Lexically Healthy: Using Text to Predict Health Bill Outcomes

Michelle Manning

UMassAmherst

Data Analytics and Computational Social Science Program

Context & Goal

The 116th Congress introduced 909 health bills in the House, but only a fraction of a percent became law. Understanding the lexical topics found within passing bills and the those that don't will enable lawmakers to more effectively create legislation that will be successful & improve healthcare.

Research Questions: Can we predict the outcome of a bill based on the topics found within the bills?

Methods & Model

In this analysis, I web scraped 739 healthcare related bills introduced in the house from congress.gov.

To discover the topical patterns in the bills, I performed a Structural Topic Model (STM).

Taking the proportion of each bill in each document, Θ, I ran multiple models to find which most accurately predicted the bill passing or not. I crossvalidated each model by running it 100 times with 10 folds. Random Forest was the best model, which I then used to predict the test set.

Results

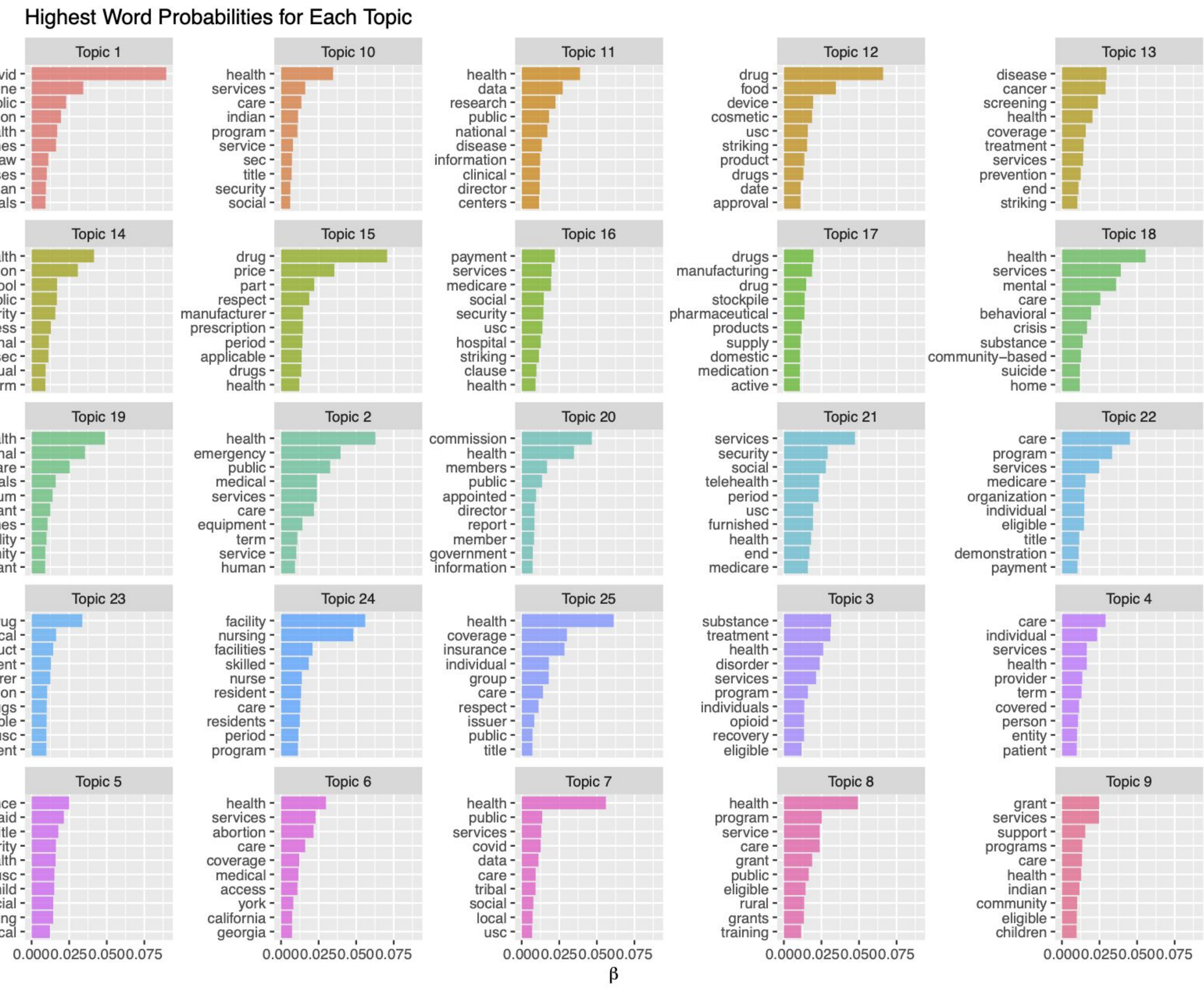

Fig 1. Word frequencies in each topic

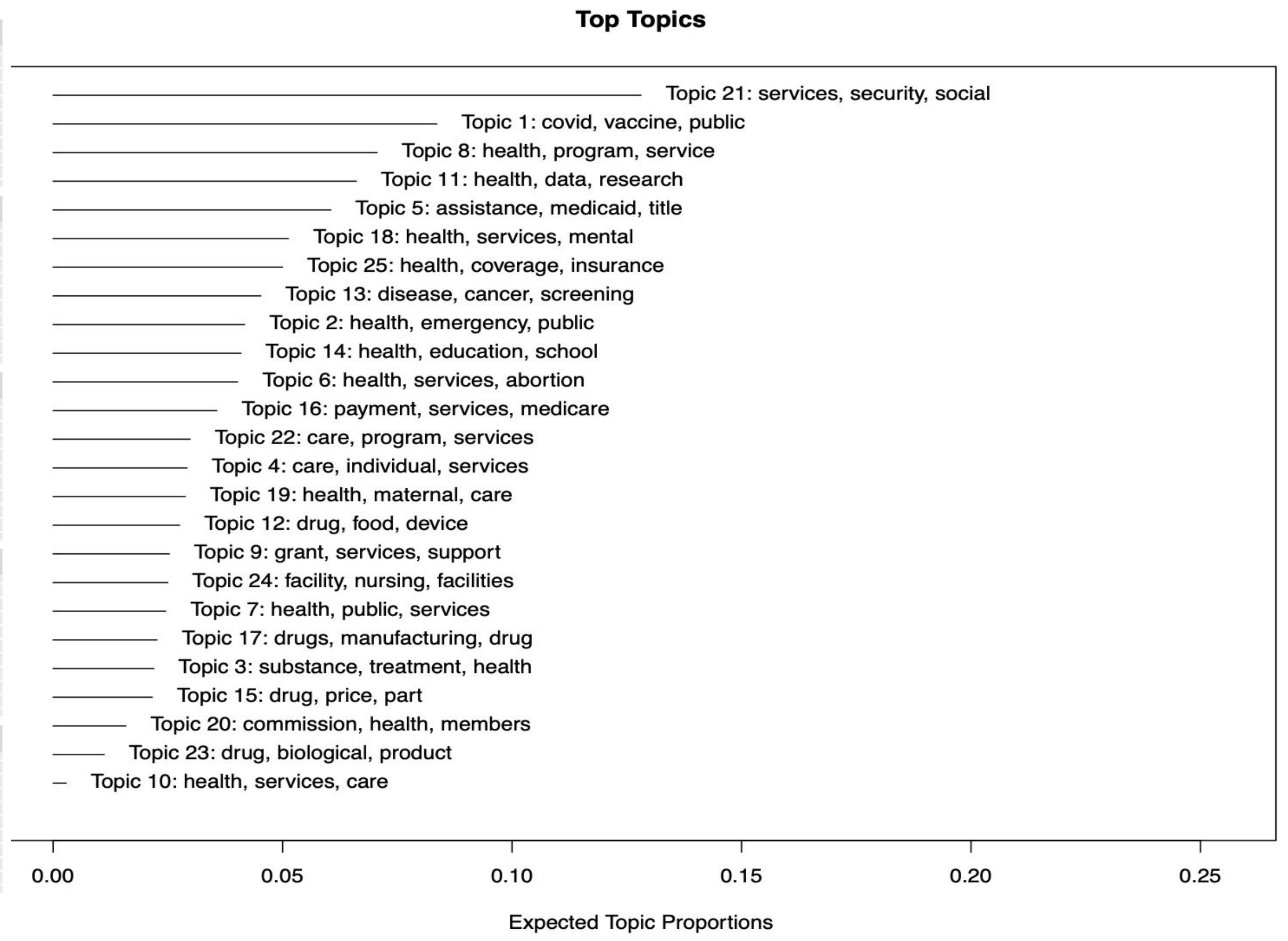

Fig 2. Topic frequency across corpus

Fig 3. Accuracy & Kappa plot of all the models.

