Vanessa (Ziwei) Xu
Professor Andrew Halterman
Text as Data
May. 11th, 2022

Final Project Report

Using Human Rights Texts to Predict Country Income Category

Economists associate a country's economy with many factors such as human resources, physical capital, natural resources and technology. However, the relationship between a country's economy or development and its human rights establishment and protection are often overlooked. In a 2017 research *Human Rights and Economic Growth*, a significant causal effect was established where "freedom and participation rights affect economic growth positively in the long term."

In my research project, I'd like to explore more in-depth how effective human rights texts published can be in predicting monetary values—countries' income classification derived from the national income levels. I'll use the Naive Bayes classification model to train over ten thousand existing human rights text data files in order to make income classification predictions with human rights texts.

Text Data:

My data input would be human rights articles published each year. I chose to use a text dataset from Harvard Dataverse developed by Christopher J. Fariss: *Human Rights Texts: Converting Human Rights Primary Source Documents into Data*. The dataset contains a large corpus of digitized primary source human rights documents published annually by agencies including Amnesty International, Human Rights Watch, the Lawyers Committee for Human Rights, and the United States Department of State.

Text Data Classification by Country Code:

After loading text data files, I generated a text corpus from them and extracted year and country code information from each file name using regular expression. After creating a dataframe with three columns containing text, country, and year, there were 13780 rows with text files. After grouping by country code, which I extracted using regular expression, I found that a lot of the files don't start with a standard three-letter ISO country code. With no faster way to do this, I decided to manually check if each file that did not start with a country code should be either deleted or name changed.

Text files such as "business_and_human_rights_2001_Human_Rights_Watch.txt", "child_soldiers_1999_Human_Rights_Watch.txt", and "globolization_comes_home_protecting_migrant_domestic_workers_rights_2007_Human_Rights_Watch.txt" are deleted because they cannot be categorized into a country. Files that start with a country's name are kept and had their name changed. For example, "switzeland_1983_Amnesty_International.txt" is manually changed to "CHE_1983.txt" and "cote_d_ivore_2010_Amnesty_International.txt" is changed to "CIV_2010.txt". There are also misspelled names such as "colimbia_1983_Amnesty_International.txt" changed to "COL_1983.txt". These files needed me to read into the text content instead of simply looking at the file names. There are also files that don't have a country name at all in their file names but are actually about a particular country. For example, "challanges_for_a_responsible_power_2008_Human_Rights_Watch.txt" is about censorship in China so I changed the name to "CHN_2008.txt". This made me read every file that doesn't start with a country name at least a few lines into the text to make sure every file is categorized correctly before doing any further tests or models. After deletion and name change, 13692 text files remained.

Validation Data

I initially wanted to directly use the GNI per capita data from the Work Bank Open Data source and I planned on using the income classification threshold to manually categorize each country each year with GNI per capita; However, this did not seem like a feasible idea because I found that there's a couple of mismatches between this data and the classification data World Bank has published. Therefore, I decided to use the classification data instead.

In the World Bank Analytical Classifications, countries were classified into four categories: Low income (L), Lower middle income (LM), Upper middle income (UM), and High income (H). The threshold started in 1987 and is different for each year. This is calculated using the GNI per capita in US dollars with the Atlas methodology. Since the validation data only start from the year of 1987, I got rid of input text data that were before 1987 and cut the dataframe down to 10744 rows in order to save time and space for future model fitting efficiency.

After loading the classification data from local directory, I converted the data using the melt function from a matrix to a long table where columns are country code, year, and class. I then cleaned the dataset by removing "X" from all entries in the year column, dropping rows with ".." as class, and dropping classification data after 2014 where no text data exists.

Then I did a left join to merge the text dataframe with validation classification values and dropped rows with empty classification values. The dataframe after merging contains 10472 rows with validation classification values as the fourth column. Then I found that there are two rows with class "LM*" and replaced them with "LM".

Text Data Distribution

After merging text and validation data and dropping irrelevant columns and rows, the remaining dataset dimension is 10,472 rows with 4 columns: country, year, text and class. The distribution of data by class is listed in Figure 1 below. Low income (L) group has the most articles of 3340 and upper middle income group (UM) has the fewest articles of 1875.

Class	Low income (L)	Lower middle income (LM)	Upper middle income (UM)	High income (H)
Proportion	0.3189458	0.3084416	0.1790489	0.1935638
Count	3340	3230	1875	2027

Figure 1

However, only looking at distribution by class isn't representative enough since they are also categorized by year. Over the years, the proportion of each group can drastically change. Below is a figure of the number of articles over time (between 1987 when the validation data started and 2014 when the text data ended).

