories ‘‘Computer Technology’ vs ‘Recreational Activity’. We need to combine documents of sub-classes of each class to form the set of documents for each class. And we need to assess the model by precision, recall and ROC curve. In this part, based on LSI and NMF, we used SVM to perform classification.

When min_df = 2

1) Hard-Margin SVM with LSI



	Precision	Recall	Accuracy
Computer technology	0.9938	0.9295	0.9622
Recreational activity	0.9349	0.9943	0.9622
Average	0.9619	0.9644	0.9622

Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

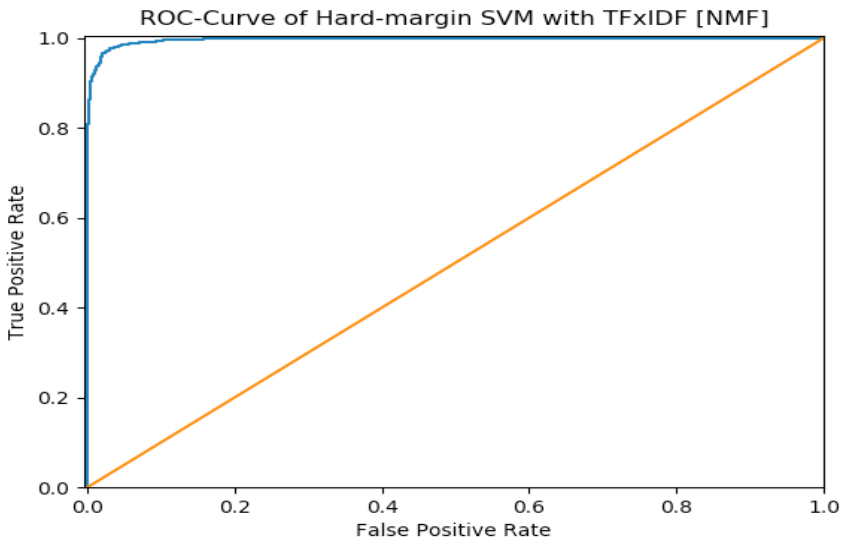
	TP	FP	FN	TN
Computer technology	1450	9	110	1581

Recreational activity	1581	110	9	1450

If We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1450	9
Predicted negative	110	1581

2) Hard-Margin SVM with NMF



	Precision	Recall	Accuracy
Computer technology	0.9916	0.9090	0.9511
Recreational activity	0.9174	0.9925	0.9511
Average	0.9507	0.9545	0.9511

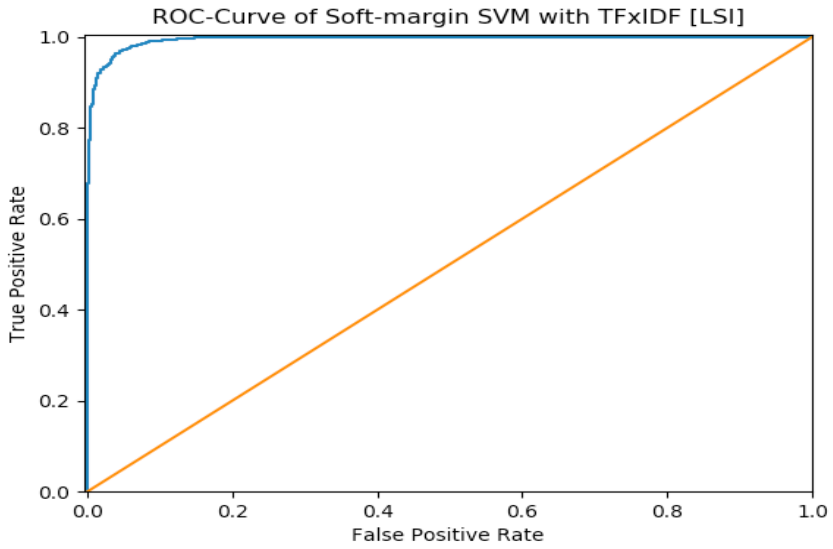
Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

	TP	FP	FN	TN
Computer technology	1418	12	142	1578
Recreational activity	1578	142	12	1418

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1418	12
Predicted negative	142	1578

3) Soft-Margin SVM with LSI



	Precision	Recall	Accuracy
Computer technology	0.9922	0.8917	0.9429
Recreational activity	0.9033	0.9931	0.9429
Average	0.9424	0.9477	0.9429

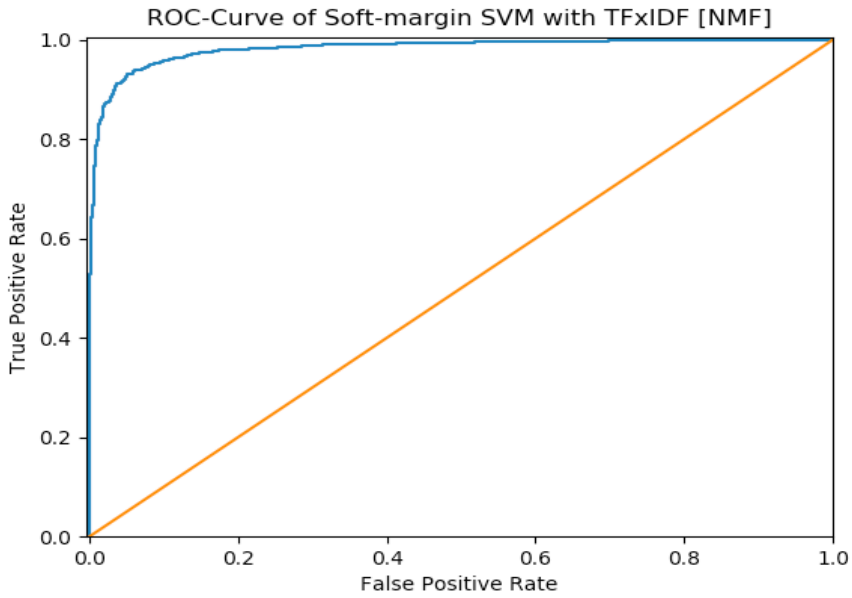
Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

	TP	FP	FN	TN
Computer technology	1391	11	169	1579
Recreational activity	1579	169	11	1391

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1391	11
Predicted negative	169	1579

4) Soft-Margin SVM with NMF



	Precision	Recall	Accuracy
Computer technology	1.0000	0.0410	0.5251
Recreational activity	0.5152	1.0000	0.5251
Average	0.5205	0.7576	0.5251

Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

	TP	FP	FN	TN
Computer technology	64	0	1496	1590
Recreational activity	1590	1496	0	64

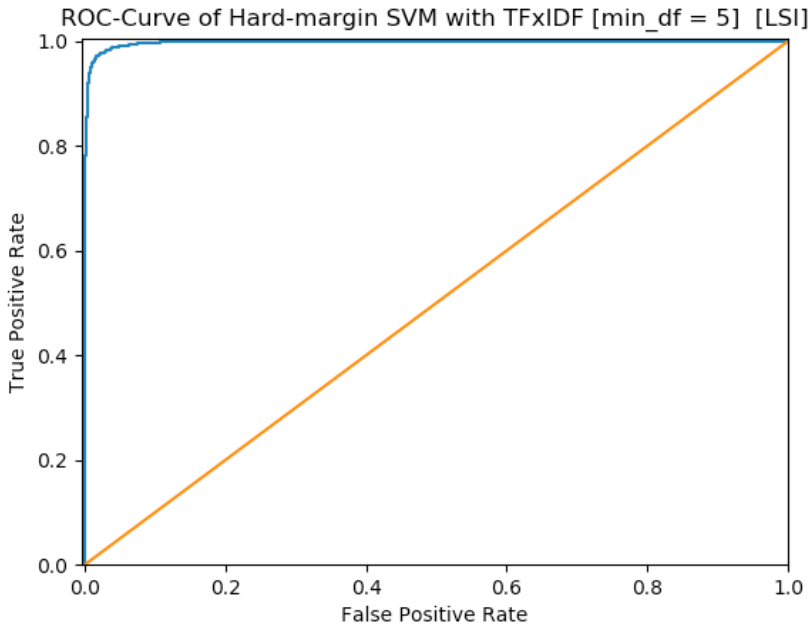
We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	64	0
Predicted negative	1496	1590

For min_df=2, the LSI model is better then NMF and hard SVM is better than soft SVM.

