The University of California, Los Angeles
Department of Electrical and Computer Engineering
ECE219 Large-scale Data Mining, Winter 2018

Project 1 Classification analysis on textual data

Xin Jiang (904589261), Zhiyuan Cao

Deadline: January 29th, 2018

Question (a) Figure 1 shows that the number of documents in each subclass is almost the same. So we get a balanced dataset to train our model.

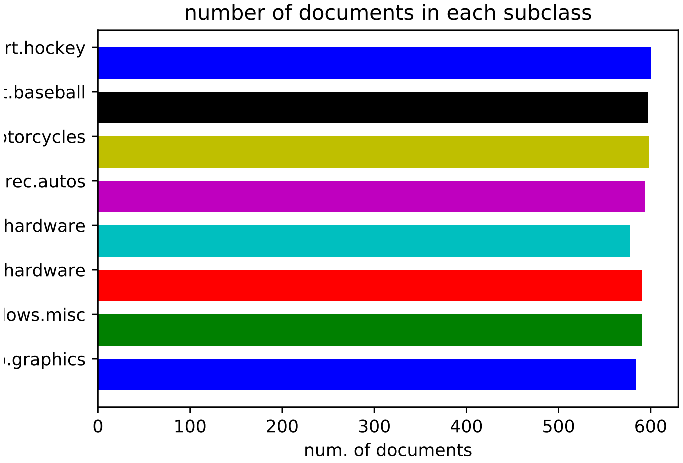

Figure 1: Number of documents in each subclass.

Question (b) With `min_df=2`, the total number of terms is 25207. There are 4732 documents in the training set, while 3150 in the test set.
With `min_df=5`, the total number of terms is
Starting from the next question, we only report the result with `mid_df=2`, and we kindly ask the grader to refer to the attached jupter notebook for results with `min_df=5`.

Question (c) The 10 most significant terms in the given four categories are shown in the following table.

`comp.sys.ibm.pc.hardware`	`comp.sys.mac.hardware`	`misc.forsale`	`soc.religion.christian`
drive	edu	1	s
scsi	line	edu	god
edu	s	2	christian
1	mac	00	t
s	subject	line	edu
2	organ	subject	christ
use	t	organ	church
line	sale	jesus	
com	use	3	subject
subject	apple	5	people

Question (d) Nothing to report in this part.

Question (e) Figure 2 shows the ROC curves and the confusion matrices of the hard-margin SVMs with data derived from LSI and NMF, respectively. The results show that the SVMs can achieve high true positive rates with tolerable false positive rate.

