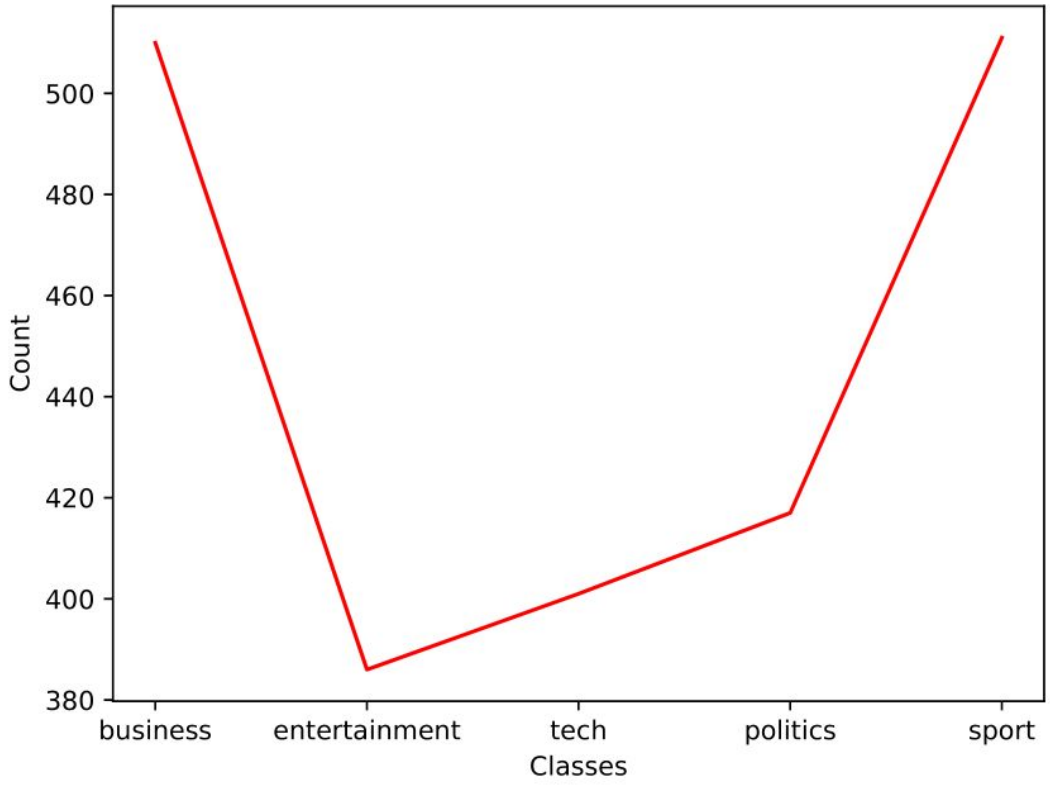
MP1 results & observations

Robert Michaud 40058095 (is me)



Distribution of BBC dataset

Effect on data

Number of word-tokens per class

- business: 131695
- entertainment: 96299
- tech: 153673
- politics: 133317
- sport: 15719

Number and % of words w/ a frequency of zero per class

- business: 18732 (14.22%)
- entertainment: 19104 (19.84%)
- tech: 19108 (12.43%)
- politics: 19726 (14.80%)
- sport: 18432 (11.73%)

Metric for BBC dataset

Although all metrics yield similar results, I believe that '**accuracy of model**' is more important. This is because we're interested in the general classification of documents in the correct class.

Because of that, there is no emphasis on any given class,

> i.e. classifying 'business' documents correctly every time is no more important than any other given class.