When we set min_df=5

1) Hard-Margin SVM with LSI



	Precision	Recall	Accuracy
Computer technology	0.9666	0.9827	0.9746
Recreational activity	0.9827	0.9667	0.9746
Average	0.9747	0.9747	0.9746

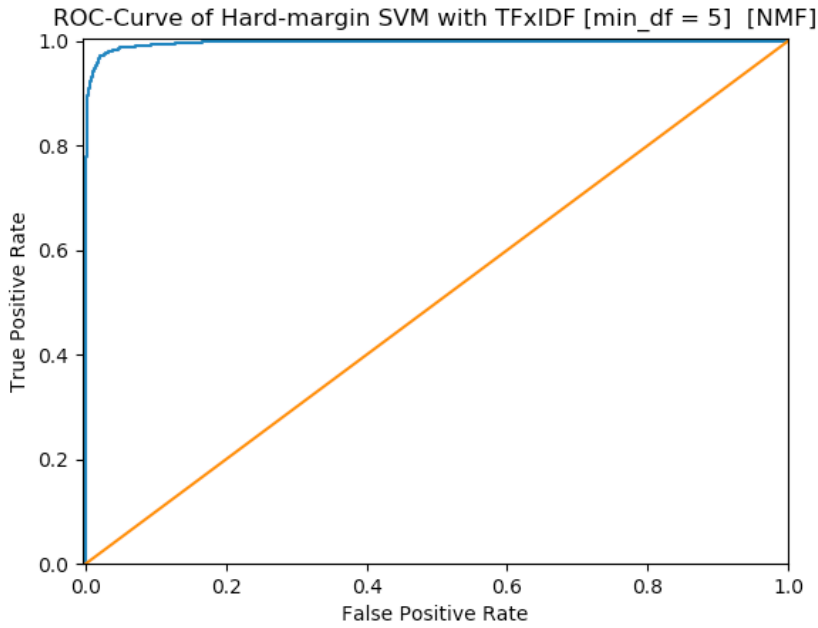
Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

	TP	FP	FN	TN
Computer technology	1533	53	27	1537
Recreational activity	1537	27	53	1533

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1533	53
Predicted negative	27	1537

2) Hard-Margin SVM with NMF



	Precision	Recall	Accuracy
Computer technology	0.9910	0.9173	0.9549
Recreational activity	0.9244	0.9918	0.9549
Average	0.9546	0.9577	0.9549

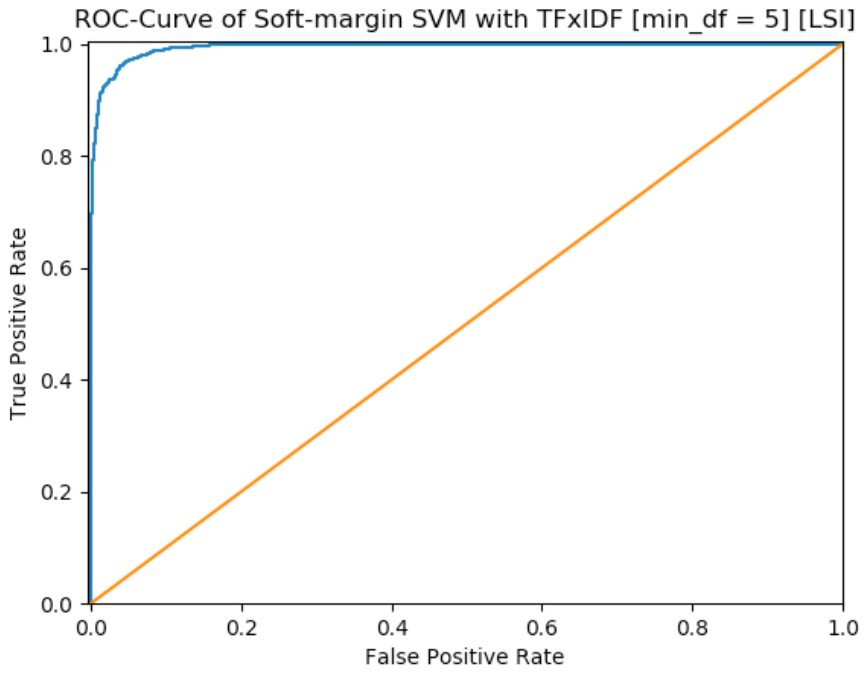
Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

	TP	FP	FN	TN
Computer technology	1431	13	129	1577
Recreational activity	1577	129	13	1431

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1431	13
Predicted negative	129	1577

3) Soft-Margin SVM with LSI



	Precision	Recall	Accuracy
Computer technology	0.9908	0.8981	0.9454
Recreational activity	0.9084	0.9918	0.9454
Average	0.9450	0.9496	0.9454

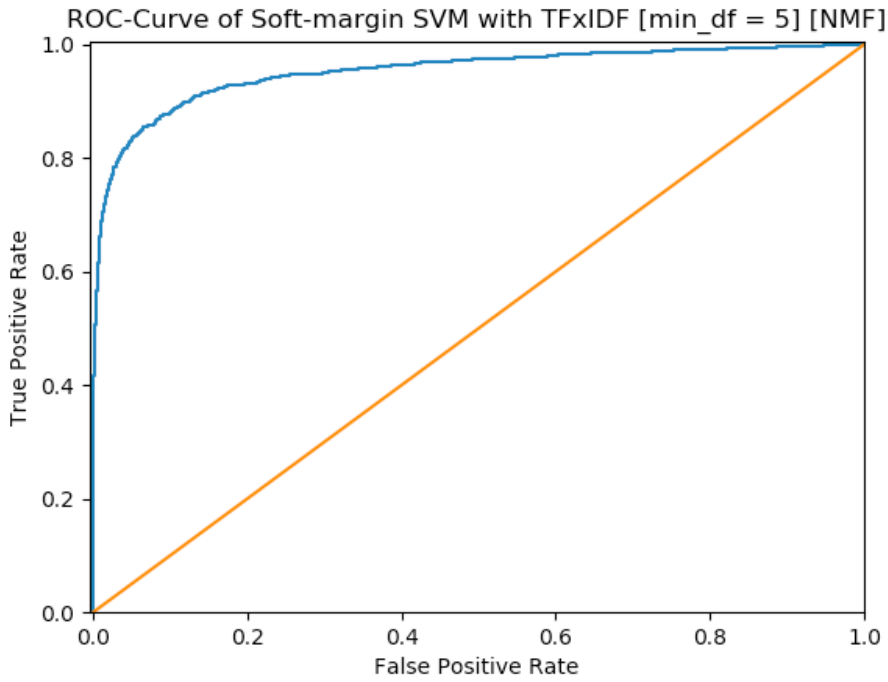
Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

	TP	FP	FN	TN
Computer technology	1401	13	159	1577
Recreational activity	1577	159	13	1401

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1401	13
Predicted negative	159	1577

4) Soft-Margin SVM with NMF



	Precision	Recall	Accuracy
Computer technology	0.9576	0.3910	0.6898
Recreational activity	0.6220	0.9830	0.6898
Average	0.6870	0.7898	0.6898

Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

	TP	FP	FN	TN
Computer technology	610	27	950	1563
Recreational activity	1563	950	27	610

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	610	27
Predicted negative	950	1563

Conclusion:

As we could see, in our model, when we choose min_df =5, the model's accuracy is better than the model with min_df = 2. When min_df is set to a constant, the LSI model is better than NMF and hard SVM is better than soft SVM.

6. Cross-validation to find the best γ

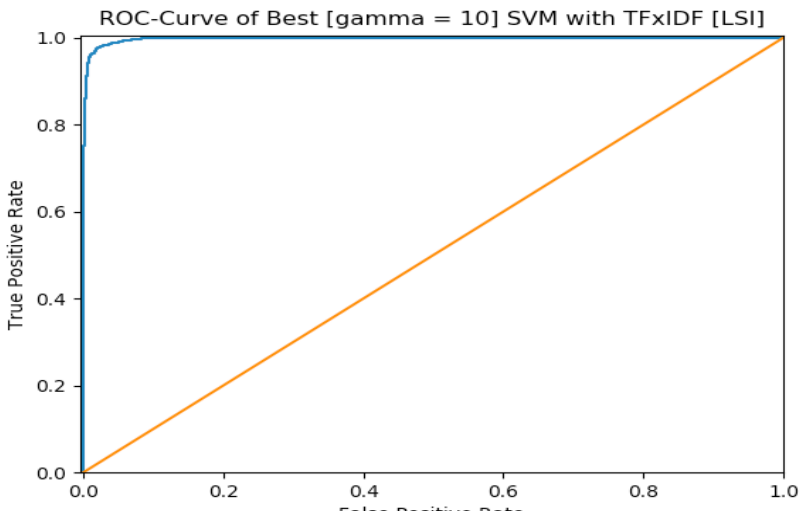
In this part, we use 5-fold cross-validation to find the parameter γ and report the confusion matrix and accuracy, recall and precision.