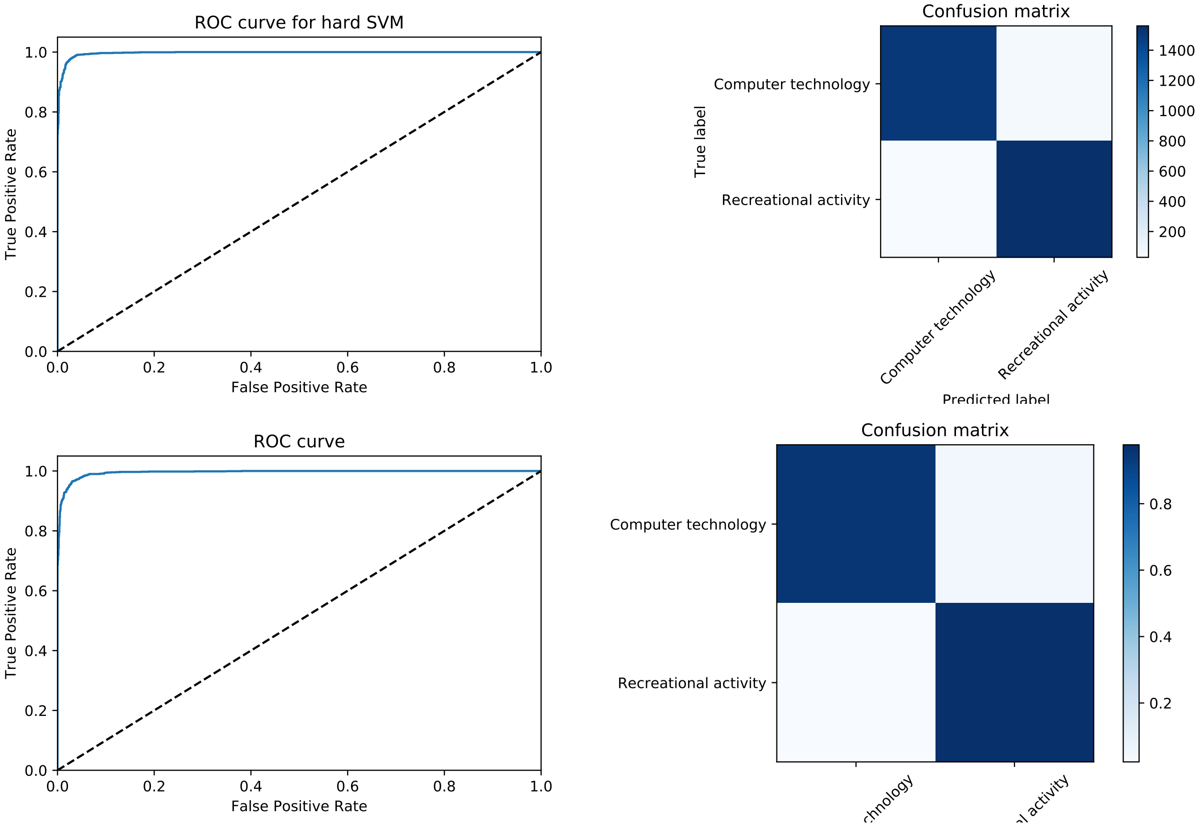

Figure 2: *Above.* The ROC curve and confusion matrix of the hard-margin SVM with LSI data. *Below.* The ROC curve and confusion matrix of the hard-margin SVM with NMF data.

The detailed statistics are shown in the following tables.

	predicted label=0	predicted label=1
actual label=0	0.97	0.03
actual label=1	0.02	0.98

Table 1: Confusion matrix for hard-margin SVM with LSI data.

	precision	recall	f1-score	support
Computer tech.	0.98	0.97	0.97	1560
Recreational act.	0.97	0.98	0.98	1590
avg / total	0.98	0.98	0.98	3150

Table 2: Detailed statistics for hard margin SVM with LSI data.

	predicted label=0	predicted label=1
actual label=0	0.95	0.05
actual label=1	0.02	0.98

Table 3: Confusion matrix for hard-margin SVM with NMF data.

	precision	recall	f1-score	support
Computer tech.	0.98	0.95	0.96	1560
Recreational act.	0.95	0.98	0.97	1590
avg / total	0.97	0.96	0.96	3150

Table 4: Detailed statistics for hard margin SVM with NMF data.

The results are not that good for soft-margin SVMs. This is partly because the training data and the test data have the same distribution, so that the small penalty does not help in classification.
The detailed statistics are shown in the following tables.

	predicted label=0	predicted label=1
actual label=0	0.00	1.00
actual label=1	0.00	1.00

Table 5: Confusion matrix for soft-margin SVM with LSI data.

	precision	recall	f1-score	support
Computer tech.	0.00	0.00	0.00	1560
Recreational act.	0.50	1.00	0.67	1590
avg / total	0.25	0.50	0.34	3150

Table 6: Detailed statistics for soft margin SVM with LSI data.

	predicted label=0	predicted label=1
actual label=0	0.00	1.00
actual label=1	0.00	1.00

Table 7: Confusion matrix for soft-margin SVM with NMF data.