MultinomialNB Performance

Default, try 1 & 2

Accuracy of model: 98.65%

Macro-average F1 of model: 98.58%

Weighted-average F1 of model: 98.65%

Smoothing = 0.0001

Accuracy of model: 97.98%

Macro-average F1 of model: 97.93%

Weighted-average F1 of model: 97.98%

Smoothing = 0.9

Accuracy of model: 98.65%

Macro-average F1 of model: 98.58%

Weighted-average F1 of model: 98.65%

Log probabilities for “Zombie”

Default, try 1 & 2

business: -11.9898798814

entertainment: -10.13237457848

tech: -12.11775496115

politics: -11.99989682463

sport: -11.03816922661

Smoothing = 0.0001

business: -20.99860663392

entertainment: -10.08892440396

tech: -21.1529417642

politics: -21.01084747010

sport: -11.2720320811

Smoothing = 0.9

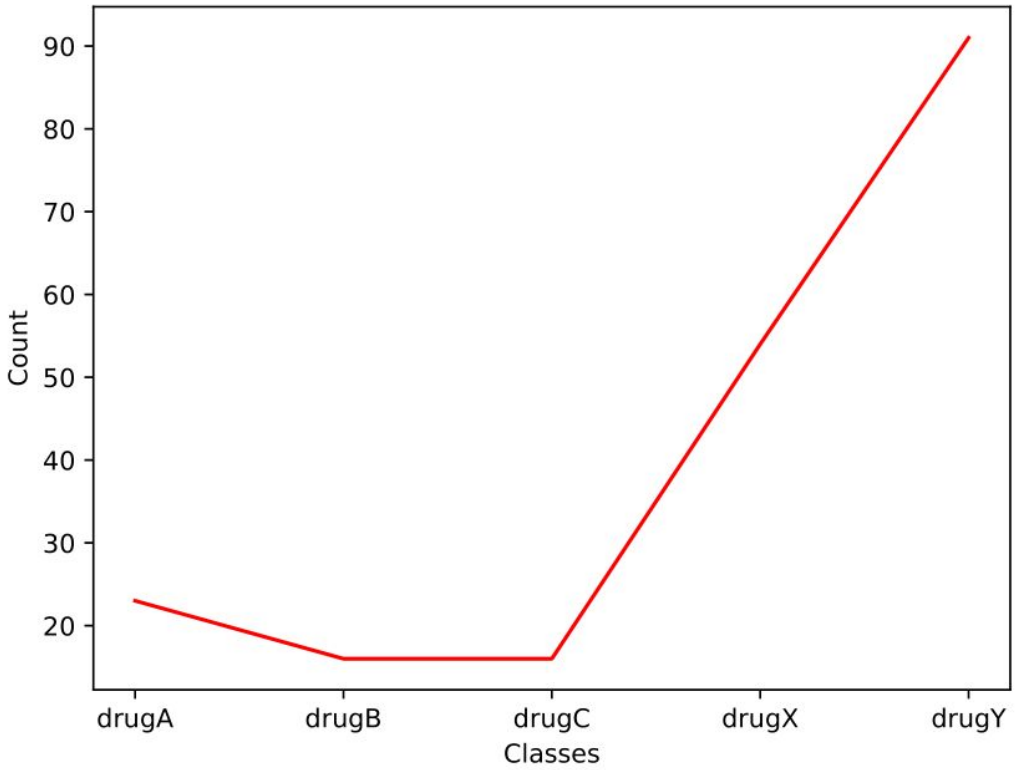
business: -12.07681085532

entertainment: -10.12889710596

tech: -12.20691617855

politics: -12.087013170571652

sport: -11.0561792221



Distribution of drug dataset

Observations

GaussianNB

```
[[15  0  0  0  0]
 [ 0 13  0  0  0]
 [ 0  0 13  1  0]
 [ 0  0  0 36  5]
 [ 4  6  5  2 50]]
```

Accuracy of model: 84.67%

Macro-average F1 of model: 84.54%

Weighted-average F1 of model: 84.66%

	precision	recall	f1-score	support
drugA	0.789	1.000	0.882	15
drugB	0.684	1.000	0.813	13
drugC	0.722	0.929	0.813	14
drugX	0.923	0.878	0.900	41
drugY	0.909	0.746	0.820	67
accuracy			0.847	150
macro avg	0.806	0.911	0.845	150
weighted avg	0.864	0.847	0.847	150