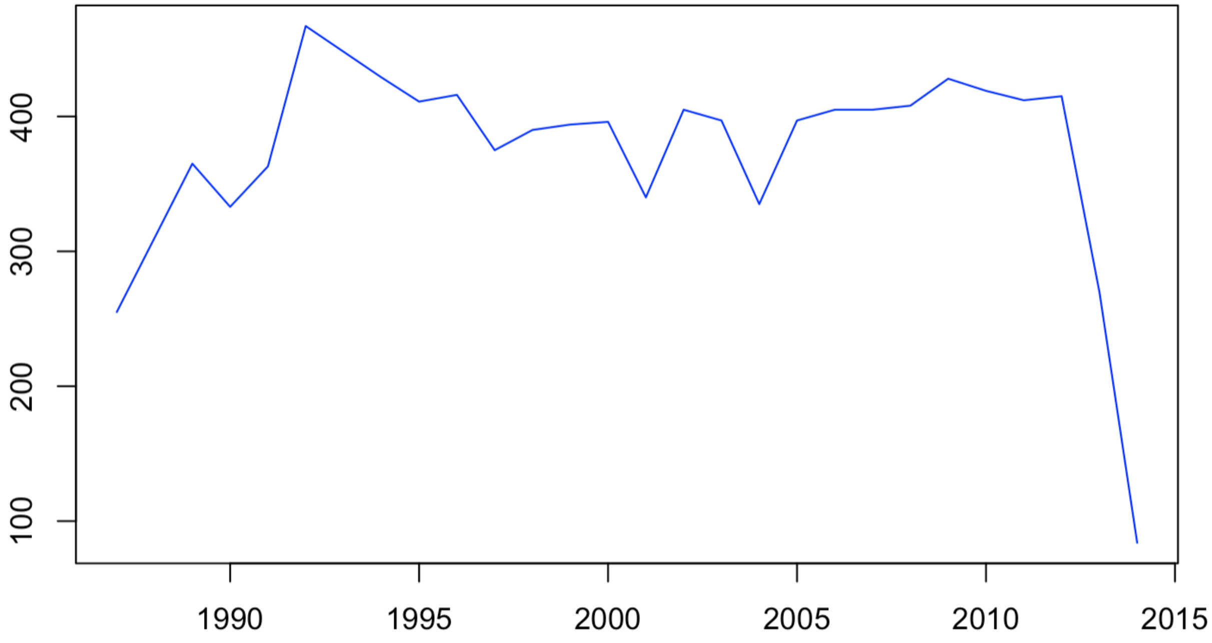

Figure 2

On average, 374 human rights articles published each year are included in this dataset between 1987 and 2014. There's an upward trend between 1987 and 1992 and remaining steady until 2012, finally the number of human rights documents decreases drastically in the following two years.

From the distribution of class and number of text files over time, we have not observed any highly imbalanced data input. However, I found that the number of documents are very imbalanced over the years after grouping by both country and year. There are multiple reasons behind this: there are a lot of missing validation data in certain country-year combination, and the number of human rights articles published differ by country and year as well. This problem may result in a lower accuracy for our classification model.

Text Pre-processing

After merging input text files with validation classification data, I did some text pre-processing. Pre-processing choices I made include converting to lowercase, replacing apostrophes with empty string, stemming, removing punctuation, removing stop words, and removing numbers. The reason for these choices would be that I want to get rid of as many unnecessary words as possible and use only features that are transformed from derivationally related forms to a common base form.

I then tokenized, filtered, and made features into a dfm for both training and testing dataset. I matched test set dfm to train set dfm features before implementing Naive Bayes classification, then converting the dfm to dataframes for future model fitting.

Naive Bayes Multiclass Classification Model

I've decided to use Naive Bayes Classification for this project. Naive Bayes is a classification algorithm designed to predict the class. It is a supervised learning technique where input data is labeled by class and trained in order to predict unlabeled data. Conditional independence is assumed given the class value, which is highly unlikely in real world problems.

However, they do tend to work surprisingly well despite the unrealistic assumptions made. In this project, text data would be used to train the model in order to predict the labeled data of four classes: H, UM, LM and L.

After a few times of failed model fitting, I found that the data frame is too large and would constantly cause errors. I decided to trim the sparse dfm by setting the minimum number of documents features occur to 0.1%. This successfully solved the problem, and I trained a Naive Bayes model on the training set using Laplace smoothing. Laplace smoothing is used in order to prevent the posterior probabilities to become zero when a word occurs zero times. Beyond avoiding computational difficulties of zero, the reason why smoothing is implemented is that there could be a lot of other words that match this category but only one mismatch would throw this whole category to 0. By adding one to the numerator and the size of the vocabulary to the denominator, we can change 0 to a rather small possibility which would not alter the result but producing a much more meaningful result.

In order to test for the best training size, I used a for loop for models with different training sizes ranging from 50% to 90% with 10% increments. I also tried setting dfm_trim with the parameter min_docfreq set to 0.0005. Including more features, running time increased significantly. However, the code managed to run successfully and below figure shows the result of two models with min_docfreq set to 0.1% (blue line) and 0.05% (pink line).

