The Hard Truth about Soft News

I. Abstract

All news stories can be divided into two categories: hard news and soft news. Hard news usually refers to up-to-the minute news and events that are reported immediately and are of an urgent nature. It mostly covers topics in politics, war, economics and international affairs. On the other hand, soft news consists of news content that is background information on someone's life, arts and entertainment, human interest stories or articles giving advice to people.

Increasingly, the lines between hard news and soft news are beginning to blur and creating a new category that is not news. Most of the newspapers today are more concerned with keeping up with the Kardashians and the new royal baby than they are with Ebola or the war in the middle-east.

It is important to be able to distinguish between hard news, soft news and not news as not only does it help one understand what kind of news is being covered in the media but also helps one know what kind of articles they want to be influenced by.

This paper uses 100,000 articles from the New York Times spanning the last three decades and attempts to build a model that effectively classifies hard news from soft news. The model goes one step further and provides what proportion of the New York Times articles are soft news and what proportion are hard news. This model can also be extended to other newspapers and media sources in order to understand what kind of topics are covered by them as part of 'news'.

A further application of this paper can also be to compare the difference in the kind of news by ownership patterns of the papers as this may give insight into why certain media sources choose to cover certain topics. At the same time, it gives consumers the information they need on the kind of news covered by their daily newspapers.

II. Methodology

The original dataset includes 100,000 New York Times articles with information including headlines, lead paragraph, body, date, categories, section, etc. In order to process the data on the machine used in this paper - a sample of 1000 articles was selected from the original dataset. The sampling process used was stripe sampling which took 500 random hard news articles and 500 random soft news articles.

Post the sampling, the data was pre-processed in order to clear the text to the bare minimum. This included removing stop words, special characters, number and capitalizations. This was done by using the script provided by Dr. Gaurav Sood.[1]

After the text was cleaned, bigrams and trigrams were generated from the text. Bigrams and trigrams are a contiguous sequences of two and three words respectively generated from the given text. These bigrams and trigrams were also sorted on frequency in order to get an idea of the most commonly used words by the New York Times in their articles. Some of these top bigrams and trigrams can be found in Appendix A.

For the purpose of this study, we would like to see different usage of words between soft news and hard news by calculating the chi square. [2]In the simplest terms, the chi square trimming provides a way of seeing if some texts occur together more than they might by chance. The chi square formula used can be stated as the following:

$$\chi^2 = \frac{(f_s * f_{\bar{h}} - f_h * f_{\bar{s}})^2}{(f_s + f_h) * (f_s + f_{\bar{s}}) * (f_h + f_{\bar{h}}) * (f_{\bar{s}} + f_{\bar{h}})}$$

In the formula, f_s and f_h represent the frequecy of the n-gram in soft news articles and hard news articles while $f_{\bar{s}}$ and $f_{\bar{h}}$ are the frequency of the n-grams not apprearing in soft news or hard news. Using this metric, the Chi square for each bigram and trigram is calculated. To include the bigrams and trigrams that explains the most difference

[1] https://github.com/soodoku/Text-as-Data

[2] http://faculty.chicagobooth.edu/matthew.gentzkow/research/biasmeas.pdf,

between soft news and hard news, only the top 500 bigrams and trigrams were included in the model. Each bigram and trigram is constructed as a binary variable and checked whether they have appeared in the body or lead paragraph of the articles.

In order to have sense of the bigrams and trigrams, here are some of the ones with the highest chi-squares.

Bigrams			
Human right	Official said	White house	Prime minister
Major league	General manager	Home run	Police said
Police office	Al Qaeda	Suprem Court	Bush administration

Trigrams			
Receiv master degre	Major leage baseball	Nation basketbal associ	Official said today
Low blood pressur	Nation hockey leagu	John F Kennedy	Junior high school
World War II	Dr Martin Luther	food drug administr	secur exchang commiss

In order to make the model more accurate in its prediction, certain bigrams and trigrams that occur frequently but are not of much value were ignored. An example of such bigrams and trigrams would be: 'also said', 'two year', 'mr mr' and 'new york time', 'new york ny','said nt want'. For a full list of all bigrams and trigrams ignored refer to Appendix B.

III. The Support Vector Machine Classifier

The Support Vector Machine Classifier (SVM) is one of the most popular and efficient classification and regression methods currently available in the programming sphere. The SVM classifier applies a linear method to the data but in a high dimensional feature space non-linearly related to the input space.[3] While some experiments were conducted

[3] http://www.jstatsoft.org/v15/i09/paper

with ridge regression and the logit model, the SVM classifier was ideal for the study as the dependent variable is a binary outcome and the goal of the paper is to generate a classifier that generalizes well, i.e. provides the same level of accuracy with data that is not yet experimented with.

Fitting the SVM Classifier:

In order to fit the classifier – the bare minimum data set was adopted - 500 observations of the sample were selected as the training data set. The SVM was initially run with a cost of 10 and a linear kernel. Once the model was fitted, cross validation was adopted and it was tuned to find the best cost function which –for the purpose of this study – was 0.1.

While the SVM classifier usually works in a linear fashion, quite often the distinction in the data may be non-linear. In order to cover that base, we fit an SVM classifier with a radial kernel and a cost of 10 initially. Once again, cross validation was adopted and the model was tuned to find the best cost and gamma which - for the purpose of this study - were 1 and 0.5 respectively.