Observations

Base-DT

```
[[19  0  0  0  0]
 [ 0 10  0  0  0]
 [ 0  0  6  7  0]
 [ 0  0  1 35  0]
 [ 0  0  1  1 70]]
```

Accuracy of model: 93.33%

Macro-average F1 of model: 88.87%

Weighted-average F1 of model: 92.88

	precision	recall	f1-score	support
drugA	1.000	1.000	1.000	19
drugB	1.000	1.000	1.000	10
drugC	0.750	0.462	0.571	13
drugX	0.814	0.972	0.886	36
drugY	1.000	0.972	0.986	72
accuracy			0.933	150
macro avg	0.913	0.881	0.889	150
weighted avg	0.934	0.933	0.929	150

Observations

Top-DT

{'criterion': 'gini', 'max_depth': 8, 'min_samples_split': 4}

```
[[19  0  0  0  0]
 [ 0 10  0  0  0]
 [ 0  0  4  9  0]
 [ 0  0  4 32  0]
 [ 0  0  1  1 70]]
```

Accuracy of model: 90.00%

Macro-average F1 of model: 83.40%

Weighted-average F1 of model: 89.50%

	precision	recall	f1-score	support
drugA	1.000	1.000	1.000	19
drugB	1.000	1.000	1.000	10
drugC	0.444	0.308	0.364	13
drugX	0.762	0.889	0.821	36
drugY	1.000	0.972	0.986	72
accuracy			0.900	150
macro avg	0.841	0.834	0.834	150
weighted avg	0.895	0.900	0.895	150

Observations

PER

```
[[ 0  0  5  8  6]
 [ 0  0  0 10  0]
 [ 0  0  1 10  2]
 [ 0  0  2 22 12]
 [ 0  0  0 12 60]]
```

Accuracy of model: 55.33%

Macro-average F1 of model: 26.67%

Weighted-average F1 of model: 49.50%

	precision	recall	f1-score	support
drugA	0.000	0.000	0.000	19
drugB	0.000	0.000	0.000	10
drugC	0.125	0.077	0.095	13
drugX	0.355	0.611	0.449	36
drugY	0.750	0.833	0.789	72
accuracy			0.553	150
macro avg	0.246	0.304	0.267	150
weighted avg	0.456	0.553	0.495	150

Observations

Base-MLP

```
[[ 0  0  0 10  9]
 [ 0  0  0 10  0]
 [ 0  0  0 10  3]
 [ 0  0  0 22 14]
 [ 0  0  0 14 58]]
```

Accuracy of model: 53.33%

Macro-average F1 of model: 23.50%

Weighted-average F1 of model: 46.05%

	precision	recall	f1-score	support
drugA	0.000	0.000	0.000	19
drugB	0.000	0.000	0.000	10
drugC	0.000	0.000	0.000	13
drugX	0.333	0.611	0.431	36
drugY	0.690	0.806	0.744	72
accuracy			0.533	150
macro avg	0.205	0.283	0.235	150
weighted avg	0.411	0.533	0.460	150

Observations

Top-MLP

{'activation': 'tanh', 'hidden_layer_sizes': (30, 50), 'solver': 'adam'}

```
[[13  4  0  2  0]
 [ 0  9  0  1  0]
 [ 0  0  3 10  0]
 [ 1  0  5 28  2]
 [ 0  0  0  7 65]]
```

Accuracy of model: 78.67%

Macro-average F1 of model: 69.16%

Weighted-average F1 of model: 78.57%

	precision	recall	f1-score	support
drugA	0.929	0.684	0.788	19
drugB	0.692	0.900	0.783	10
drugC	0.375	0.231	0.286	13
drugX	0.583	0.778	0.667	36
drugY	0.970	0.903	0.935	72
accuracy			0.787	150
macro avg	0.710	0.699	0.692	150
weighted avg	0.802	0.787	0.786	150