When Min_df=2

[gamma = 0.001] 5-Fold Average Accuracy: 0.94421304
[gamma = 0.01] 5-Fold Average Accuracy: 0.96745837
[gamma = 0.1] 5-Fold Average Accuracy: 0.96999537
[gamma = 1.0] 5-Fold Average Accuracy: 0.97591124
[gamma = 10] 5-Fold Average Accuracy: 0.97654459
[gamma = 100] 5-Fold Average Accuracy: 0.97569870
[gamma = 1000] 5-Fold Average Accuracy: 0.97379774

Best Accuracy is 0.97654459 when gamma = 10

SVM with TFxIDF [LSI] gamma = 10



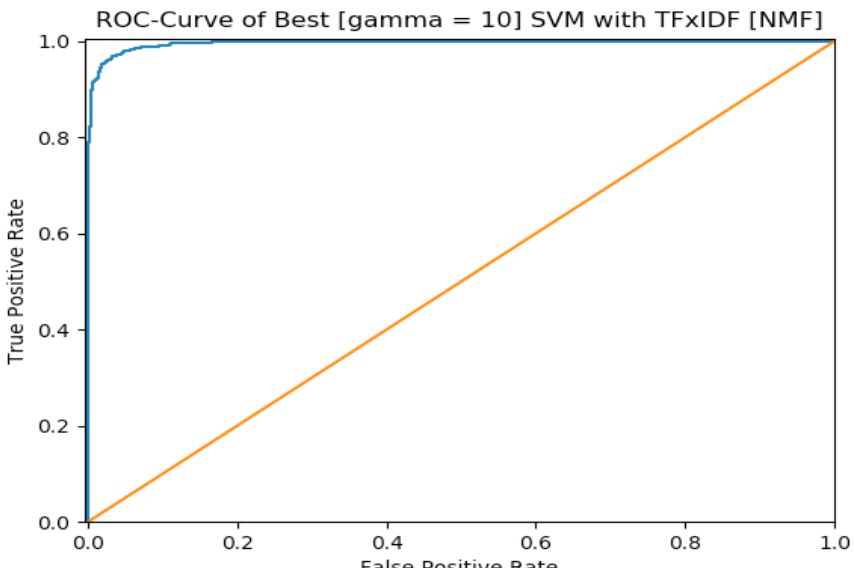
	Precision	Recall	Accuracy
Computer technology	0.9818	0.9679	0.9752
Recreational activity	0.9690	0.9824	0.9752
Average	0.9752	0.9754	0.9752

Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

	TP	FP	FN	TN
Computer technology	1510	28	50	1562
Recreational activity	1562	50	28	1510

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1510	28
Predicted negative	50	1562



SVM with TFxIDF [NMF] gamma = 10

	Precision	Recall	Accuracy
Computer technology	0.9738	0.9538	0.9644
Recreational activity	0.9556	0.9748	0.9644
Average	0.9643	0.9647	0.9644

Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

	TP	FP	FN	TN
Computer technology	1488	40	72	1550
Recreational activity	1550	72	40	1488

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1488	40
Predicted negative	72	1550

When min_df = 5

[gamma = 0.001] 5-Fold Average Accuracy: 0.94695967

[gamma = 0.01] 5-Fold Average Accuracy: 0.96640173

[gamma = 0.1] 5-Fold Average Accuracy: 0.97020656

[gamma = 1.0] 5-Fold Average Accuracy: 0.97506535

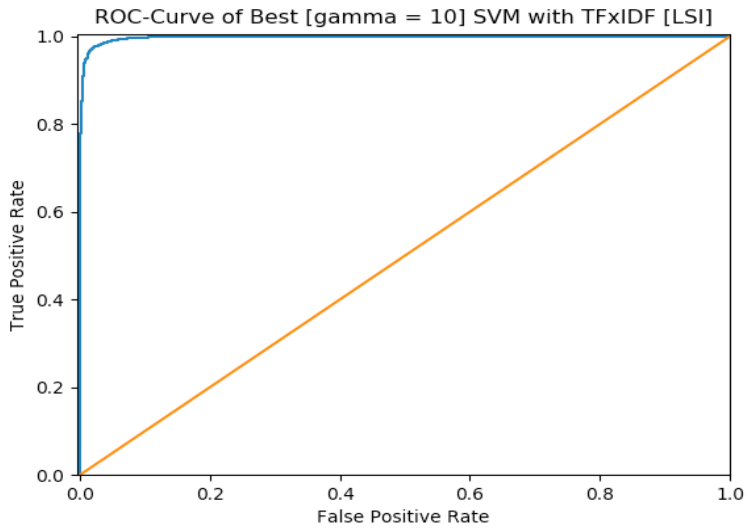
[gamma = 10] 5-Fold Average Accuracy: 0.97612198

[gamma = 100] 5-Fold Average Accuracy: 0.97548840

[gamma = 1000] 5-Fold Average Accuracy: 0.95838086

Best Accuracy is 0.97612198 when gamma = 10

SVM with TFxIDF [LSI]



	Precision	Recall	Accuracy
Computer technology	0.9812	0.9679	0.9749
Recreational activity	0.9690	0.9818	0.9749
Average	0.9749	0.9751	0.9749

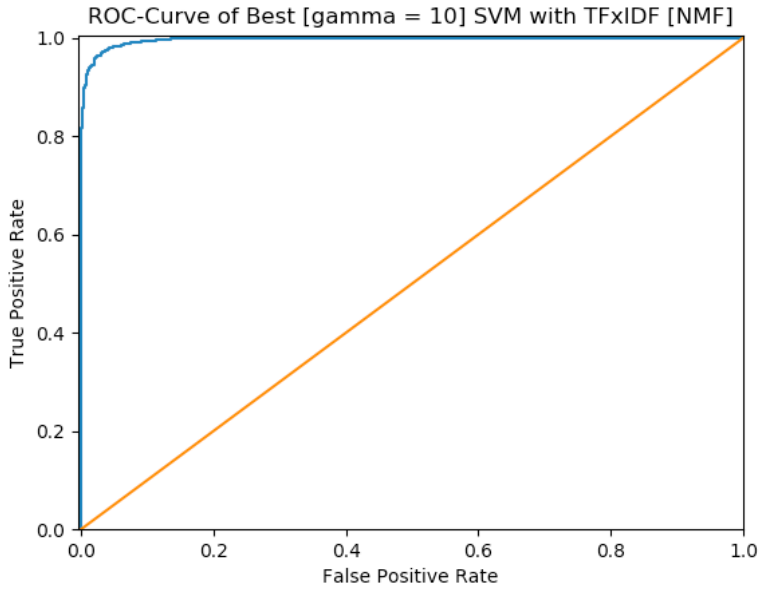
Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

	TP	FP	FN	TN
Computer technology	1510	29	50	1561
Recreational activity	1561	50	29	1510

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1510	29
Predicted negative	50	1561

Best gamma (gamma = 10) SVM with TFxIDF [NMF]



	Precision	Recall	Accuracy
Computer technology	0.9784	0.9564	0.9679
Recreational activity	0.9582	0.9792	0.9679
Average	0.9678	0.9683	0.9679

Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

	TP	FP	FN	TN
Computer technology	1492	33	68	1557

Recreational activity	1557	68	33	1492

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1492	33
Predicted negative	68	1557

Conclusion:

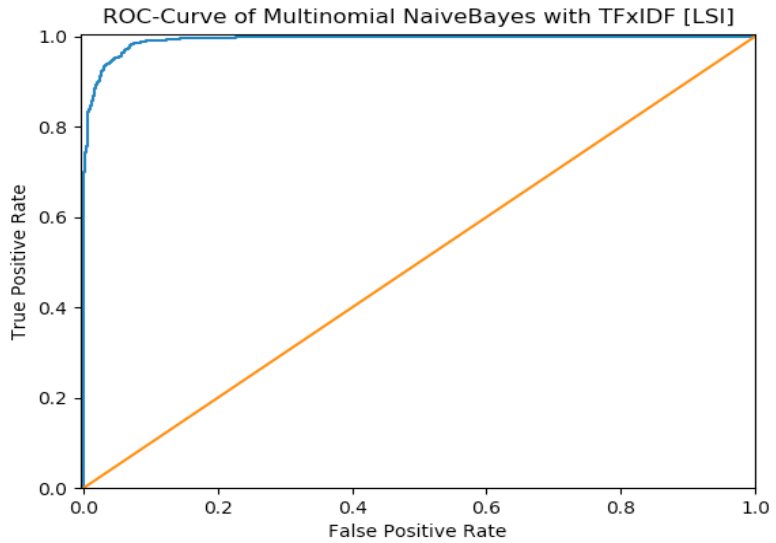
As we could see, in our model, when we choose min_df =5 , the model's accuracy is nearly equal to the model for min_df = 2. When we set min_df as a constant, the LSI model is slightly better than NMF

7. Naïve Bayes algorithm

Next, we use naïve Bayes algorithm for the same classification task.

When min_df=2

Naïve Bayes with LSI



	Precision	Recall	Accuracy
Computer technology	0.9927	0.8718	0.9333
Recreational activity	0.8876	0.9937	0.9333
Average	0.9328	0.9402	0.9333

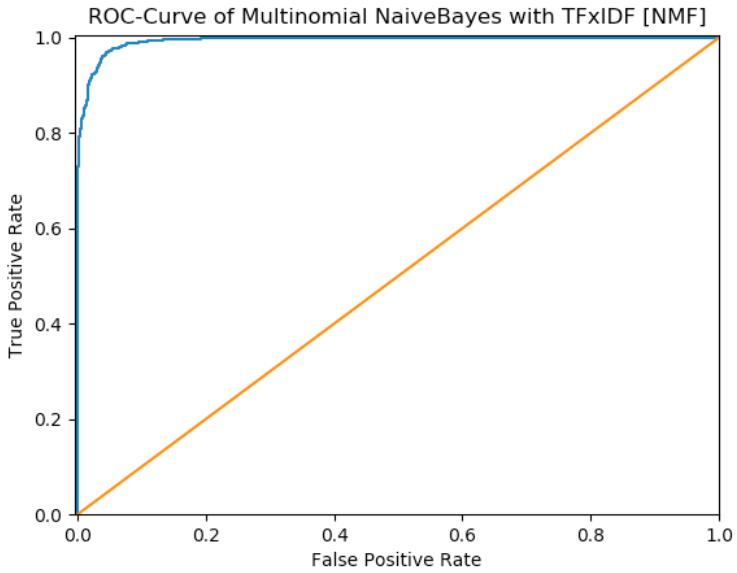
Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