IV. Results

The results of both the SVM classifiers that were fitted are expressed in the following confusion matrices. [4] The true result means that the news article is classified as soft news article while false result refers to hard news.

1. **Linear Kernel, best cost = 0.1**

		Observed	
		True	**False**
Predicted	**True**	**144**	**49**
	False	**95**	**212**

Accuracy: 0.712 or 71.2%

[4] Refer to Appendix C for formulae used to calculate various metrics of the confusion matrix

This means that of all the observations, this model has classified 71.2% of the observations correctly.

Error Rate: 0.288 or 28.8%

This means that of all the observations, this model has misclassified 28.8% of all the observations.

Sensitivity or True Positive Ratio: 0.603 or 60.3%

This means that 60.3% of the total of actual soft news was correctly recognized as soft news.

Specificity or False Positive ratio: 0.3402 or 34.02%

This means that 34.02% of the false observations were recognized as true or misclassified.

Bayes Error Rate: 0.4716 or 47.16%

This means that the lowest possible error rate for this classifier is 47.16% or that 47.16% of the observations would be misclassified at the very least.

2. **Radial kernel, best cost 1 and gamma 0.5**

	Observed		
		True	**False**
Predicted	**True**	**188**	**170**
	False	**51**	**91**

Accuracy: 0.558 or 55.8%

This means that of all the observations, this model has classified 55.8% of the observations correctly.

Error Rate: 0.442 or 44.2%

This means that of all the observations, this model has misclassified 44.2% of all the observations.

Sensitivity or True Positive Ratio: 0.787 or 78.7%

This means that 78.7% of the total of actual soft news was correctly recognized as soft news.

Specificity or False Positive ratio: 0.412 or 41.2%

This means that 41.2% of the false observations were recognized as true or misclassified.

Bayes error rate: 0.5995 or 59.95%

This means that the lowest possible error rate for this classifier is 59.95% or that 59.95% of the observations would be misclassified at the very least.

V. Inference

From the above results, the models with linearl kernel and with radial kernel appear to have different strength in predicting different news. We can see that the model with a linear kernel is better at predicting hard news than the one with radial kernel. But the radial kernel model turns out to have better results in predicting soft news than the linear kernel one. Thus the choice of model may depend on what kind of news we are trying to classify. Yet in general, it may be estimated that overall the model with the linear kernel and 0.1 as the best cost gives the most ideal result with an accuracy of over 70%.

Also, there are a few ways in which the accuracy of prediction can be improved:

i. In this project, we are trying to balance between the prediction of both kinds of news articles. For the prediction of soft news to be improved, we could have a larger number of soft news articles than hard news articles in the sample set instead of having equal number of both kinds of articles.
ii. Since some bigrams and trigrams that do not make sense such as “last year”, “said would” and “one time” are ignored, there could be potential problems associated with the choice of bigrams and trigrams.
iii. As only 1000 articles were selected as the sample, a larger data set of 100,000 articles will be helpful with getting more accurate prediction.

VI. Appendix A

young woman	young man name
young men	york time west
young man	york new York
young children	york law firm
york univers	york citi mr
york ny	yesterday new york
york new	year old said
yield serv	year old graduat
yesterday afternoon	year new york
year people	year first time
year last	year ago want
year compani	year ago said
would receiv	would work hard
would go	would like see
would give	world war ii
would cost	world trade organ
would continu	world trade center
world war	work new york
world seri	wing metropolitan museum
world said	west street new
world championship	west new york
world champion	visit new york
white hous	ve got lot
west street	vanilla ice cream
west coast	upstat new york
week last	upper west side

VII. Appendix B

also said','two year','mr mr','first time','year mr','said yesterday','next sunday','staten island' ,'year earlier','york time','would use','like mr','said statement','four year', 'said dr','last week','year ago','last month','told report', 'percent percent','three year','would like','said interview','year old','even though','said one', 'next week','ca nt','said ms','ask mr','said ms','told mr','son mr','state mr','tuesday pm', 'six month','day mr','said mani','new hampshire', 'said could','ask mr','lower manhattan','million american','recent months','said expect', 'show mr','york state','m sure','mr thompson','said believ','said also','next month','said peopl','five year','said told', 'time squar','north carolina','like mani','said plan','said time','said think','two day','sever year','believ mr','know mr', 'let alon','said two','today would','time mr','could go','said recent','said said','said still','recent year','one anoth', 'also found','four month','jersey citi','one time','say would','week ago','new york','fifth avenu', 'week last', 'new york time','new york ny','said ca nt','two year ago','said nt want','last four year', 'said last week','four year ago','new york mr','last three year','last two month','said d like', 'said nt know','sever year ago'

VIII. Appendix C

Accuracy: TP+TN/TP+TN+FP+FN

Error Rate: FP+FN/ TP+TN+FP+FN

Sensitivity, TPR: TP/TP+FN

Specificity, FPR: TN/FP+FN

BER, ½(TPR+TNR)

Where,

TP – True Positive

TN- True Negative

FP – False Positive

FN – False Negative

BER – Bayes' Error Rate