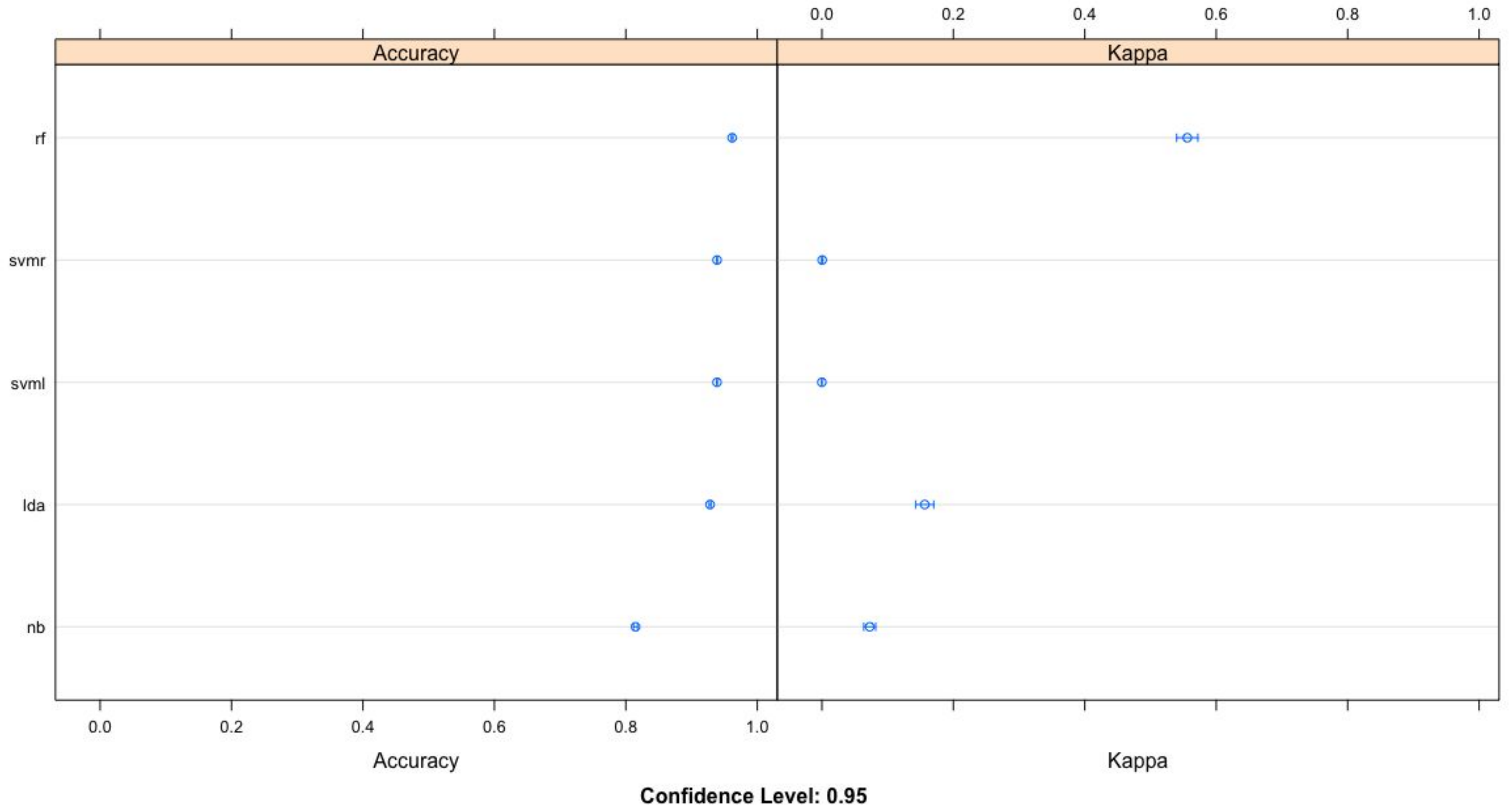

Table 1. Random Forest mtry summary

mtry	Accuracy	Kappa
2	0.9405542	0.04206502
13	0.9617895	0.55592357
24	0.9611727	0.58852032

Table 2. Confusion Matrix

	0	1
0	172	5
1	1	6

Table 3. Random Forest Summary. Note the high Precision, Recall, & F1 scores.

Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision
0.9884	0.6364	0.9771	0.7778	0.9771

Recall	F1	Prevalence	Detection Rate	Detection Prevalence
0.9884	0.9828	0.9402	0.9293	0.9511

Model Comparison

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
lda	0.8364	0.9107	0.9286	0.9282	0.9455	1	0
nb	0.6182	0.7818	0.8182	0.8148	0.8545	0.9464	0
svmr	0.9286	0.9286	0.9455	0.9388	0.9455	0.9636	0
svml	0.9107	0.9286	0.9455	0.9387	0.9455	0.9464	0
rf	0.8929	0.9464	0.9636	0.9618	0.9818	1	0

Table 4. Accuracy outcomes for models. The SVMs were close, but RF was most accurate.

Conclusions

The Random Forest was 96 % accurate based on the proportion of topics found. With a Precision, Recall, & F1 scores above 97%, the Random Forest model was able to predict pretty well the outcomes of bills based on bill topics.

Future Research

This model focuses on the lexical patterns that affect a bill passing. Other factors could influence this are the party in power, the bill sponsor's party, or the year it was introduced. Incorporating these factors into the model could make it more useful.

References

- https://machinelearningmastery.com/machine-learning-in-r-step-by-step/
- https://juliasilge.com/blog/evaluating-stm/
- https://machinelearningmastery.com/machine-learning-in-r-step-by-step/
- https://community.rstudio.com/t/how-to-prevent-data-leakage-between-training-and-test-set-when-i-have-repeated-measures/30666

Acknowledgements

Dr. Justin Gross, Omer Yalcin,, Larri Miller, Leah Dion & the Slack Study Group

Learn more about me