CMPUT 497: Assignment 5 – Report

Shouyang Zhou
University of Alberta
Edmonton, Alberta, Canada
shouyang@ualberta.ca

1 Classify Text Type

Exploratory (Training) Data Analysis

Given that a Naïve Bayes classifier demonstrates significant bias, I first inspect the qualities of the training dataset since these characteristics will factor into my model tuning and cross validation concerns. I do this using Excel, constructing a pivot table of some basic averages and counts from the training dataset.

	% of Labels	Avg. Len	Sum of Len.
Business	22.89	1948	759,844
Entertain.	17.02	1905	552,493
Politics	18.78	2689	860,778
Sport	22.48	1898	726,965
Tech	18.84	2986	958,672

* Len being character length of the text samples.

By label, the classes exhibit a moderate to minor class imbalance. Some classes such as politics and tech demonstrate longer articles which may concentrate their tokens towards certain subjects and themes more than other classes.

Overall these metrics show that there is no dominating class within the dataset. There may be some concern on the amount of data contrasting entertainment and tech, but in general, these problems are not significant.

Feature Representation

I decided to use a binary variant of naïve Bayes focusing on unigrams. That is, each text is represented by a binary vector mapping the existence of terms in the text to a true/false or 1/0 value. The universe of terms is generated from a subset of the training text. This decision was based on the need to generate a compact and generalizable set of features. This representation is small enough such that bulk training and prediction must be computationally efficient while maintaining enough features to adequately classify texts from each class.

I believe the use of unigrams represents the best trade-off of data representativeness for compactness. The use of higher n-gram models compilates many aspects such as model generalization, the feature space size, and the effects of smoothing. The training dataset represents ~ 1700 small articles, alone this wouldn't provide enough n-grams to achieve decent coverage of the relevant n-grams possible in each class. I suspect that this would lead to an overfitting of the data given its sparseness, and the introduction of more noise since the weight of the smoothing is more important.

Furthermore, I believe that intuitively the domain specific terms are likely to be more prevalent and relevant than that of domain specific phrasing as represented in higher n-grams. It is more likely that domain specific phrasing is a product of domain specific terms hence unigrams should be enough.

Lastly, I use a binary encoding of terms to further reduce noise. Intuitively using the observed term to class counts introduces noise from the author's preference for certain terms, and the length of at which they talk about a given subject. In generalizing the core relation of an observed term to a label-class, I believe it is more important that a term is more often used across texts from that label. This is especially confounded by the small amount of texts per label-class in the training da-

taset and the high bias of the naïve Bayes classifier.

Feature Selection

Under this feature representation, there are still ~30,000 candidate unigram features to select from using the entire training set. Under my preprocessing scheme below, this is reduced to ~23,000 via generalizing numbers and lemmatizing tokens. It would be quite computationally costly to generate and test via this feature set.

Examining the frequency distribution of terms, they appear to fall under an exponential distribution. I decide the prune the universe of features using via skipping the top n expressed terms, effectively treating this as a stop list, and obtaining maximally k features after skipping the top n terms. Intuitively, this should approximately capture the most common generalizable terms used between the various classes.

I constructed my model during reading week and before I saw the feature selection methods in the information retrieval text. I had achieved a satisfactory level of performance below in the parameter tuning section, so I did not implement mutual information index as a feature selection mechanism. I suspect it may have proven useful in further honing my universe of terms.

Parameter Tuning

I tune my design decisions in the cross-validation procedure. I experimented with the use of preprocessing, i.e. lemmatizing and abstracting numbers, and my stop list and maximal feature length. My stopping criterion was to achieve higher than 90% average accuracy in the cross-validation phase. After tuning, this represents skipping the top 100 most common terms, and using the 500 terms that follow.

Smoothing

My model uses “add-half” smoothing term on both the label and per feature-label estimates which provides milder smoothing than Laplace smoothing. This is slightly different than the smoothing described in the text, which adds a numerator term and a denominator term to probability calculation, instead I add a constant amount to the calculated probabilities.

Initially this was by mistake, but I decided to keep the smoothing as is after comparing the results from the cross validation. This alternative smoothing method increased mean accuracy from 85% to 94%. Intuitively, regular smoothing would smooth the probability distributions of the model, whereas my method simply adds a small constant to avoid zero probabilities. My goal was avoiding zero probabilities and to include a milder smoothing to that of Laplace hence I am comfortable with keeping this unintentional change especially given the higher training accuracy. This would lean the model towards more bias in the bias-variance tradeoff in contrast to the Laplace smoothing. From the training accuracy results above, and my experiences with smoothing in the previous experiments, I suspect that add-n smoothing in general adds too much variance to these models. Technically, I doubt that the resultant values are strictly probabilities anymore although for pragmatic concerns this is not an issue like the assumptions the naïve Bayes method on the structure of text.