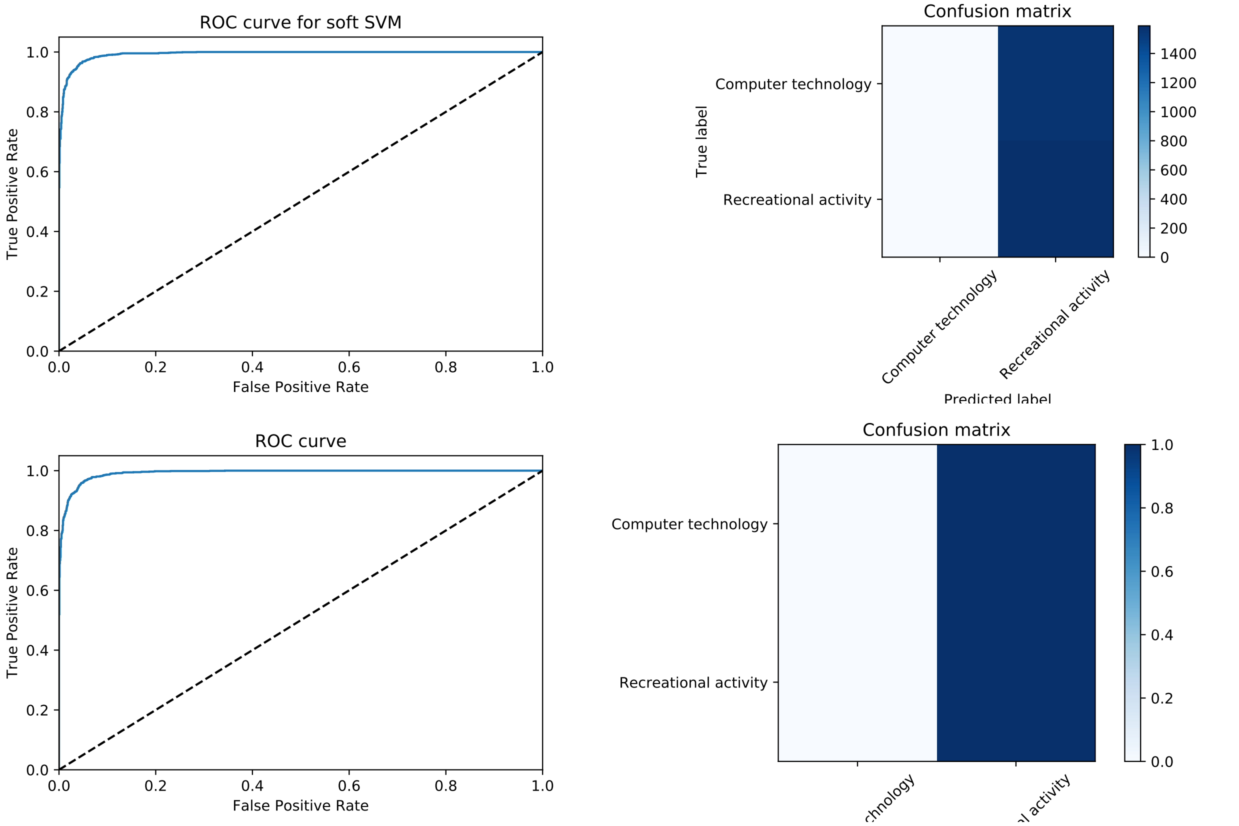

Figure 3: *Above.* The ROC curve and confusion matrix of the soft-margin SVM with LSI data. *Below.* The ROC curve and confusion matrix of the soft-margin SVM with NMF data.

	precision	recall	f1-score	support
Computer tech.	0.00	0.00	0.00	1560
Recreational act.	.0.50	1.00	0.67	1590
avg / total	0.25	0.50	0.34	3150

Table 8: Detailed statistics for soft margin SVM with NMF data.

Question (f) We report the accuracy result for different penalty terms in the soft-margin SVMs in the following table. In our experiment, $k = 1000$ is the best choice with both LSI and NMF data. Figure 4 shows the ROC curves and confusion matrices with the best choice.

penalty	0.001	0.01	0.1	1	10	100	1000
accuracy	0.4577	0.4577	0.4578	0.9344	0.9672	0.9714	0.9598

The detailed statistics are shown in the following tables.

	predicted label=0	predicted label=1
actual label=0	0.97	0.03
actual label=1	0.02	0.98

Table 9: Confusion matrix for best soft-margin SVM with LSI data.