	TP	FP	FN	TN
Computer technology	1360	10	200	1580
Recreational activity	1580	200	10	1360

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1360	10
Predicted negative	200	1580

Naïve Bayes with NMF



	Precision	Recall	Accuracy
Computer technology	0.9773	0.9397	0.9594
Recreational activity	0.9430	0.9786	0.9594
Average	0.9592	0.9602	0.9594

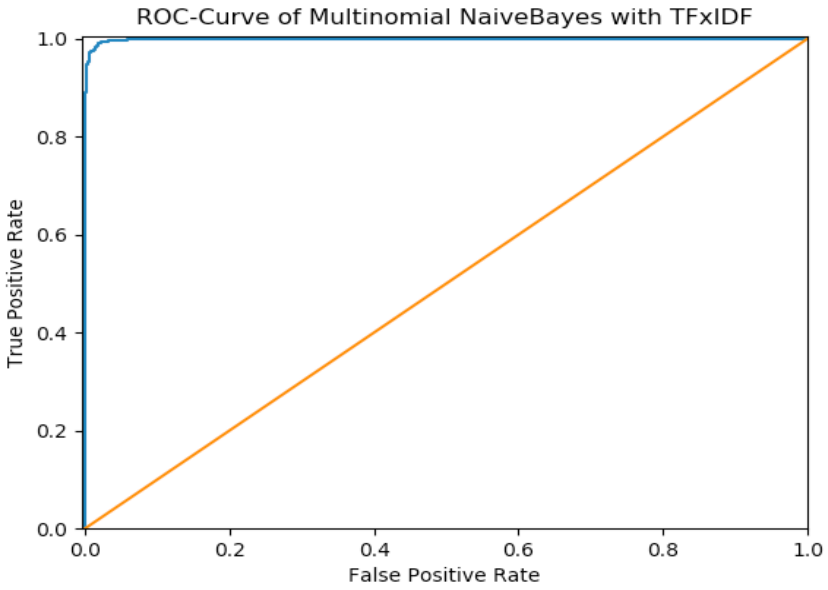
Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

	TP	FP	FN	TN
Computer technology	1466	34	94	1556
Recreational activity	1556	94	34	1466

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1466	34
Predicted negative	94	1556

Naïve Bayes with original TFxIDF



	Precision	Recall	Accuracy
Computer technology	0.9890	0.9814	0.9854
Recreational activity	0.9819	0.9893	0.9854
Average	0.9854	0.9855	0.9854

Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

	TP	FP	FN	TN
Computer technology	1531	17	29	1573

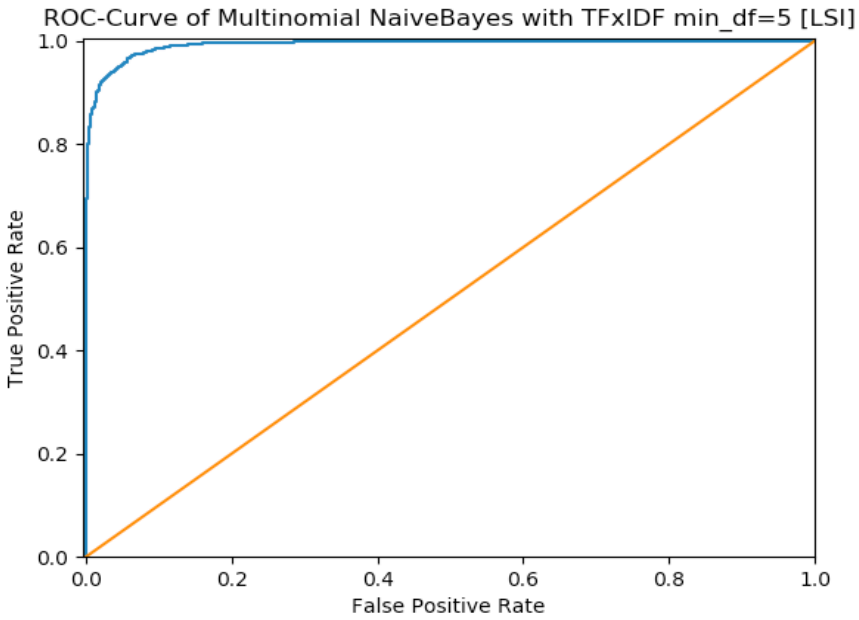
Recreational activity	1573	29	17	1531

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1531	17
Predicted negative	29	1573

When min_df=5

Naive Bayes with LSI



	Precision	Recall	Accuracy
Computer technology	0.9967	0.7686	0.8841
Recreational activity	0.8146	0.9975	0.8841
Average	0.8830	0.9056	0.8841

Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

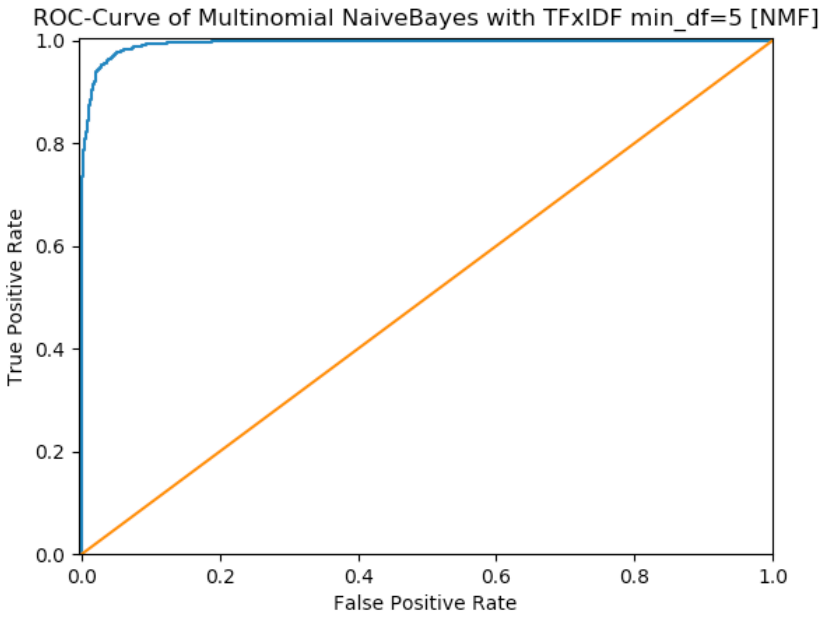
	TP	FP	FN	TN

Computer technology	1199	4	361	1586
Recreational activity	1586	361	4	1199

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1199	4
Predicted negative	361	1586

Naïve Bayes with NMF



	Precision	Recall	Accuracy
Computer technology	0.9811	0.9333	0.9581
Recreational activity	0.9376	0.9824	0.9581
Average	0.9579	0.9594	0.9581

Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

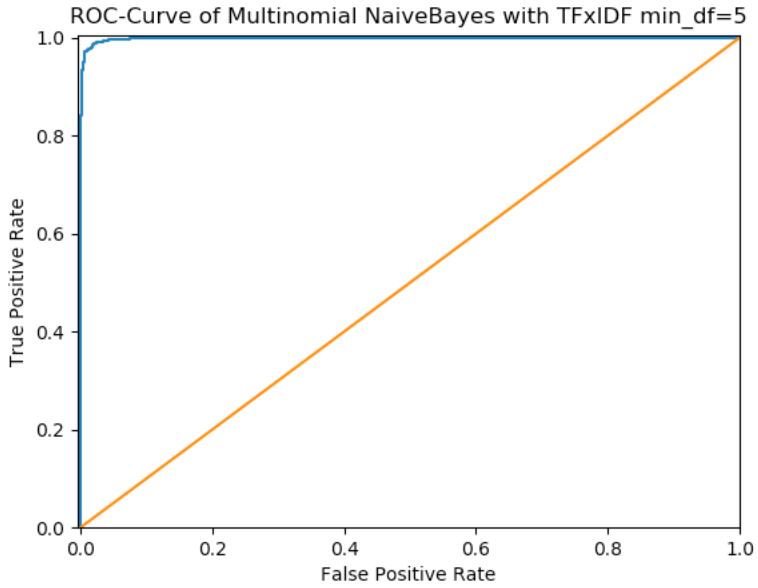
	TP	FP	FN	TN
Computer technology	1456	28	104	1562

Recreational activity	1562	104	28	1456

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1456	28
Predicted negative	104	1562

Naïve Bayes with Original TFxIDF



	Precision	Recall	Accuracy
Computer technology	0.9771	0.9840	0.9806
Recreational activity	0.9842	0.9774	0.9806
Average	0.9807	0.9806	0.9806

Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

	TP	FP	FN	TN
Computer technology	1535	36	25	1554
Recreational activity	1554	25	36	1535

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1535	36
Predicted negative	25	1554

