Bringing Back Black Boxes: Classification of TV news using neural nets *

Stanford CS224N Custom Project

Shun Yamaya
Department of Political Science
Stanford University
syamaya@stanford.edu

Jennifer Wu
Department of Political Science
Stanford University
jwu19@stanford.edu

Abstract

In this project, we aim to 1) measure the partisan slant and 2) classify topic of news television shows. For partisan slant, we fine-tune a RoBERTa model with the text from members of Congress and presidents, using the partisanship of the speaker to label the partisan slant of the text. To categorize the topic of the news, we also use a RoBERTa and fine-tune it with topics derived from Vanderbilt Television News Archives' one-sentence summaries. We find that these neural net models perform slightly better (or currently on par, as it is still training) than typical methods used in the social sciences, such as ridge and elastic net.

1 Introduction

Television has been America's premier news source for decades (Mitchell et al., 2016; pew, 2021). Social science research consistently finds that what TV news says and how they say it strongly affects people's political issue priorities, voting behavior, and values at large (Iyengar and Kinder, 2010; DellaVigna and Kaplan, 2007; Kim, 2023). Thus, a long literature has sought to measure the types of topics the media covers (White, 1950; Funkhouser, 1973) and how "partisan" the tone of the text is (Martin and Yurukoglu, 2017). We seek to advance this line of work of automated classification and measurement of political text using deep learning techniques learned in CS 224N.

2 Related Work

Our work lies in the intersection of social sciences and natural language processing. First, to categorize the topic of text, the predominant approach in the social sciences using topic models to learn lower-dimensional representations of documents (Roberts et al., 2016), but the interpretability of such "topics" can be quite opaque. For example, the model will frequently learn correlations in speech patterns that do not appear to relate to substantive topics of interest, but are ultimately difficult to verify. Moreover, specifically in the domain of news reporting, topic models often behave more like "event detection algorithms," where it discovers events (e.g. a famous plane crash) rather than a substantive topic (e.g. accidents). This discrepancy is because of the nature of the algorithm being unsupervised. Yet in practice, analysts are often more interested in the latter over the former.

Few scholars have implemented supervised machine learning models on n-gram representations of media text, and to a much lesser extent, neural methods on embedding representations (Turkel et al., 2021). This is in part because of the lack of consistent topic labels to media text. We make progress on this by generating topic labels from the Vanderbilt Television News Archive (VTNA), which employs human coders to provide keyword summaries of nightly news segments. We group these short summaries into 20 substantive topics and fine-tune a pre-trained neural net on these labels to classify television segments. Our goal is to straightforwardly use deep learning models to obtain more accurate estimates of topic categories, and benchmark those predictions against predictions from classic machine-learning models such as ridge, lasso, and elastic net.

In our second contribution, we measure the "partisanship" of a text using neural nets. When measuring a latent trait like "partisanship" from language, political scientists and economists have asked the following thought experiment—given some text, how likely would a human coder think it was spoken by a Republican or Democratic politician? (Peterson and Spirling, 2018; Gentzkow et al., 2019) Previous work uses a machine learning models to learn words (or n-gram features) that are highly predictive of party labels. Thus, when predicting the "partisanship" of text coming entities without formal party labels like media companies, Martin and Yurukoglu (2017) train a statistical model on Congressional floor speeches that predicts the partisanship (or vote-based ideology measure) of the speaker, then applies the trained model on new blocks of text.

In principle, our method extends that of Martin and Yurukoglu (2017) by applying a neural method to same problem, but our approach also differs in a few, important ways. First, while we train our model on binary party labels, Martin and Yurukoglu (2017) train their model on congressional votes-based ideology estimates (Lewis et al., 2022) [2]. In political science and social sciences broadly, such ideology estimates are popular, low-dimensional summaries of congressional voting records which people interpret as to be the traditional left-right political spectrum. We do not train our model on these vote-based measures, partially for model interpretability. However, we use these vote-based measures for validating our model, by correlating our neural partisan measures against existing political ideology estimates. Political ideology should be correlated with both congressional voting patterns and speech in Congress. Second, we expand the training data beyond Congressional floor speeches (1994-2022). We include the set of political speech to press releases and electronic communications by Members of Congress as well as Presidential speeches and addresses. These data efforts will be detailed in the next section, but this greatly increases the number of training data with party labels. Finally, Martin and Yurukoglu (2017) restrict their model to the 1000 phrases most predictive of partisanship, as obtained from Gentzkow et al. (2019). In our view, this is similar to a manual version of lasso, and since we already are predicting party labels we omit this step and employ a data-driven approach.