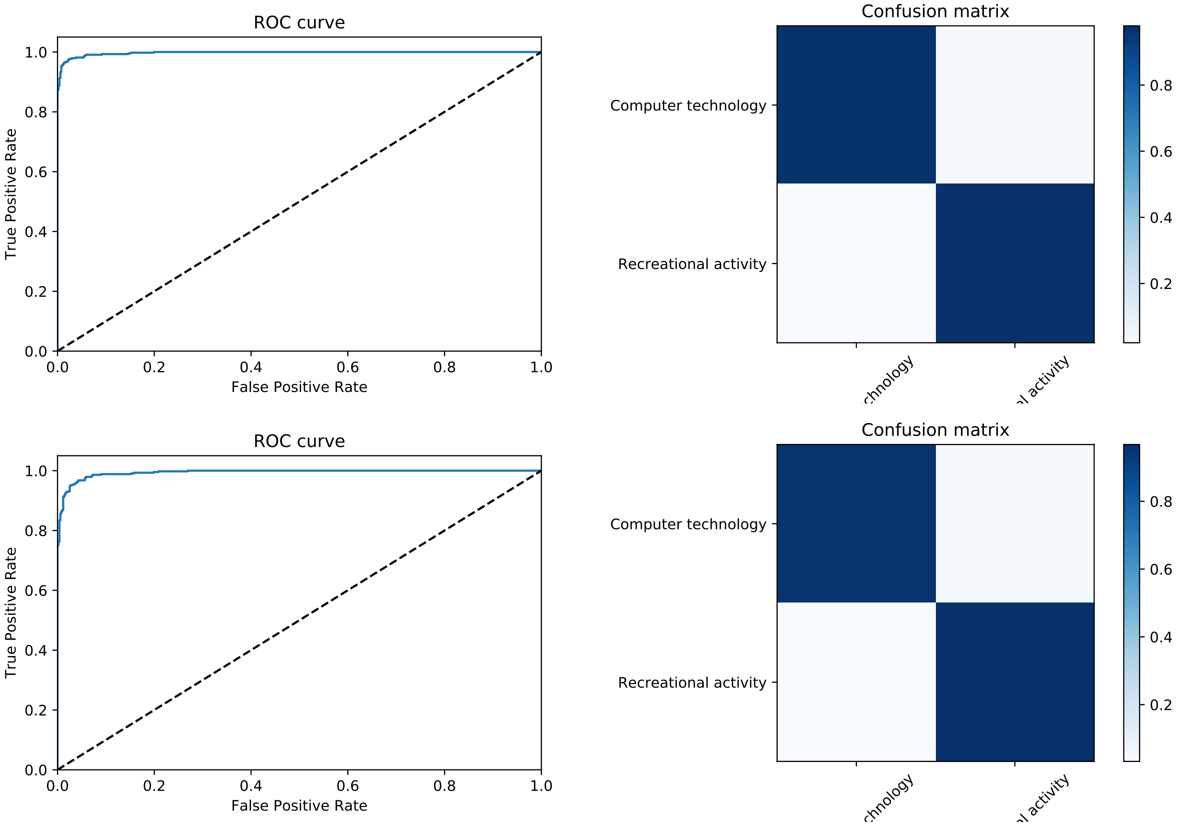

Figure 4: *Above.* The ROC curve and confusion matrix of the best soft-margin SVM with LSI data. *Below.* The ROC curve and confusion matrix of the best soft-margin SVM with NMF data.

	precision	recall	f1-score	support
Computer tech.	0.98	0.97	0.97	513
Recreational act.	0.97	0.98	0.97	433
avg / total	0.97	0.97	0.97	946

Table 10: Detailed statistics for best soft-margin SVM with LSI data.

	predicted label=0	predicted label=1
actual label=0	0.95	0.05
actual label=1	0.03	0.97

Table 11: Confusion matrix for best soft-margin SVM with NMF data.

	precision	recall	f1-score	support
Computer tech.	0.97	0.95	0.96	513
Recreational act.	0.95	0.97	0.96	433
avg / total	0.96	0.96	0.96	946

Table 12: Detailed statistics for best soft-margin SVM with NMF data.

Question (g) We report the ROC curves and confusion matrices in Figure 5.