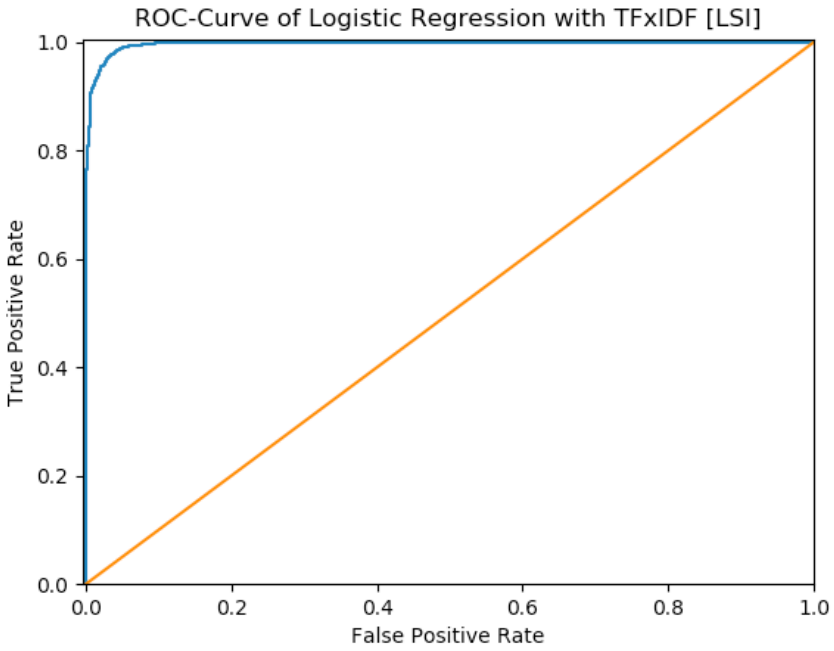
Conclusion:

As we could see here, in Bayes model, when we choose min_df =2, the model's accuracy is better than the model with min_df = 5. When we set min_df as a constant, the NMF model is better than LSI model.

8.logistic regression classifier

When min_df=2

Logistic Regression with LSI when min_df = 2



	Precision	Recall	Accuracy
Computer technology	0.9804	0.9603	0.9708
Recreational activity	0.9618	0.9811	0.9708
Average	0.9707	0.9711	0.9708

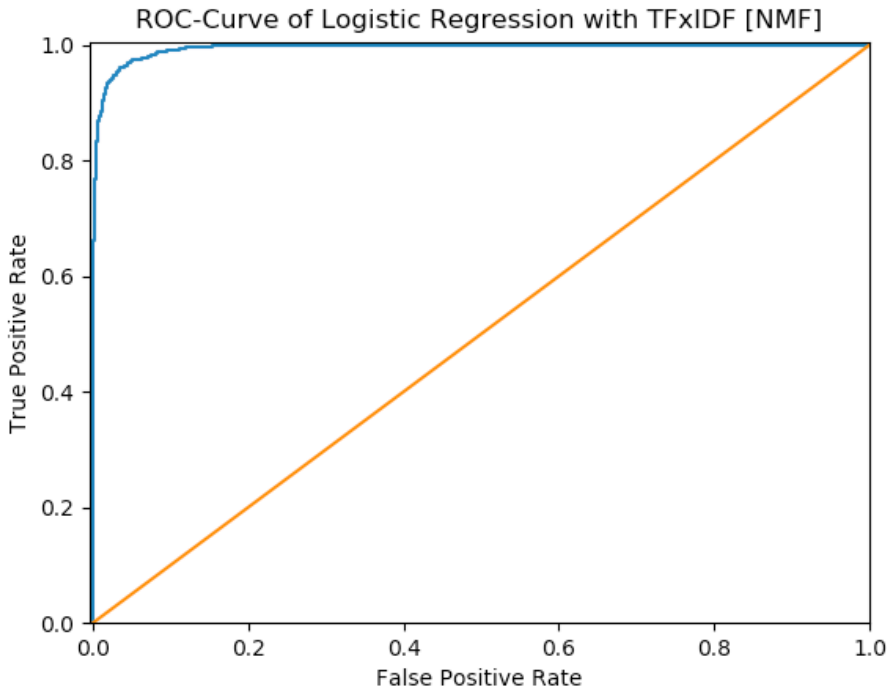
Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

	TP	FP	FN	TN
Computer technology	1498	30	62	1560
Recreational activity	1560	62	30	1498

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1498	30
Predicted negative	62	1560

Logistic Regression with NMF when min_df = 2



	Precision	Recall	Accuracy
Computer technology	0.9657	0.9551	0.9610
Recreational activity	0.9564	0.9667	0.9610
Average	0.9609	0.9610	0.9610

Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

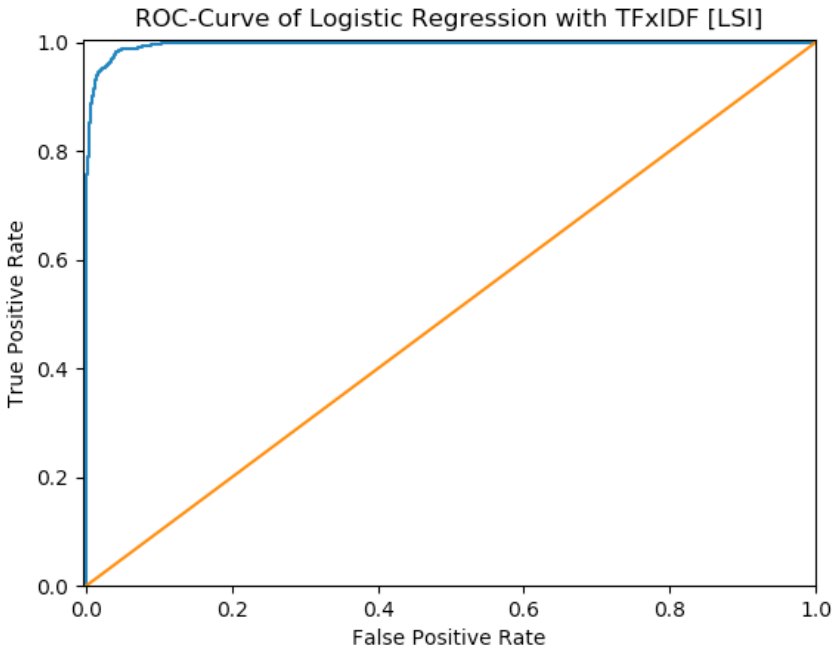
	TP	FP	FN	TN
Computer technology	1490	53	70	1573
Recreational activity	1537	70	53	1490

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1490	53
Predicted negative	70	1537

When min_df=5

Logistic Regression with LSI when min_df = 5



	Precision	Recall	Accuracy
Computer technology	0.9784	0.9596	0.9695
Recreational activity	0.9611	0.9792	0.9695
Average	0.9694	0.9698	0.9695

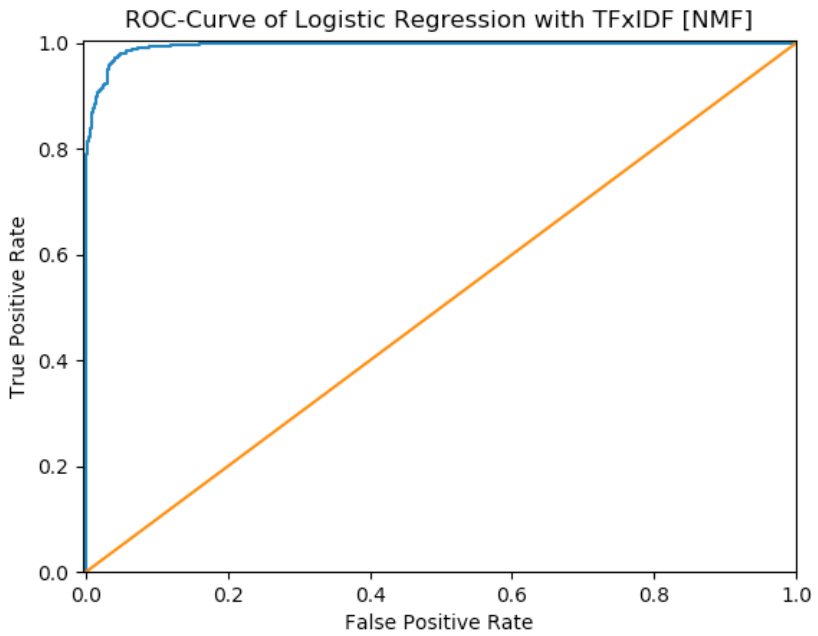
Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

	TP	FP	FN	TN
Computer technology	1497	33	63	1557
Recreational activity	1557	63	33	1497

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1497	33
Predicted negative	63	1557

Logistic Regression with NMF when min_df = 5



	Precision	Recall	Accuracy
Computer technology	0.9683	0.9603	0.9648
Recreational activity	0.9613	0.9692	0.9648
Average	0.9647	0.9648	0.9648

Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

	TP	FP	FN	TN
Computer technology	1498	49	62	1541
Recreational activity	1541	62	49	1498

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1498	49
Predicted negative	62	1541

Conclusion:

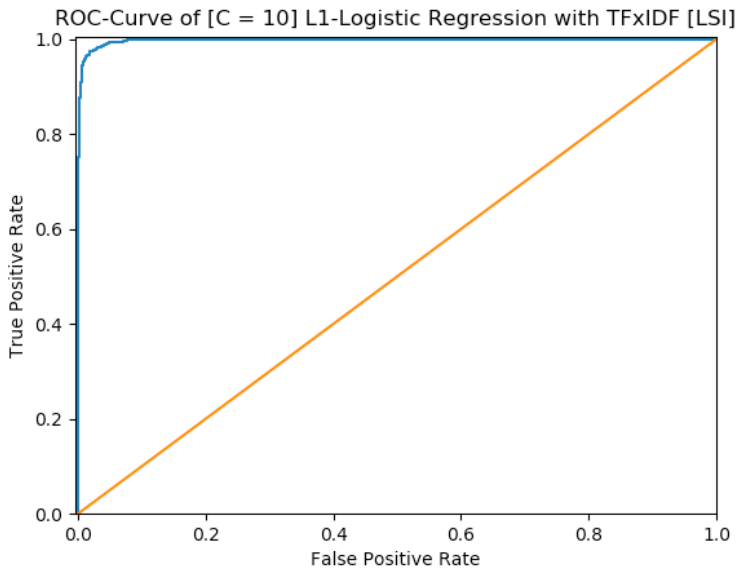
As we could see, in our model, when we choose min_df =2 , the model's accuracy is better than the model for min_df = 5. When we treat min_df as a constant, the LSI model is better than NMF .