3 Empirical approach

3.1 Data sources and overview of processing

Internet Archives TV News Data: We use television closed captions from CNN, MSNBC, Fox News, ABC, CBS, and NBC. These data are our main target of interest—we want to measure the topic prevalence and partisan tone in these text. The first three companies are 24-hour cable news networks, and the latter three are traditional broadcast network companies. These data are sourced from Internet Archives TV News, and spans near-24 hour coverage between 2010 to 2022. We download the data via a researcher-restricted API—these captions contain both news and advertising content. To take out the advertising content, we merge the data with the Advertising Inventory Files developed by

[2] Martin and Yurukoglu (2017) also train a version of their model using binary classes in Appendix figure A5, but do not discuss it in length

GDELT. These files use the caption codes embedded in the video files to indicate advertising air time, which we use to delete words with corresponding time stamps. Finally, for the training data, we align and divide the captions with the segment start and end time stamps from the VTNA (more on this data source below). For cable news data (which we do not have the VTNA time stamps), we divide the captions where there is a 30-second or longer gap in news captions. Some manual spot-checking suggests that the 30-second threshold corresponds well with advertisements breaks.

Vanderbilt Television News Archives: To create topic labels, we use data from the Vanderbilt Television News Archives (VTNA). Since the 1970s, the VTNA has hired research assistants to watch news show from CNN (*Anderson Cooper 360* or *AC 360*), FNC (*The Fox Report with Shepard Smith* prior to 2013, and *Special Report with Bret Baier* on weekdays and *The Fox Report* on weekends after 2013), ABC (Daily evening news), CBS (Daily evening news), and NBC (Daily evening news) every day, and to record the exact start and end time for each segment at the second-level. The research assistants also provide keyword summaries of each segment. These summaries and exact timestamps (at the second level) allow us to break up news captions into segments that correspond exactly to a particular segment with a coherent topic, as judged by a human coder. This is especially helpful for US broadcast news (ABC, NBC, and CBS) because unlike cable news channels where each television segment (broken apart by ads) usually covers roughly one topic, broadcast news consecutively covers multiple topics over a short period of time. In manual spot checks by watching video segments on the internet archives, the aligning between the VTNA and caption data seems to be working well.

Moreover, we turn the short summaries in 20 coherent topics. These labels are provided below in Appendix A, Table 6 with a random sample of one-sentence summaries. These twenty labels were constructed by applying a low-dimensional ($k < 25$) topic model on the aforementioned topic-segmented captions to get a sense of the range of possible topics, then using a keyword dictionary method to categorize titles into bins. Here too, we randomly sampled 100 titles and verified that we agreed with all of the title-categories classifications.

Congressional and Presidential Data: For pre-training and fine-tuning a model to measure partisan tone, we further gather a variety of text where there is a clear partisan speaker associated. First, following the literature, we gather congressional floor speeches from `govinfo.gov` from 1994 until the end of 2022. Second, we collect the official e-newsletters from members of Congress, using `dcinbox.com` from 2009 until early 2023. Third, we collect Congressmember press-releases from January 1, 2021 until January 1, 2023 with the ProPublica Congress API. Finally, we have presidential proclamations, campaign documents, state of the union statements, and press statements from the 43rd to 46th presidents from `presidency.ucsb.edu`. We link each text in these four broad categories of political text to the party of their respective speaker. We only keep text by authors from the Democrat or Republican party.

3.2 Proposed Neural Method

For both classifying topic and predicting partisanship, we finetune a base RoBERTa model (Liu et al., 2019). We attempted to use the RoBERTa large model as well, but we did not have enough GPU resources to finetune it. Since RoBERTa takes in max 512 tokens, we chunk our data into segments with max 400 words, and record the prediction for those words. If greater than 512 tokens is produced for a chunk, we truncate it. In the future, we plan on chunking the data while keeping together sentences, and ensuring that we do not have to truncate any of our data. We would also like to train on different type of models, and implement continued pre-training on some documents.