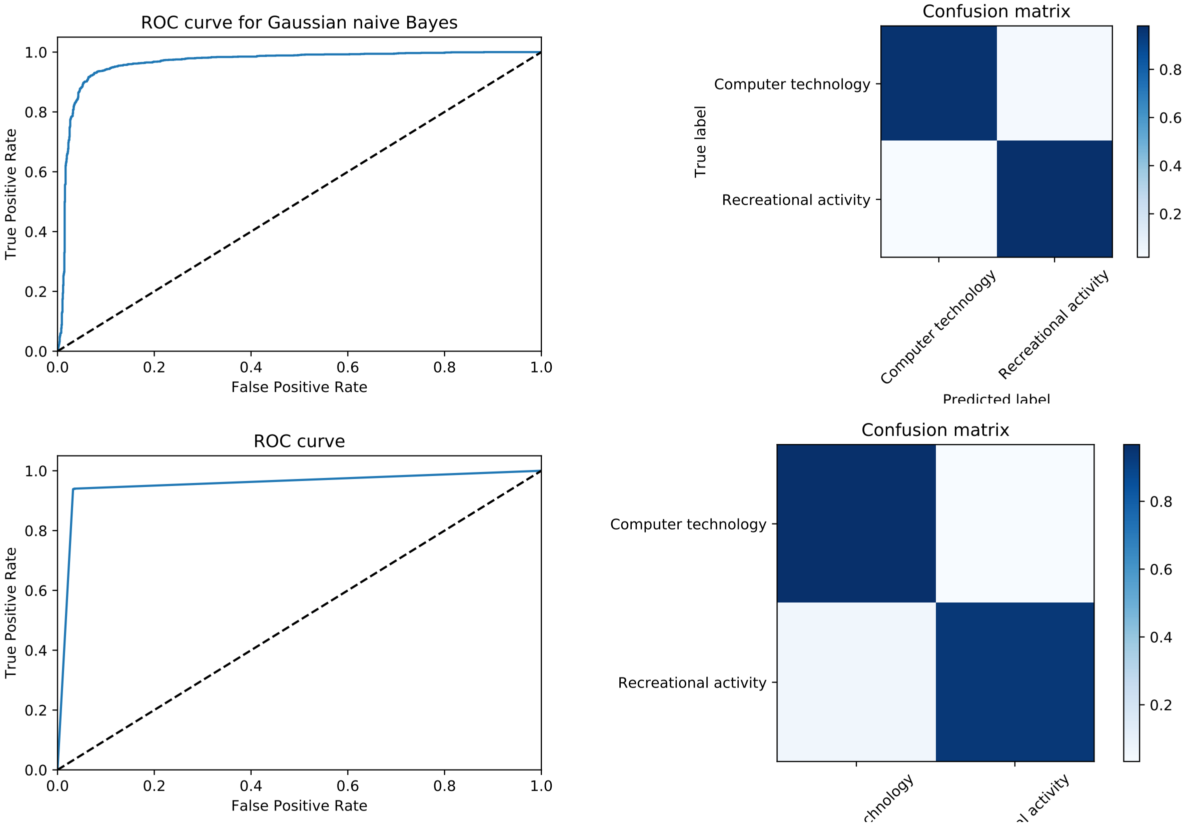

Figure 5: *Above.* The ROC curve and confusion matrix of the Gaussian naive Bayes classifier with LSI data. *Below.* The ROC curve and confusion matrix of the Gaussian naive Bayes classifier with NMF data.

The detailed statistics are shown in the following tables.

	predicted label=0	predicted label=1
actual label=0	0.87	0.13
actual label=1	0.05	0.95

Table 13: Confusion matrix for naive Bayes classifier with LSI data.

	precision	recall	f1-score	support
Computer tech.	0.95	0.87	0.91	1560
Recreational act.	0.88	0.95	0.92	1590
avg / total	0.92	0.91	0.91	3150

Table 14: Detailed statistics for naive Bayes classifier with LSI data.

	predicted label=0	predicted label=1
actual label=0	0.97	0.03
actual label=1	0.06	0.94

Table 15: Confusion matrix for naive Bayes classifier with NMF data.

	precision	recall	f1-score	support
Computer tech.	0.94	0.97	0.95	1560
Recreational act.	0.97	0.94	0.95	1590
avg / total	0.95	0.95	0.95	3150

Table 16: Detailed statistics for naive Bayes classifier with NMF data.

Question (h) We report the ROC curves and confusion matrices in Figure 6.

