

## Lab Sheet 7 – Text Classification

Kavishka Gamage – 17000475

Screenshot of the result of Classifier evaluation demo.py

```
Actual counts: [('spam', 10), ('ham', 10)]
Predicted counts: [('spam', 11), ('ham', 9)]
Predicted:
      spam ham
Actual: spam    5  5
       ham     6  4
Accuracy: 0.45
Manually computed accuracy: 0.45
Precision: 0.45
Manually computed precision: 0.45
Recall: 0.5
Manually computed recall: 0.5
F1 score: 0.48
Manually computed F1 score: 0.47

(base) G:\Github\NLP\UCSC>
```

Screenshot of the result of feature extraction demo.py

```
(base) G:\Github\NLP\UCSC>C:\Users\Kavishka\anaconda3\python.exe "g:/Github/NLP/UCSC/Labsheets/2_5_Text_Classification/feature_extraction_demo.py"
[[0 0 1 0 1 0 1 0 1]
 [1 1 1 0 2 0 2 0 0]
 [0 1 1 0 1 0 1 1 1]
 [0 0 1 1 0 1 0 0 0]]
[[0 0 1 0 0 0 1 0 0]]
['and', 'beautiful', 'blue', 'cheese', 'is', 'love', 'sky', 'so', 'the']
0 and beautiful blue cheese is love sky so the
0 0 0 1 0 1 0 1 0 1
1 1 1 1 0 2 0 2 0 0
2 0 1 1 0 1 0 1 1 1
3 0 0 1 1 0 1 0 0 0
0 and beautiful blue cheese is love sky so the
0 0 0 1 0 0 0 1 0 0
0 and beautiful blue cheese is love sky so the
0 0.00 0.00 0.40 0.00 0.40 0.00 0.40 0.00 0.60
1 0.44 0.35 0.23 0.00 0.56 0.00 0.56 0.00 0.00
2 0.00 0.43 0.29 0.00 0.35 0.00 0.35 0.55 0.43
3 0.00 0.00 0.35 0.66 0.00 0.66 0.00 0.00 0.00
0 and beautiful blue cheese is love sky so the
0 0.0 0.0 0.63 0.0 0.0 0.0 0.77 0.0 0.0
0 and beautiful blue cheese is love sky so the
0 0.0 0.0 1.0 0.0 1.0 0.0 1.0 0.0 1.0
1 1.0 1.0 1.0 0.0 2.0 0.0 2.0 0.0 0.0
2 0.0 1.0 1.0 0.0 1.0 0.0 1.0 1.0 1.0
```

Screenshot of the result of Classification.py

### Classes in news categories

```
['alt.atheism', 'comp.graphics', 'comp.os.ms-windows.misc', 'comp.sys.ibm.pc.hardware', 'comp.sys.mac.hardware', 'comp.windows.x', 'misc.forsale', 'rec.autos', 'rec.motorcycles', 'rec.sport.baseball', 'rec.sport.hockey', 'sci.crypt', 'sci.electronics', 'sci.med', 'sci.space', 'soc.religion.christian', 'talk.politics.guns', 'talk.politics.mideast', 'talk.politics.misc', 'talk.religion.misc']
```

### Result of SVM classifier and bow feature set

```
svm_bow =>  
  
Accuracy: 0.64  
Precision: 0.67  
Recall: 0.64  
F1 Score: 0.65
```

### Result of SVM classifier and TFIDF feature set

```
svm_tfidf =>  
  
Accuracy: 0.77  
Precision: 0.77  
Recall: 0.77  
F1 Score: 0.76
```

### Result of SVM classifier and avg word2vec feature set

```
svm_avgwv =>  
  
Accuracy: 0.56  
Precision: 0.57  
Recall: 0.56  
F1 Score: 0.55
```

### Result of SVM classifier and TFIDF weighted avg wordvector feature set

```
svm_tfidfwv =>  
  
Accuracy: 0.53  
Precision: 0.54  
Recall: 0.53  
F1 Score: 0.51
```

### Result of MNB classifier and bow feature set

```
mnb_bow =>  
  
Accuracy: 0.67  
Precision: 0.73  
Recall: 0.67  
F1 Score: 0.65
```

### Result of MNB classifier and TFIDF feature set

```
mnb_tfidf =>  
  
Accuracy: 0.72  
Precision: 0.78  
Recall: 0.72  
F1 Score: 0.7
```

## Result Interpretation

### Comparing SVM and MB performance with related to bow and TFIDF feature set :

Here for bow feature set MNB model has performed better than SVM classifier also there is a significant improvement in precision of MNB model.

