Final Paper

Arun Varghese

April 28, 2016

1 Introduction

This paper attempts to quantify and measure the political bias of newspapers in the South-Western Indian state of Kerala, written in the Malayalam language. This is done in the style of Gentzkow and Shapiro 2010 [1]. Gentzkow and Shapiro create an index to measure the similarity of news outlet language to the language of congressional Democrats and Republicans. I will perform a similar exercise here using the text from the proceedings of the Kerala Legislative Assembly and the text of online Malayalam news articles.

2 Background

2.1 Kerala and its Politics

Kerala is a state on the South-Western tip of India, formed in 1956 after independence through the merger of multiple Malayalam-speaking regions of British India. Malayalam is an agglutinative, South Indian language of the Dravidian language family.

Politically, Kerala is notable for strong leftist leanings. In 1957, it become the first political state in the world to democratically elect a communist government. Communist parties remain active and government in the state is dominated by two coalition fronts: the Communist Party of India (Marxist)-led Left Democratic Front (LDF) and the India National Congress-led United Democratic Front (UDF). Since the state's founding, these two coalitions have alternatively held majority power in the legislature. This provides a nice analogue to the two ended bias spectrum (Republican versus Democrat) used in Gentkow and Shaprio 2010.

2.2 Niyamasabha

The law-making body of the state, the Kerala Legislative Assembly, is known as the Niyamasabha (Malayalam for hall of laws). The Niyamasabha consists of 140 elected members known as Members of Legislative Assembly (MLA), each representing one of the 140 constituencies of Kerala state. The current assembly is the 13th since the formation of Kerala state, with the last general assembly election occurring in 2011 [2]. Since 2011, the UDF has held majority power in the Niyamasabha.

2.3 Kerala news media

Print media is widely produced and consumed in Kerala. Dozens of newspapers are published in the state, both in Malayalam and in English. Many of these newspapers have begun publishing online versions. This will be my source for news article text.

It is also a widely-held belief that the various Malayalam newspapers have their own political leanings [5]. Indeed some newspapers have, in their past, been affiliated with certain political parties and others are officially owned by political parties themselves. It will be interesting to see if my automatic method produces results in line with public opinion.

2.3.1 Malayalam newspapers

For this project, I will analyze the text of the three largest Malayalam-language newspapers: Malayala Manorama, Mathrubhumi, and Deshabhimani. The Malayala Manorama is the largest circulating newspaper in Kerala. Wikipedia lists its political alignment as 'Centre-right' where 'right' means an alignment with the Indian National Congress/UDF [6]. Mathrubhumi is the second most widely read newspaper in Kerala [7]. Wikipedia does not list a political alignment but some online discussion suggests that it too is UDF-leaning [8]. Finally, Deshabhimani, the third most widely read newspaper in Kerala, is owned by the Communist Party of India (Marxist) and appropriately, Wikipedia lists its political alignment as 'Left-wing, Marxist' [9].

3 Data

3.1 Legislative text

The text of statements given by MLAs in June 2015 and February 2016 were scraped from the Niyamasabha's official website where partial transcripts of legislative speech are available [3]. These time periods where chosen for convenience. Specifically, we take statements given during the 'Question and Answer' period of recent assembly sessions. During this 'Question and Answer' period, MLAs are allowed to submit open questions to the floor. Typically these statements are submitted by more than one MLA and multiple questions (usually 3 or 4) are posed in a single statement. All together, I collected 4321 statements. Transcripts from other periods of assembly sessions are not readily available and are more difficult to collect. So here, I will use only 'Question and Answer' text. For the purposes of this project, I will proceed assuming that this does not bias my sample of legislative language.

3.2 News article text

The text of news articles published in April 2016 were scraped from the online versions of various Malayalam-language newspapers. Headline news articles and politics articles were specifically targeted. Each article will be considered a document. Approximately 20 articles will be collected for each newspaper I examine.

4 Procedures

Using text from transcripts of legislative speech, I wish to build a model that predicts the political slant of a piece of text. My goal then is to use this model to assign a political slant score to Malayalam newspaper articles. I approach this problem in two, separate ways. First, I build a regression model to assign a continuous slant score to news articles. The average slant scores of a newspaper's articles will be its slant score. As an alternative, I build a classification model and use the class probabilities as a slant score for news articles. Then, I again average these individual slant scores to arrive at a newspaper-level slant score.

4.1 Constructing a label

To train my model, I need to label my legislative text with its political slant. For my classification model, this is straightforward. MLA party/coalition-affiliation is available online so I can match each statement directly to a coalition. Empirically, MLAs always submit statements with members of the same coalition so I do not have any issues where multiple coalitions are mapped to a single statement.