9. Optimize model (Regularization)

Now, by adding a regularization term to the optimization objective. We try both L1 and L2 norm regularizations and sweep through different regularization coefficients, ranging from very small ones to large ones.

C from the 0.001 to 10, and we choice the best score to output

L1-Logistic Regression with LSI when min_df = 2



	Precision	Recall	Accuracy
Computer technology	0.9817	0.9654	0.9740
Recreational activity	0.9666	0.9824	0.9740
Average	0.9739	0.9742	0.9740

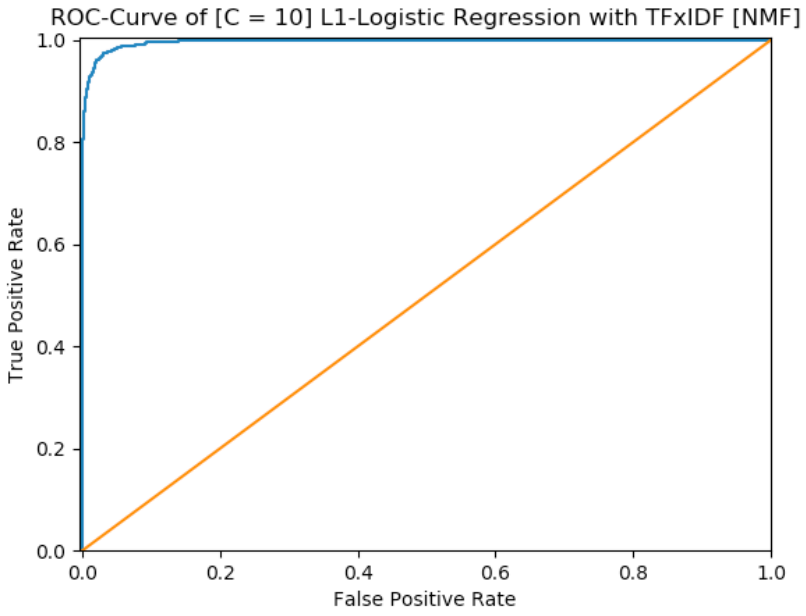
Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

	TP	FP	FN	TN
Computer technology	1506	28	54	1562
Recreational activity	1562	54	28	1506

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1562	28
Predicted negative	54	1562

L1 Logistic Regression with NMF when min_df = 2



	Precision	Recall	Accuracy
Computer technology	0.9790	0.9551	0.9676
Recreational activity	0.9570	0.9799	0.9676
Average	0.9675	0.9680	0.9676

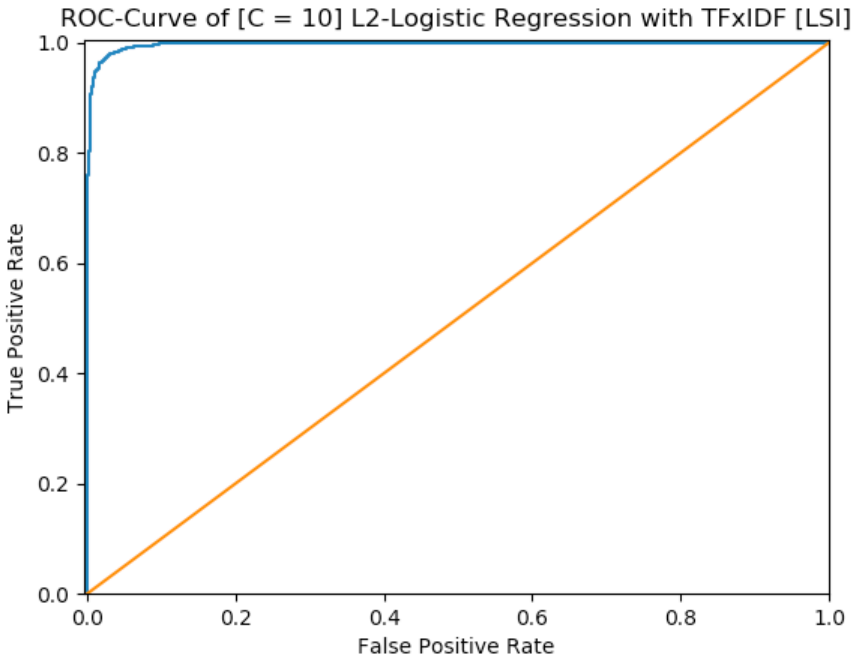
Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

	TP	FP	FN	TN
Computer technology	1490	32	70	1558
Recreational activity	1558	70	32	1490

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1490	32
Predicted negative	70	1558

L2 Logistic Regression with LSI when min_df = 2



	Precision	Recall	Accuracy
Computer technology	0.9817	0.9635	0.9730
Recreational activity	0.9648	0.9824	0.9730
Average	0.9729	0.9733	0.9730

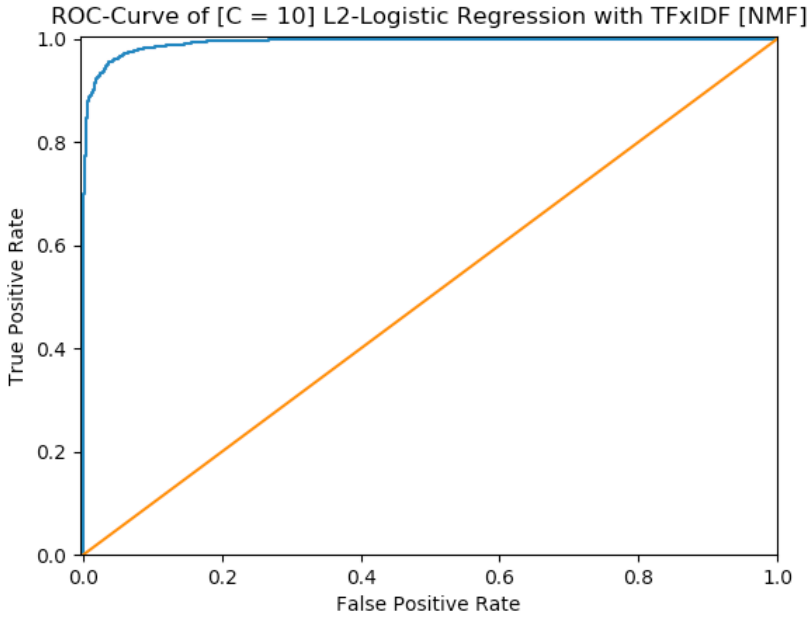
Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

	TP	FP	FN	TN
Computer technology	1503	28	57	1562
Recreational activity	1562	57	28	1503

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1503	28
Predicted negative	57	1562

L2 Logistic Regression with NMF when min_df = 2



	Precision	Recall	Accuracy

Computer technology	0.9659	0.9455	0.9565
Recreational activity	0.9476	0.9673	0.9565
Average	0.9564	0.9568	0.9565

Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

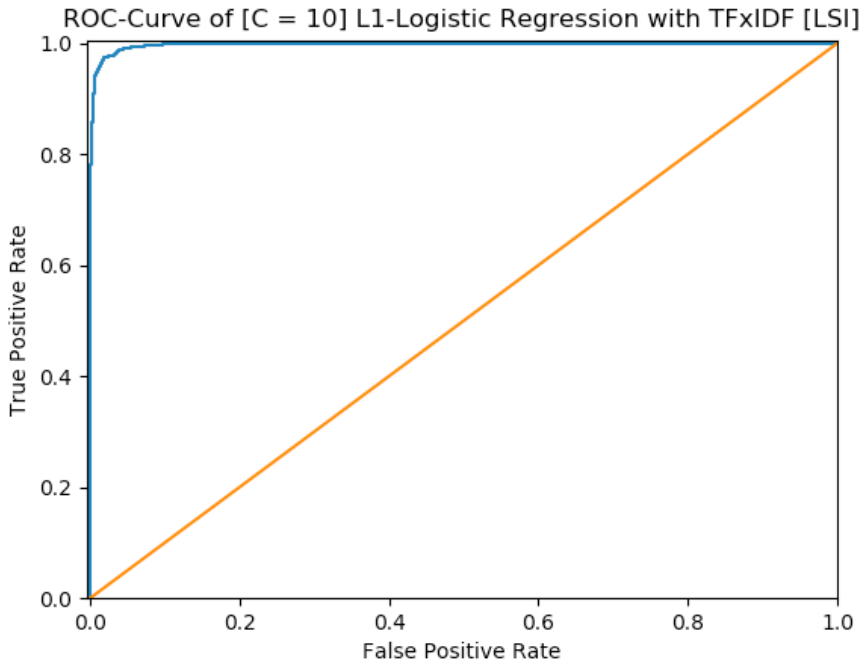
	TP	FP	FN	TN
Computer technology	1475	52	85	1538
Recreational activity	1538	85	52	1475