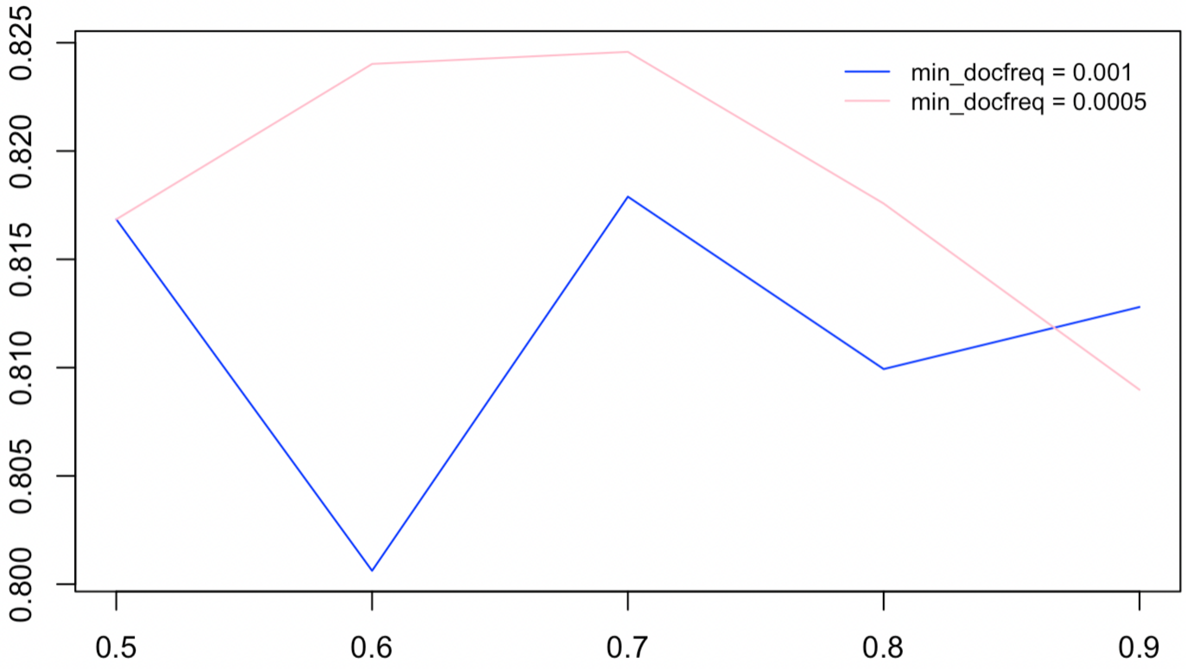

Figure 3

The x axis represents the training size from 50% to 90% and the y axis is the accuracy after evaluating the model on the test set. With the exception of the training size of 90%, accuracy of the model is higher with min_docfreq set to 0.05% because there will be fewer words excluded in feature selection.

The highest accuracy achieved from all models trained is 82.458% (rounding to three decimal places) where training size is 70% and min_docfreq set to 0.05%. The confusion matrix where the highest accuracy achieved is shown below.

	High income (H)	Low income (L)	Lower middle income (LM)	Upper middle income (UM)
High income (H)	538	7	14	33
Low income (L)	11	879	95	6

Lower middle income (LM)	28	133	752	82
Upper middle income (UM)	52	10	80	421

Figure 4

The accuracy calculation derives from the sum of all True Positive and True Negative divided by sum of all True Positive, False Positive, True Negative, and False Negative. I chose this evaluation metric because accuracy gives an overall measure of how effective the model is in correctly predicting class on the entire dataset. With the class distribution listed above in Figure 1, the classes are not highly imbalanced where one class is highly populated whereas others with only few units. Therefore, I believe accuracy would be an effective evaluation metric.

Conclusion & Future Work

I believe that the Naive Bayes Classification model with Laplace smoothing was a good classifier using human rights text data to predict the income classification of countries. This proves once again that there's a strong correlation between human rights establishment or protection and a country's income level. This research can also be extended to more recent human rights files. After establishing this relationship, in future work, more research can be done to predict countries' poverty level and extreme poverty headcount using human rights texts. Topic modeling can also be explored in order to find out what human rights issues cause low income countries to stay low. This will a particularly meaningful and impactful research area where more research should be done.

Bibliography:

Christopher J. Fariss; Fridolin J. Linder; Zachary M. Jones; Charles D. Crabtree; Megan A. Biek; Ana-Sophia M. Ross; Taranamol Kaur; Michael Tsai, 2015, "Human Rights Texts: Converting Human Rights Primary Source Documents into Data", https://doi.org/10.7910/DVN/IAH8OY, Harvard Dataverse, V3

"How Does the World Bank Classify Countries?" *How Does the World Bank Classify Countries? – World Bank Data Help Desk*, https://datahelpdesk.worldbank.org/knowledgebase/articles/378834-how-does-the-world-bank-classify-countries.

Koob, Sigrid Alexandra, et al. *Human Rights and Economic Growth*. https://www.humanrights.dk/sites/humanrights.dk/files/media/migrated/final_human_rights_and_economic_growth_-_an_econometric_analysis.pdf.