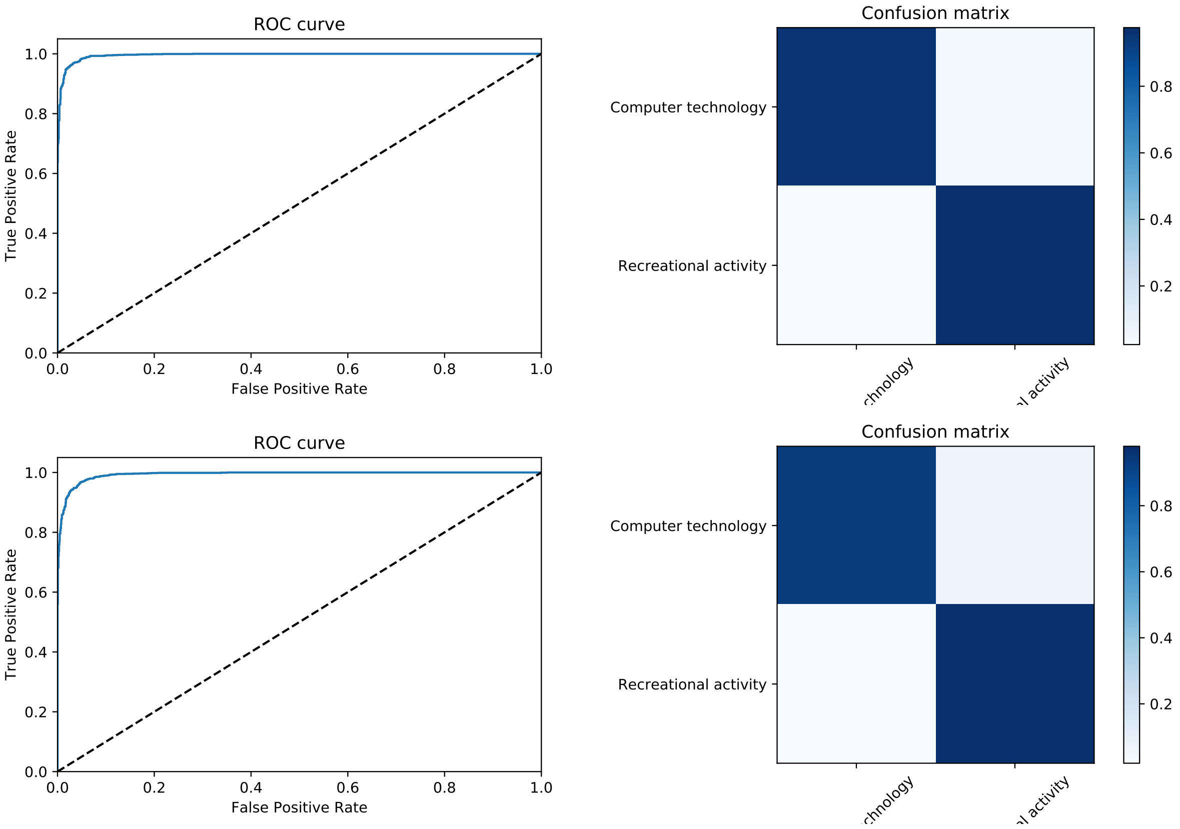

Figure 6: *Above.* The ROC curve and confusion matrix of the logistic regression classifier with LSI data. *Below.* The ROC curve and confusion matrix of the logistic regression classifier with NMF data.

The detailed statistics are shown in the following tables.

	predicted label=0	predicted label=1
actual label=0	0.93	0.07
actual label=1	0.02	0.98

Table 17: Confusion matrix for logistic regression classifier with LSI data.

	precision	recall	f1-score	support
Computer tech.	0.98	0.83	0.95	1560
Recreational act.	0.94	0.98	0.96	1590
avg / total	0.96	0.96	0.96	3150

Table 18: Detailed statistics for logistic regression classifier with LSI data.

	predicted label=0	predicted label=1
actual label=0	0.93	0.07
actual label=1	0.02	0.98

Table 19: Confusion matrix for logistic regression classifier with NMF data.

	precision	recall	f1-score	support
Computer tech.	0.98	0.93	0.95	1560
Recreational act.	0.94	0.98	0.96	1590
avg / total	0.96	0.96	0.96	3150

Table 20: Detailed statistics for logistic regression classifier with NMF data.

Question (i) The following table shows the accuracy results for different penalties as well as different data. The result shows that in general the larger penalty coefficient will generate a better performance, and overall ℓ_2 regularization works better than ℓ_1.
In general, ℓ_2 regualrization prefers small errors for all examples, while ℓ_1 regularization is in favor of sparse solutions and built-in feature selection.

Question (j) We first report the result for multiclass naive Bayes classifier in Figure 7.