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1475	52
Predicted negative	85	1538

When min_df=5

L1 Logistic Regression with LSI min_df = 5



	Precision	Recall	Accuracy

Computer technology	0.9798	0.9654	0.9730
Recreational activity	0.9665	0.9805	0.9730
Average	0.9729	0.9732	0.9730

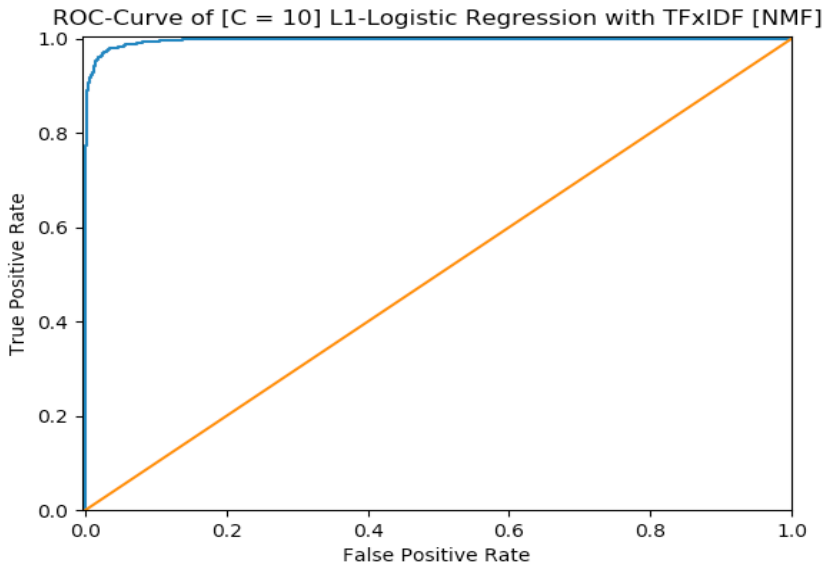
Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

	TP	FP	FN	TN
Computer technology	1506	31	54	1559
Recreational activity	1559	54	31	1506

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1506	31
Predicted negative	54	1559

L1 Logistic Regression with NMF when min_df = 5



	Precision	Recall	Accuracy
Computer technology	0.9810	0.9603	0.9711
Recreational activity	0.9618	0.9818	0.9711

Average	0.9710	0.9714	0.9711

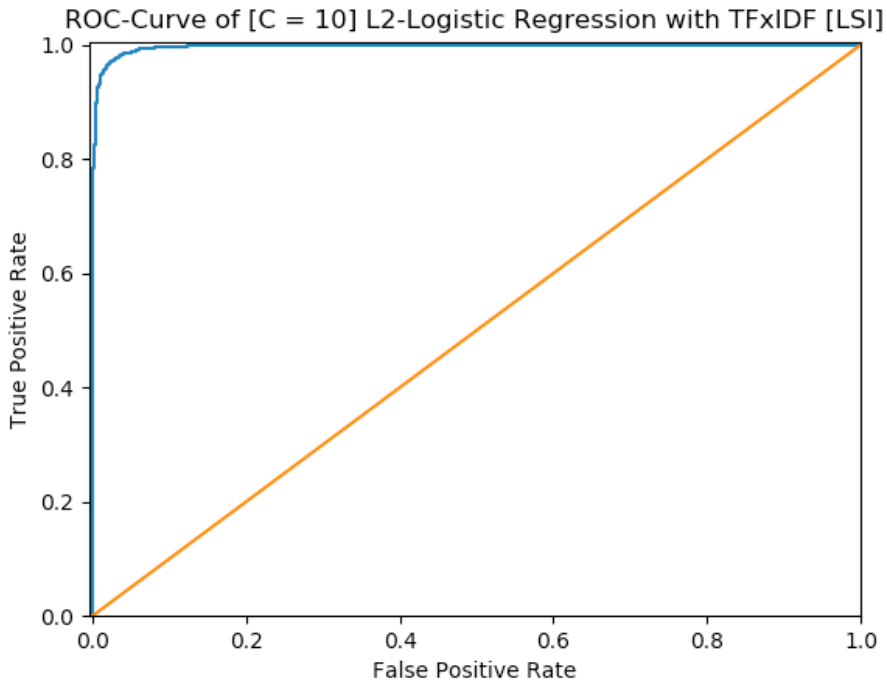
Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

	TP	FP	FN	TN
Computer technology	1498	29	62	1561
Recreational activity	1561	62	29	1498

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1498	29
Predicted negative	62	1561

L2 Logistic regression with LSI when min_df = 5



	Precision	Recall	Accuracy
Computer technology	0.9811	0.9647	0.9733

Recreational activity	0.9660	0.9818	0.9733
Average	0.9733	0.9735	0.9733

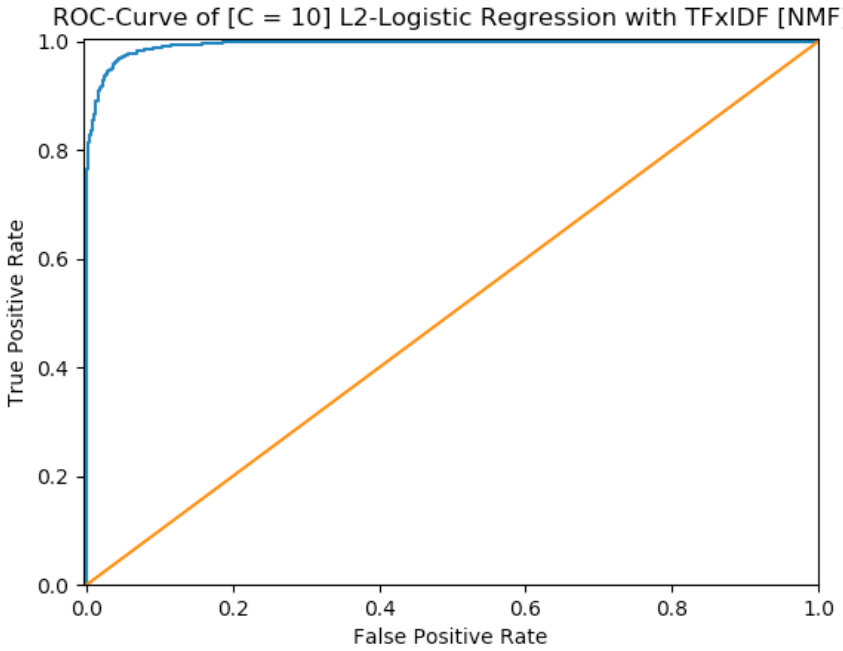
Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

	TP	FP	FN	TN
Computer technology	1505	29	55	1561
Recreational activity	1561	55	29	1505

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1505	29
Predicted negative	55	1561

L2 Logistic regression with NMF when min_df = 5



	Precision	Recall	Accuracy
Computer technology	0.9711	0.9487	0.9606
Recreational activity	0.9508	0.9732	0.9606

Average	0.9605	0.9610	0.9606

Confusion Matrix is as following (at Computer Technology row we treat Computer Technology as positive, vice versa):

	TP	FP	FN	TN
Computer technology	1480	44	80	1546
Recreational activity	1546	80	44	1480

We set computer technology as positive and recreational activity as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	1480	44
Predicted negative	80	1546

Conclusion:

In summary, l1 norm and l2 norm can get similar best accuracy and there is no fixed best combination of parameters. No matter which regularization method we use, smaller C tends to perform worse. Parameter ‘min_df = 2’ tend to get better results than ‘min_df = 5’ in most classifiers, especially the number of true positive and true negative. It is probably because ‘min_df = 5’ removes so much information from the documents. And, LSI method is a little better than NMF.

