Evaluating Gender Equality in China's Parental Leave Policies through Supervised Machine Learning

Xiao Sun

August 5, 2024

I aim to measure "gender equality" in news coverage of China's parental leave policies. I used my dataset, which includes news articles from 2015 to 2024 from three mainstream news outlets: Global Times, China Daily, and Xinhua News Agency. The dataset consists of the full text of each news article, the date, and the media source. I converted the articles into quasi-sentences and saved all the information as a CSV file.

I applied two random forest (RF) models to this task—one using bag-of-words vectorization (Model 1) and another using word embedding vectorization (Model 2). Below is a detailed description of the process for each model:

(1) I created a binary variable to identify whether an article mentions equal-sharing of care responsibilities between men and women. (2) I simulated supervised learning classification by manually coding a stratified sample of the text data from the three news outlets. (3) I tokenized the sample text data using the bag-of-words method, treating similar and dissimilar words as unique tokens. (4) I split the data into training, validation, and test sets. (5) I used the training data to build the RF model through the RandomForestClassifier function. (6) I tuned and evaluated the model using the validation set according to the best set of parameters. (7) I evaluated the model's true out-of-sample performance using the test set. (8) Finally, I applied the model to the full dataset.

Different from step (3) in Model 1, I used the word embedding method for Model 2, assuming that the meaning of a word is given by its contexts. This method reduces the high-dimensional sparse bag-of-words vectorization to lower dimensions while maintaining the semantic and syntactic relationships between words. I used Doc2Vec to build vocabularies and train the model, checking word and document similarities to validate the word embedding method. The similarities were very high, close to 0.99, and the documents shared similar agendas. Except for step 3, other steps were similar to those for Model 1.

To compare the performance of the two models, I checked their accuracy, precision, recall, F1 scores, confusion matrices, and ROC curves. The scores in Table 1 indicate that, overall, Model 1 outperforms Model 2 in terms of prediction precision, classification accuracy, and balanced performance.

Table 1. Two models’ scores on validation set

|  |  |  |
| --- | --- | --- |
|  | Model 1 | Model 2 |
| precision | 0.8333333333333334 | 0.5263157894736842 |
| recall | 0.3333333333333333 | 0.6666666666666666 |
| accuracy | 0.6666666666666666 | 0.5757575757575758 |
| f1 | 0.47619047619047616 | 0.5882352941176471 |

Model 1 has a higher ROC AUC score (0.77) compared to Model 2 (0.60), indicating better overall classification performance and reliability in positive predictions. Considering their confusion matrices, Model 1 has high True Negatives (17) and low False Positives (1), indicating that the model is effective at correctly identifying negatives. In contrast, Model 2 has balanced True Positives (10) and False Negatives (5), indicating that the model has better recall, effectively identifying more actual positives. Given the context of predicting gender equality, if capturing all relevant cases of gender equality is crucial (e.g., in social research or policy-making where missing a positive case could be detrimental), Model 2 might be more suitable despite its lower precision and accuracy. However, if precision and overall classification performance are more critical, Model 1 would be the better choice.

Figure 1. Model 1 performance

A blue and white chart

Description automatically generatedA graph of a curve

Description automatically generated with medium confidence

(a) confusion matrix (b) ROC curve

Figure 2. Model 2 performance

A blue and white squares with numbers

Description automatically generatedA graph of a curve

Description automatically generated with medium confidence

1. confusion matrix (b)ROC curve