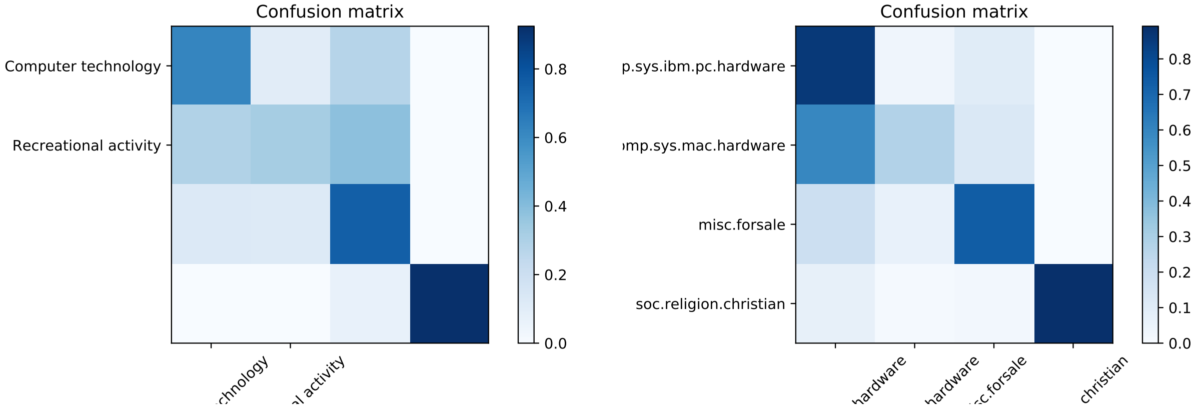

Figure 7: *Left.* The confusion matrix of the multiclass naive Bayes classifier with LSI data. *Right.* The confusion matrix of the multiclass naive Bayes classifier with NMF data.

The detailed statistics are shown in the following tables.

0.62	0.10	0.28	0.00
0.29	0.32	0.38	0.00
0.12	0.12	0.76	0.00
0.00	0.00	0.07	0.92

Table 21: Confusion matrix for multiclass naive Bayes classifier with LSI data.

	precision	recall	f1-score	support
comp.sys.ibm.pc.hardware	0.60	0.62	0.61	392
comp.sys.mac.hardware	0.59	0.32	0.42	385
misc.forsale	0.51	0.76	0.96	390
soc.rel.christian	1.00	0.92	0.96	398
avg / total	0.68	0.66	0.65	1565

Table 22: Detailed statistics for multiclass naive Bayes classifier with LSI data.

0.86	0.04	0.10	0.00
0.59	0.28	0.13	0.00
0.19	0.07	0.74	0.00
0.07	0.01	0.02	0.89

Table 23: Confusion matrix for multiclass naive Bayes classifier with NMF data.

	precision	recall	f1-score	support
comp.sys.ibm.pc.hardware	0.50	0.86	0.63	392
comp.sys.mac.hardware	0.70	0.28	0.39	385
misc.forsale	0.74	0.74	0.94	390
soc.rel.christian	1.00	0.89	0.94	398
avg / total	0.74	0.69	0.68	1565

Table 24: Detailed statistics for multiclass naive Bayes classifier with NMF data.

Figure 8 shows the confusion matrices for SVM (one-vs-one).
The detailed statistics are shown in the following tables.

0.84	0.10	0.06	0.00
0.12	0.83	0.05	0.00
0.05	0.05	0.89	0.00
0.01	0.01	0.01	0.98

Table 25: Confusion matrix for SVM (one-vs-one) with LSI data.

	precision	recall	f1-score	support
comp.sys.ibm.pc.hardware	0.82	0.84	0.83	392
comp.sys.mac.hardware	0.84	0.83	0.83	385
misc.forsale	0.88	0.89	0.89	390
soc.rel.christian	0.99	0.98	0.99	398
avg / total	0.89	0.88	0.88	1565

Table 26: Detailed statistics for SVM (one-vs-one) with LSI data.

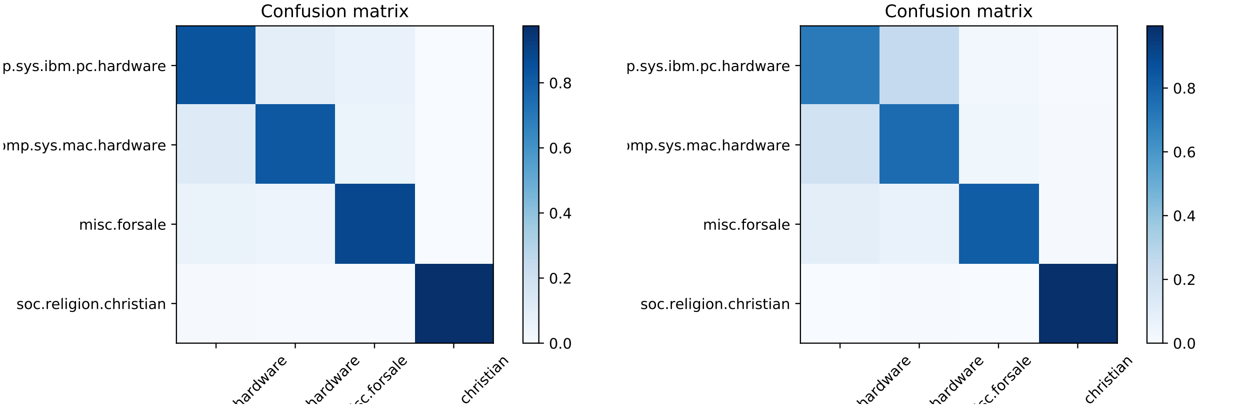

Figure 8: *Left.* The confusion matrix of SVM (one-vs-one) with LSI data. *Right.* The confusion matrix of SVM (one-vs-one) with NMF data.