10. Multi-class Classification

In this part, we aim to perform Naïve Bayes classification and multiclass SVM classification (with both One VS One and One VS the rest methods described above) and report the confusion matrix and calculate the accuracy, recall and precision of your classifiers

Multinomial NaiveBayes with TFIDF when min_df is 2 [LSI]

	Precision	Recall	Accuracy
comp.sys.ibm.pc.hardware	0.7717	0.6811	0.8696
comp.sys.mac.hardware	0.8177	0.7688	0.9010
misc.forsale	0.7131	0.8667	0.8799

soc.religion.christian	0.9765	0.9397	0.9789
Average	0.8141	0.8197	0.9073

Confusion Matrix is as following (at comp.sys.ibm.pc.hardware row we treat comp.sys.ibm.pc. hardware as positive, vice versa):

	TP	FP	FN	TN
comp.sys.ibm.pc.hardware	267	79	125	1094
comp.sys.mac.hardware	296	66	89	1114
misc.forsale	338	136	52	1039
soc.religion.christian	374	9	24	1158

We set comp.sys.ibm.pc.hardware as positive and others as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	267	79
Predicted negative	125	1094

Multinomial NaiveBayes with TFIDF when min_df is 2 [NMF]

	Precision	Recall	Accuracy
comp.sys.ibm.pc.hardware	0.8054	0.7602	0.8939
comp.sys.mac.hardware	0.8226	0.7948	0.9073
misc.forsale	0.8000	0.8615	0.9118
soc.religion.christian	0.9752	0.9874	0.9904
Average	0.8510	0.8508	0.9259

Confusion Matrix is as following (at comp.sys.ibm.pc.hardware row we treat comp.sys.ibm.pc. hardware as positive, vice versa):

	TP	FP	FN	TN
comp.sys.ibm.pc.hardware	298	72	94	1101
comp.sys.mac.hardware	306	66	79	1114

misc.forsale	336	84	54	1091
soc.religion.christian	393	10	5	1157

We set comp.sys.ibm.pc.hardware as positive and others as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	336	84
Predicted negative	54	1091

Multinomial NaiveBayes with TFIDF when min_df is 5 [LSI]

	Precision	Recall	Accuracy
comp.sys.ibm.pc.hardware	0.7237	0.8087	0.8748
comp.sys.mac.hardware	0.9022	0.6468	0.8958
misc.forsale	0.8739	0.7821	0.9176
soc.religion.christian	0.7928	1.0000	0.9335
Average	0.8094	0.8232	0.9054

Confusion Matrix is as following (at comp.sys.ibm.pc.hardware row we treat comp.sys.ibm.pc. hardware as positive, vice versa):

	TP	FP	FN	TN
comp.sys.ibm.pc.hardware	317	121	75	1052
comp.sys.mac.hardware	249	27	136	1153
misc.forsale	305	44	85	1131
soc.religion.christian	398	104	0	1063

We set comp.sys.ibm.pc.hardware as positive and others as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	317	121
Predicted negative	75	1052

Multinomial NaiveBayes with TFIDF when min_df is 5 [NMF]

	Precision	Recall	Accuracy
comp.sys.ibm.pc.hardware	0.7558	0.7500	0.8767
comp.sys.mac.hardware	0.7836	0.7714	0.8914
misc.forsale	0.8321	0.8513	0.9201
soc.religion.christian	0.9824	0.9824	0.9911
Average	0.8388	0.8385	0.9198

Confusion Matrix is as following (at comp.sys.ibm.pc.hardware row we treat comp.sys.ibm.pc. hardware as positive, vice versa):

	TP	FP	FN	TN
comp.sys.ibm.pc.hardware	294	95	98	1078
comp.sys.mac.hardware	297	82	88	1198
misc.forsale	332	67	58	1108
soc.religion.christian	391	7	7	1160

We set comp.sys.ibm.pc.hardware as positive and others as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	294	95
Predicted negative	98	1078

Multi-calss SVM ovo with TFIDF when min_df is 5 [LSI]

	Precision	Recall	Accuracy
comp.sys.ibm.pc.hardware	0.6198	0.8776	0.8345
comp.sys.mac.hardware	0.9462	0.4571	0.8601
misc.forsale	0.7366	0.8821	0.8920
soc.religion.christian	1.0000	0.8970	0.9738

Average	0.7784	0.8257	0.8901

Confusion Matrix is as following (at comp.sys.ibm.pc.hardware row we treat comp.sys.ibm.pc. hardware as positive, vice versa):

	TP	FP	FN	TN
comp.sys.ibm.pc.hardware	344	211	48	962
comp.sys.mac.hardware	176	10	209	1170
misc.forsale	344	123	46	1052
soc.religion.christian	357	0	41	1167

We set comp.sys.ibm.pc.hardware as positive and others as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	344	211
Predicted negative	48	962

Multi-class SVM ovr with TFIDF when min_df is 5 [LSI]

	Precision	Recall	Accuracy
comp.sys.ibm.pc.hardware	0.7738	0.8903	0.9073
comp.sys.mac.hardware	0.8839	0.9104	0.9272
misc.forsale	0.9055	0.8846	0.9482
soc.religion.christian	0.9974	0.9523	0.9872
Average	0.8844	0.8901	0.9425

Confusion Matrix is as following (at comp.sys.ibm.pc.hardware row we treat comp.sys.ibm.pc. hardware as positive, vice versa):

	TP	FP	FN	TN
comp.sys.ibm.pc.hardware	349	102	43	1071
comp.sys.mac.hardware	312	41	73	1139

misc.forsale	345	36	45	1139
soc.religion.christian	379	1	19	1166

We set comp.sys.ibm.pc.hardware as positive and others as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	349	102
Predicted negative	43	1071

Multi-class SVM ovr with TFIDF when min_df is 5 [NMF]

	Precision	Recall	Accuracy
comp.sys.ibm.pc.hardware	0.6752	0.8750	0.8633
comp.sys.mac.hardware	0.8448	0.7351	0.9016
misc.forsale	0.9265	0.8077	0.9361
soc.religion.christian	0.9895	0.9497	0.9847
Average	0.8419	0.8590	0.9214

Confusion Matrix is as following (at comp.sys.ibm.pc.hardware row we treat comp.sys.ibm.pc. hardware as positive, vice versa):

	TP	FP	FN	TN
comp.sys.ibm.pc.hardware	343	165	49	1008
comp.sys.mac.hardware	283	52	102	1128
misc.forsale	315	25	75	1150
soc.religion.christian	378	4	20	1163

We set comp.sys.ibm.pc.hardware as positive and others as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	343	165

Predicted negative	49	1008

Multi-calss SVM ovo with TFIDF when min_df is 2 [LSI]

	Precision	Recall	Accuracy
comp.sys.ibm.pc.hardware	0.5470	0.8469	0.7859
comp.sys.mac.hardware	1.0000	0.1429	0.7819
misc.forsale	0.6215	0.9051	0.8390
soc.religion.christian	1.0000	0.8417	0.9597
Average	0.6842	0.7921	0.8435

Confusion Matrix is as following (at comp.sys.ibm.pc.hardware row we treat comp.sys.ibm.pc. hardware as positive, vice versa):

	TP	FP	FN	TN
comp.sys.ibm.pc.hardware	332	275	60	898
comp.sys.mac.hardware	55	0	330	1180
misc.forsale	353	215	37	960
soc.religion.christian	335	0	63	1167

We set comp.sys.ibm.pc.hardware as positive and others as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	332	275
Predicted negative	60	898

Multi-class SVM ovr with TFIDF when min_df is 2 [LSI]

	Precision	Recall	Accuracy
comp.sys.ibm.pc.hardware	0.7795	0.8929	0.9099
comp.sys.mac.hardware	0.8852	0.8208	0.9297

misc.forsale	0.9008	0.8846	0.9470
soc.religion.christian	0.9973	0.9422	0.9847
Average	0.8851	0.8907	0.9428

Confusion Matrix is as following (at comp.sys.ibm.pc.hardware row we treat comp.sys.ibm.pc. hardware as positive, vice versa):

	TP	FP	FN	TN
comp.sys.ibm.pc.hardware	350	99	42	1074
comp.sys.mac.hardware	316	41	69	1139
misc.forsale	345	38	45	1137
soc.religion.christian	375	1	23	1166

We set comp.sys.ibm.pc.hardware as positive and others as negative to construct the confusion matrix as follow:

	Ground Truth	Ground False
Predicted positive	350	99
Predicted negative	42	1074

Multi-class SVM ovr with TFIDF when min_df is 2 [NMF]

	Precision	Recall	Accuracy
comp.sys.ibm.pc.hardware	0.7119	0.8699	0.8792
comp.sys.mac.hardware	0.8497	0.8078	0.9176
misc.forsale	0.9188	0.8128	0.9355
soc.religion.christian	0.9973	0.9397	0.9840
Average	0.8576	0.8695	0.9291

Confusion Matrix is as following (at comp.sys.ibm.pc.hardware row we treat comp.sys.ibm.pc. hardware as positive, vice versa):

	TP	FP	FN	TN

comp.sys.ibm.pc.hardware	341	138	51	1035
comp.sys.mac.hardware	311	55	74	1125
misc.forsale	317	28	73	1147
soc.religion.christian	374	1	24	1166

Confusion Matrix is as following (at comp.sys.ibm.pc.hardware row we treat comp.sys.ibm.pc. hardware as positive, vice versa):

	Ground Truth	Ground False
Predicted positive	341	138
Predicted negative	51	1035

Conclusion:

In conclusion, generally in multiclass classification tasks, the performance of SVM (both “One VS One” and “One VS Rest”) and Naïve Bayes is different in different cases. In Bayes, the accuracy in min_df =2 and 5 has a little different but generally equal. Comparing two SVM multiclass classification methods, “One VS Rest” runs faster than “One VS One” and gets a better result according to accuracy, precision and recall. LSI method is better than NMF and when min_df is 5, the accuracy is better.