Preprocessing

While the data for the assignment has a degree of prepressing, I add some additional measures to compress my feature space.

I use the Treebank word and Punkt sentence tokenizer in NLTK and replace numbers in the dataset with a generic token. The treebank and Punkt tokenizers have been recommended by the authors of NLTK, they represent general tokenizers suitable for English. The Treebank tokenizer may be especially suitable given that it was devised from another news-source / journalism data source like the BBC.

I replace numbers to generalize numeric features in the dataset. I expect that denominations / units may be a good feature. For example, in sport scoring measures such as “3-1” are common and in contrast, in business denominations in percent and millions may be more common.

Lastly, I use a rough pass lemmatizer to reduce noise from the phrasing of the text. Consider for example that an entertainment article may contain the words “star”, “stars”, “starring”, ‘starred”, the

root word "star" is the relevant feature not the actual word form. The Wordnet lemmatizer requires the pos tag of the word to function. I found it computationally inefficient to generate the POS tag string per text and as such, I assume that each token is a verb in the lemmatization step. In general, this should leave many non-verb tokens unaffected. While it is likely that this introduces some noise, the drastic reduction in terms (from 30,000 to 23,000) suggests that this will likely improve model performance.

Test Results

I conducted my analysis of the model-estimates on samples from "testBBC.csv" via Excel from the generated output file. My model misclassifies 39 samples from the test dataset representing an accuracy of 94%. See next for replications of Figure 4.5 and 4.6 from the chapter. I recommend viewing the Excel sheet tab titled "Confusion Matrix" for the calculations. The report formatting here will make the tables taken from that file rather small. Because the model accuracy was fairly, high the macro averaged and micro averaged precision and recall measures where very similar. Please see the "Total" confusion matrix for the micro averaged precision and recall metrics.

I will report upon these figures in the next section, as part of the error analysis.

	Actual					
Estimate	**BUS**	**ENT**	**POL**	**SPO**	**TECH**	**Pre.**
business	147	5	3	2	3	0.92
entertainment		105	1		2	0.97
politics	8	5	116	1		0.89
sport			1	160	1	0.99
tech	2	2	3		101	0.94
Re.	0.94	0.90	0.94	0.98	0.94	
F1	0.93	0.93	0.91	0.98	0.94	

Business	**Actual**	**Not-Actual**
Estimate	147	13
Not-Estimate	10	498
Recall / Precision	0.94	0.92
Entertainment	**Actual**	**Not-Actual**
Estimate	105	3
Not-Estimate	12	548
Recall / Precision	0.90	0.97
Politics	**Actual**	**Not-Actual**
Estimate	116	14
Not-Estimate	8	530
Recall / Precision	0.94	0.89
Sport	**Actual**	**Not-Actual**
Estimate	160	2
Not-Estimate	3	503
Recall / Precision	0.98	0.99
Tech	**Actual**	**Not-Actual**
Estimate	101	7
Not-Estimate	6	554
Recall / Precision	0.94	0.94
Total	**Actual**	**Not-Actual**
Estimate	629	39
Not-Estimate	39	2633
Recall / Precision	0.94	0.94
Macro Average		
Recall	0.94	
Precision	0.94	

2 Error Analysis

I will report on the combined set of errors from the test and eval datasets. First, I will hand label the results from the eval dataset and report a confusion matrix for that set. This will be in the form of another Excel based on the output of "evalBBC.csv".

Manual Labelling (Eval – Dataset)

Via Excel, I will hide the estimator's labeling and skim through the text samples until I form my own labelling assignment per sample. I will report the results in a confusion matrix. Because, I may interpret the articles contents subjectively, I will also report an adjusted accuracy figure after re-

viewing the estimator's labelling and reassessing my labelling.

Eval Dataset Results

	Actual					
Estimate	BUS	ENT	POL	SPO	TECH	Pre.
business	14	1	3	1	2	0.67
entertainment		9				1.00
politics	3	3	15			0.71
sport	1			9		0.90
tech	1	1	3		5	0.50
Re.	0.74	0.64	0.71	0.90	0.71	
F1	0.70	0.78	0.71	0.90	0.59	

The raw accuracy was 74%, after reading the estimator assigned labelling, my adjusted accuracy was 90%.