For my regression model, I need a continuous measure of slant. To do this, I will look at the MLAs who submitted each statement. Vote share data by constituency from the 2011 assembly general election is available online. I will match each MLA to their constituency and assign them a slant score between 0 and 1 equal to the vote share in their constituency. A slant score equal to 1 will correspond to a constituency voting 100% for a UDF-coalition party while a slant score of 0 will correspond to a constituency voting 100% for a LDF-coalition party. In this way, I will obtain a continuous slant score for each MLA.

This method is motivated by the 'Median Voter Theorem' [4], from formal political science. This theorem states that a majority voting system will elect the candidate most preferred by the median voter and thus the winning candidate's political leanings should be roughly reflected by the vote share in his constituency. Because most of my legislative statements were submitted by multiple MLAs, I will score each statement according to the average slant score of its submitters.

4.1.1 Dataset balance

After assigning slant scores, I observed that a majority of documents in my corpus were submitted by LDF-affiliated MLAs. In fact, only 24% of my documents were submitted by UDF MLAs and the average slant across all documents was 0.45, slightly LDF-leaning. Given that UDF is the current coalition in power, I hypothesize that this may be because LDF MLAs submit questions and concerns to the floor more frequently while UDF MLAs are more conservative in submitting questions to the floor.

This imbalance is potentially harmful to my predictive models but I found that it did not affect performance during model evaluation in my regression task. It did however lead to a poorer model in my classification task. Therefore to train my classification model, I constructed a balanced training set by randomly undersampling documents of the LDF class such that they were equal in number to UDF class documents. I trained my regression model on the original, unbalanced training set.

4.2 Feature extraction

All words, bi-grams and tri-grams and their frequencies were extracted from my legislative text corpus, vectorized and labelled with the MLAs who submitted the statement as well as the corresponding slant score. As mentioned before, each statement consists of multiple questions. Here, I consider the text of each statement a document.

To the best of my knowledge, a Malayalam stop-word list is not readily available online. Therefore, I constructed a novel stop-word list including words in my corpus appearing more than 50 times. This figure was calibrated to try and remove words

I manually identified as stop-words. These stop-words were not included in the n-grams extracted above. Words appearing only once in my corpus were also removed as they would likely only add noise to my models. In fact, removal of these once-appearing words was a larger reduction of my vocabulary than the removal of stop-words. This is possibly due to Malayalam's agglutinativity, as it allows unique words to be produced more easily.

4.3 Feature selection

Chi-squared feature selection was performed on my n-gram features. In chi-squared feature selection, a chi-squared test for independence is performed for every combination of n-gram feature occurrence and coalition-labeling. Intuitively, the chi-squared test checks whether an n-gram and coalition-labeling co-occur more often than would be observed by chance, given their respective frequencies in the dataset. This allows us to quantitatively assess which n-grams are most representative of each coalition label. In the regression task, the top 132,000/134,315 most informative n-grams were selected and the rest dropped. In the classification task, the top 59,000/65,124 most informative n-grams were selected and the rest dropped. These cut-offs were manually calibrated to maximize performance during model evaluation.

After chi-square feature selection, my n-gram count matrix was normalized with a TF-IDF weighting. TF-IDF stands for term frequency inverse document frequency and is a score assigned to each term or n-gram in a document. It is equal to that n-gram's frequency within the document multiplied by the inverse document frequency, one over the number of documents that n-gram appears in. This multiplication by IDF means that n-grams which appear in many other documents, and therefore are common and less informative, have their impact weighted downwards towards zero.

4.4 Models

4.4.1 Regression model

Again, the target for my regression model was the continuous UDF-ness slant score derived from vote share data. The training set for this regression model was the unbalanced, normalized count matrix described in section 4.3. I trained a few different models using various regression techniques and used 5-fold cross-validation to evaluate each, using r-squared, the coefficient of determination, as my performance measure.

In 5-fold cross-validation, the data is randomly partitioned into 5 parts. Then, a classifier is trained on 4 of the parts and tested on the last. This is repeated such that all 5 of the partitions are tested on once. We can evaluate the performance of the model by taking an average performance measure across the 5 runs. If the model is overfit to any particular portion of the dataset, this is penalized in the average performance figure. In the end, a simple linear regression model performed the best and achieved an average r-squared of 0.19 across the 5 cross validation runs.

4.4.2 Classification model

The target for my classification model was each document's coalition-affiliation. This was binary, either UDF or LDF. The training set for the classification model was the same count matrix used for the regression model except balanced on class as described in section 4.1.1.

Again, I trained multiple models using different classification techniques and used 5-fold cross validation to evaluate each one. Here, my performance measure was accuracy, the percentage of observations correctly classified. The best performing model was one trained using Logistic Regression and achieved an accuracy of 0.77– an improvement over the 0.50 that could be scored by randomly guessing.