0.71	0.19	0.09	0.01
0.15	0.77	0.06	0.01
0.05	0.04	0.90	0.01
0.00	0.01	0.00	0.99

Table 27: Confusion matrix for SVM (one-vs-one) with NMF data.

	precision	recall	f1-score	support
comp.sys.ibm.pc.hardware	0.78	0.71	0.74	392
comp.sys.mac.hardware	0.76	0.77	0.77	385
misc.forsale	0.85	0.90	0.87	390
soc.rel.christian	0.98	0.99	0.98	398
avg / total	0.84	0.84	0.84	1565

Table 28: Detailed statistics for SVM (one-vs-one) with NMF data.

Figure 9 shows the confusion matrices for SVM (one-vs-rest).
The detailed statistics are shown in the following tables.

0.82	0.11	0.06	0.01
0.08	0.85	0.06	0.00
0.05	0.05	0.89	0.01
0.01	0.01	0.01	0.98

Table 29: Confusion matrix for SVM (one-vs-rest) with LSI data.

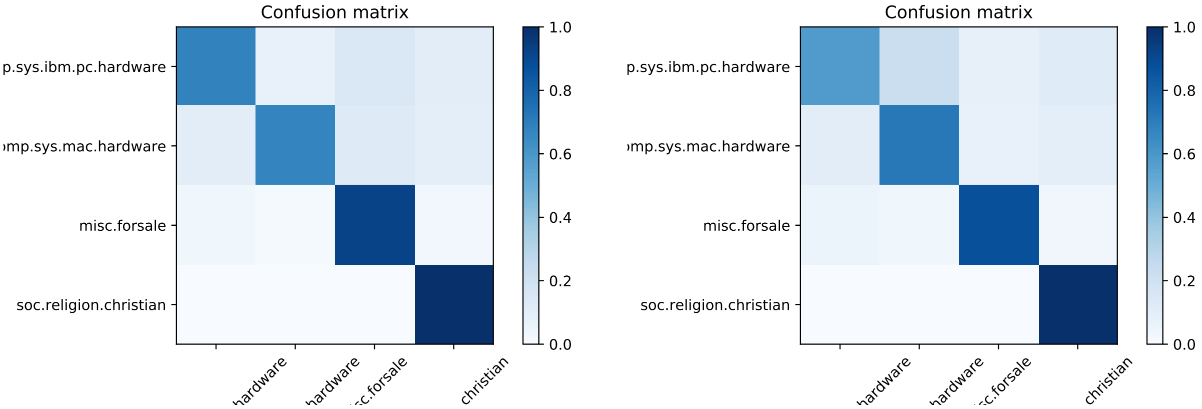

Figure 9: *Left.* The confusion matrix of SVM (one-vs-rest) with LSI data. *Right.* The confusion matrix of SVM (one-vs-rest) with NMF data.

	precision	recall	f1-score	support
comp.sys.ibm.pc.hardware	0.85	0.82	0.84	392
comp.sys.mac.hardware	0.84	0.85	0.85	385
misc.forsale	0.87	0.89	0.88	390
soc.rel.christian	0.98	0.98	0.98	398
avg / total	0.89	0.89	0.89	1565

Table 30: Detailed statistics for SVM (one-vs-rest) with LSI data.

0.52	0.29	0.12	0.06
0.04	0.83	0.07	0.05
0.03	0.05	0.90	0.02
0.00	0.01	0.00	0.99

Table 31: Confusion matrix for SVM (one-vs-rest) with NMF data.

	precision	recall	f1-score	support
comp.sys.ibm.pc.hardware	0.88	0.52	0.66	392
comp.sys.mac.hardware	0.70	0.83	0.76	385
misc.forsale	0.82	0.90	0.86	390
soc.rel.christian	0.88	0.99	0.94	398
avg / total	0.82	0.81	0.80	1565

Table 32: Detailed statistics for SVM (one-vs-rest) with NMF data.