Overall Trends

In both datasets, errors in differing labels had different affinities for their true label. The business articles appear to have affinities for the politics and technology, and all other labels have an affinity to be mislabeled as business articles. Entertainment articles show some affinity to business and politics, non-entertainment articles are seldom mislabeled as entertainment. Political articles are at times mislabeled to business or technology, although it is more common for business or entertainment articles to be mislabeled into entertainment ones. Sports articles are generally very well isolated, there is a very minor cross labelling of business and sport. Technology articles share some affinity to business articles, generally business, entertainment, and political articles are candidates to be mislabeled as technology articles.

By precision, both entertainment and sport standout in both datasets. Politics is a consistently lower performing label in both datasets.

By recall, sport is a well performing label while entertainment is the lowest in both datasets.

Rationalizing Errors

From my labelling of the evaluation dataset, I find that the general trends of error denote some natural relation between the various subject domain. This is then confounded by the binary feature encoding and the high bias of the naïve Bayes classifier. I'll will present some unit examples to demonstrate this.

From my hand labelling, I found several article introductions that can be more ambiguous for a reader than the classifier. Consider the snippets (My :Label / Estimator Label):

- *(1) jackson film absolute disaster a pr expert has told the michael jackson child abuse trial that the tv documentary at the heart of the case was an absolute disaster . (entertainment / politics)*
- *(2) uk firms embracing e-commerce uk firms are embracing internet trading opportunities as never before e-commerce minister mike o brien says. a government-commissioned study ranked the uk third in its world index of use of information and communication technology (ict). the report suggests 69% of uk firms are now using broadband and that 30% of micro businesses are trading online. (business / tech)*

Similarly, I found the same in the test dataset (Actual Label / Estimator Label):

- *(3) uk s national gallery in the pink the national gallery home to some of the uk s greatest artworks has seen a big jump in visitor numbers. five million visitors made the london gallery - which houses treasures like raphael s madonna of the pinks - the uk s most visited museum in 2004. it recorded a 13.8% rise in numbers and was the country s second most visited tourist attraction behind blackpool pleasure beach.. (entertainment / business)*
- *(4) campbell to be lions consultant former government communications chief alastair campbell will act as a media consultant to sir clive woodward s 2005 lions on their tour to new zealand. campbell who left downing street earlier this year will advise on media strategy before and during the tour. (sport / politics)*

In all these snippets one theme is the use of cross domain terms. One can find likely inclusions in the universe of features my feature selection scheme. For example, "film" contrasted to "trial", "internet" versus "trading", "gallery" and "million" or "lions" to the two term phrase "dowing street". Given a binary feature encoding, I suspect that small differences in the details reported, and the class probabilities generated these differences.

One exemplary article starts and ends with the following:

- *(article head) us blogger fired by her airline a us airline attendant suspended over inappropriate images on her blog - web diary - says she has been fired. ellen simonetti known as queen of the sky wrote an anonymous semi-fictional account of her life in the sky. she was suspended by delta in september. in a statement she said she was initiating legal action against the airline for wrongful termination . a delta spokesperson confirmed on*

wednesday that ms simonetti was no longer an employee.

- *(article tail) delta has been hit recently by pressures of rising fuel costs and fierce competition. it has said it needs to cut between 6 000 and 7 000 jobs and reduce costs by $5bn (Â£2.7bn) a year. analysts had warned recently that the airline might have to seek chapter 11 bankruptcy prevention. last week it struck a $1bn cost-cutting deal with its pilots which could save it from bankruptcy. the deal would see pilots accept a 32% pay cut in return for the right to buy 30 million delta shares unions said. and on monday it negotiated a deal to defer about $135m in debt which was due next year until 2007. the airline also said it had agreed the terms of a $600m loan from american express.*

This article is a technology article but mislabeled as a business article. From the head of the article, the reporter wishes to address the social issues related to blogging, but they also include this concise briefing on the business aspect of Delta Airlines in the article. Given the binary feature encoding, the use of business keywords such as "cost", "competition", and "shares" likely overshadowed the actual meaning conveyed by the article. This was a good business summary that concisely addressed key events and metrics of Delta airlines that in the eyes of the estimator, dominates the actual discussion! Compare and contrast the possible tech and business key terms between the two sections!

In general though, I suspect that if two labels were allowed to be assigned to each article then the accuracy of this task would be near 100%. More often, the articles share some element of one label and another because the subject matter touches upon both. The nature of misclassifications is dependent upon the nature of the training data, and simply the amount of content the author devotes the differing aspects of the subject.