4.4.3 Model training

After selecting final regression and classification models, I fit each on their respective training sets. Then for each model, I mapped my corpus of news article text onto the count matrix space of the appropriate training sets. This meant, for each article, filling out a vector marking the count of each n-gram identified during feature selection and then weighting this count by to the inverse document frequency calculated on the training set. I did this separately for the corpus of each of the three newspapers I analyzed. Finally, I tested my models on this newspaper text data and predicted political slant scores.

5 Results

A table reporting the slant expectation for each newspaper as well as the slant score predicted by both the regression and classification models is given below.

Newspaper	Slant expectation	Regression model	Classification model
Malayala Manorama	UDF bias	0.44	0.46
Mathrubhumi	UDF bias	0.44	0.44
Deshabhimani	LDF bias	0.44	0.44

As can be seen above, there was almost no variation in the predictions of either model. Except for one case, the predicted slant for every newspaper was approximately 0.44, a LDF-leaning score. Perhaps this score would be reasonable for Deshabhimani because we expected it to have an LDF bias but because all newspapers have this score, it appears that these models are just not picking up any signal in the data.

I also calculated predictions using some of the models that I set aside during cross-validation. Some of these models shifted scores to be more UDF-leaning but did so for every newspaper–there was still no variation in the predictions made. Running the model with less strict feature selection led to similarly uninformative predictions.

This robust failure in prediction suggests fundamental issues with the input data rather than with model building. In the section below, I explore a few reasons why these models lacked predictive power.

6 Discussion

The first explanation is that the input data, the text from 'Question and Answer' sessions did not contain biased words or phrases. This seems possible. The statements given were not very long and often contained many procedural and general question words–there may not have been a lot of room to use politically charged words or phrases. It's also possible that the 'Question and Answer' session is a fairly neutral time during a legislative hearing and that MLAs do not use highly partisan language. I did not find any strong evidence pointing in either direction during my background research.

A second explanation is that my training dataset was too small. I collected only 4321 statements that all together, were about 8MB in size. This was done on purpose to limit the computational burdens of this project. It could be that I have not collected enough legislative text to identify words or phrases with strong political slant signal.

Finally, it is possible that this specific method of bias detection is not appropriate given Malayalam's linguistic structure and the fact that a Malayalam word stemmer does not exist. Malayalam does follow western conventions for whitespace but because it is agglutinative, some words take compounded forms without whitespace in between and it is hard to isolate the root words. This means that a word that is truly predictive of political slant might occur in multiple, slightly different forms. This method would be unable to aggregate those occurrences and would lose the power to pick up on its signal. Stemming, transforming a compound word to its root words, would have been a solution to this problem. However, I was not able to find or build a Malayalam stemmer that functioned well.

7 Conclusion

In this paper, I attempted to predict the political slant of Malayalam-language newspapers by comparing the text of news articles against a corpus of text taken from the proceedings of Kerala's legislative body, which I was able to label with a political slant score. I built two predictive models, one treating the task as a regression problem and the other treating the task as a classification problem. Neither model was able to produce slant predictions that agreed with expectations. In fact, neither model produced predictions that appeared meaningful at all. This lack of predictive power is possibly due to the training dataset being uninformative or too small or due to the linguistic nature of Malayalam requiring different analysis methods.

I remain interested in the goal of this project, measuring political bias of Malayalam newspapers, so in future work I hope to address some of the shortcomings of this project and build on the structure I have laid out here. More generally, I am excited by the application of natural language processing methods in social science contexts so I will carry lessons learned over the course of this project into subsequent work.

References

[1] Matthew Getzkow and Jesse Shapiro. *What Drives Media Slant? Evidence from US Daily Newspaper.* Econometrica, Vol. 78, No. 1 (January, 2010), 35-71

[2] Kerala Legislature,
`https://en.wikipedia.org/wiki/Kerala_Legislature`

[3] Niyamasabha,
`http://www.niyamasabha.org/`

[4] Randall G. Holcombe. `Public Sector Economics` Upper Saddle River: Pearson Prentice Hall, p. 155.

[5] Lijo Isac's Answer to Malayalam language: Is Malayala Manorama politically biased?
`https://www.quora.com/Malayalam-language/Is-Malayala-Manorama-politically-biased`

[6] Malayala Manorama,
`https://en.wikipedia.org/wiki/Malayala_Manorama`

[7] Mathrubhumi,
`https://en.wikipedia.org/wiki/Mathrubhumi`

[8] Mathrubhumi Reviews,
`http://www.mouthshut.com/review/Mathrubhumi-review-snltltmrsm`

[9] Deshabhimani,
`https://en.wikipedia.org/wiki/Deshabhimani`