As for TFIDF feature set, SVM model has performed better compared to all classifier and feature set combination.

Overall MNB model has given high precision in both bow and TFIDF feature set compare to SVM classifier.

### Comparing feature set bow, TFIDF, avg word vectors and TFIDF word vectors in SVM classifier

TFIDF feature set has given highest results compared to other features. It looks like for this case simple feature extraction methods like bow, TFIDF has given highest result compared to advanced technique like word2vec.

## SVM TFIDF model Error Analysis Interpretation

Misclassified instances for Actual Label: alt.atheism ,Predicted Label: soc.religion.christian

```
Actual label: alt.atheism
Predicted label: soc.religion.christian
Document:-
I would like a list of Bible contadictions from those of you who dispite being free from Christianity are well versed in the Bible.

Actual label: alt.atheism
Predicted label: soc.religion.christian
Document:-
They spent quite a bit of time on the wording of the Constitution. They picked words whose meanings implied the intent. We have already looked in the dictionary to define the word. Isn't this sufficient? But we were discussing it in relation to the death penalty. And, the Constitution need not define each of the words within. Anyone who doesn't know what cruel is can look in the dictionary (and we did).

Actual label: alt.atheism
Predicted label: soc.religion.christian
Document:-
Our Lord and Savior David Keresch has risen! He has been seen alive! Spread the word! -----

Actual label: alt.atheism
Predicted label: soc.religion.christian
Document:-
"This is your god" (from John Carpenter's "They Live," natch)
```

Model identify document as Christian when there words like god, bible, death present. Having high weight for these word might be the case for misclassified instances.

Misclassified instances for Actual Label: talk.politics.misc Predicted Label: talk.politics.guns

```
Actual Label: talk.politics.misc
Predicted Label: talk.politics.guns
Document:-
After the initial gun battle was over, they had 50 days to come out peacefully. They had their high priced lawyer, and judging by the posts here they had some public support. Can any one come up with a rational explanation why the didn't come out (even after they negotiated coming out after the radio sermon) that doesn't include the Davidians wanting to commit suicide/murder/general mayhem?

Actual Label: talk.politics.misc
Predicted label: talk.politics.guns
Document:-
Yesterday, the FBI was saying that at least three of the bodies had gunshot wounds, indicating that they were shot trying to escape the fire. Today's paper quotes the medical examiner as saying that there is no evidence of gunshot wounds in any of the recovered bodies. At the beginning of this siege, it was reported that while Keresch had a class III (machine gun) license, today's paper quotes the government as saying, no, they didn't have a license. Today's paper reports that a number of the bodies were found with shoulder weapons next to them, as if they had been using them while dying -- which doesn't sound like the sort of action I would expect from a suicide. Our government lies, as it tries to cover over its incompetence and negligence. Why should I believe the FBI's claims about anything else, when we can see that they are LYING? This system of government is beyond reform.

Actual Label: talk.politics.misc
Predicted label: talk.politics.guns
Document:-
If you look through this newsgroup, you should be able to find Clinton's proposed "Wiretapping" Initiative f
or our computer networks and telephone systems. This 'Initiative' has been up before Congress for at least t
he past 6 months, in the guise of the "FBI Wiretapping" bill. I strongly urge you to begin considering your future. I
strongly urge you to get your application for a passport in the mail soon.

Actual Label: talk.politics.misc
Predicted Label: talk.politics.guns
Document:-
Well, for one thing most, if not all the Davidians (depending on whether they could show they acted in self-defense and there were no illegal weapons), could have gone on with their life as they were living it. No one was forcing them to give up their religion or even their legal weapons. The Davidians had survived a change in leadership before so even if Keresch himself would have been convicted and sent to jail, they still could have carried on. I don't think the Davidians were insane, but I don't see a reason for mass suicide (if the fire was intentional set by some of the Davidians.) We also don't know that, if the fire was intentionally set from inside, was it a generally know plan or was this something only a inner circle knew about, or was it something two or three felt they had to do with or without Keresch's knowledge/blessing, etc.? I don't know much about Masada. Were some people throwing others over? Did mothers jump over with their babies in their arms?
```

Model tend to identify document as guns when 'gun',' jail' word present. This might be the case for misclassified instances.