Partisan neural net: We randomly sample 20% of congressmembers per Congress to hold out for the final evaluation. To finetune the RoBERTa base model, we do a 80-20 training-test split on the remaining data. We set a learning rate of $2\mathrm{e}{-5}$, batch size of 16, and use the AdamW optimizer. Because of the size of our dataset, we evaluate the model using classification error rate and at every 100 steps. There are 2 labels total – either Democrat or Republican. It takes roughly one to two and a half hours to run every 100 steps and evaluate. At 1600 steps, we switched to evaluating every 500 steps, which decreased the training time. Unfortunately we did not finish training the full model, so we present results based off of a model from step 6000, epoch .05. This model took around 36 hours to train.

Topic neural net: Likewise, we hold out 20% of our labeled segments for topic classification and do a 80-20 training-test split on the remaining segments to finetune with RoBERTa. We set a learning rate of $2\mathrm{e}{-5}$, batch size of 16, use the AdamW optimizer, and do this for 3 epochs. We set an early stopping patience of 5 with a threshold of 0. We evaluate the model using classification error rate and at every 500 steps. There are 20 possible labels. The entire model took roughly 10 hours to run.

4 Results

In the case of the topic classification task, the training data ($N = 57,215$) consists of labeled caption segments, with the labels coming from the VTNA dataset; for partisanship prediction, the training data are speeches and press releases from politicians ($N = 1,512,703$). For both tasks, we hold out 20 percent of our training data and train our models on the remaining 80 percent, and in particular for the partisanship task, we randomly sample 20% of congress members per Congress to hold out.

4.1 Topic Classification Results

We first report results from our baseline method: using ridge, lasso, and elastic net on a document term matrix of the training data. Punctuation and English stopwords are removed, and terms are turned to lower case and stemmed according to Porter method. Finally, words that don't appear in at least 50 documents and words that appear in more than 90 percent of documents are pruned.

Baseline Results: In the first three rows of table 1, we show the overall accuracy and the multi-class AUC value (Hand and Till, 2001) of our baseline models. We can see that the baseline models seem to preform similarly. In Appendix B, Table 7, we show the precision, recall, and specificity of the elastic-net model with $\alpha = 0.1$. The specificity of all groups for that model (lack of false positives) is high, but the precision and recall (proportion of true positives correctly predicted, proportion of true positives out of predicted positives) are both low across all categories.

Model	Overall accuracy	Multi-class AUC
Ridge	0.636	0.810
Elastic Net ($\alpha = 0.1$)	0.642	0.797
Lasso	0.624	0.801
Fine-tuned RoBERTa	0.708	0.841

Table 1: Accuracy statistics for topic models on hold-out set

Neural Results: Our final model had a training loss of 0.965, a validation loss of 1.141, and an evaluation accuracy of 0.667. To get holdout set predictions, we take the mean probability across chunks for a segment, and then assign the segment to the most probable topic. As seen in the fourth row of table 1, our fine-tuned RoBERTa model has a higher accuracy and multi-class AUC than all our baseline models when predicting on the hold-out set.

Still, we expected the model to perform better than 70% with the lowest recall of .164 for the "environment" category (see Appendix B, Table 8 for more statistics) on classifying the topic of news segment. This may be because we are chunking the text to make sure document information is not loss, and aggregating up at the end. The accuracy for the hold out set of non-chunked documents is .651 (roughly 8576 documents, or 60.0% of total documents in our hold out set), while the chunked data have a .793 accuracy. This discrepancy may be because the chunked data contains more meaningful words, however. We speak more about how we hope to address this in the future in the Evaluation section.

4.2 Partisanship Classification Results

Baseline: For the binary classification task, we show the results from the ridge regression. The overall accuracy is 0.691, with an AUC of 0.775. In Table 2, we present further accuracy statistics,

Category	Precision	Recall	Specificity
Democrat	0.668	0.713	0.715
Republican	0.715	0.670	0.668

Table 2: Classification accuracy statistics for ridge regression

Since we are ultimately interested in the continuous probability measure, we validate the probabilities by using the congressional votes-based ideology measure of the legislators that are held out. For each "test" legislator, we calculate the average probability their speech is categorized Republican. We correlate that against their voting score. For voting scores, we rely on two sources: DW-NOMINATE (Lewis et al., 2022) and MC3-GGUM (Duck-Mayr and Montgomery, 2023). Essentially, both take congressional vote records as starting points and extract some latent dimension where a lower value is interpreted to a more "liberal" voting pattern. DW-NOMINATE assigns one score to each legislator across time, while MC3-GGUM fits a more complex Bayesian model for each Congressional cycle to better account for legislators that behave in a more extreme fashion. For the MC3-GGUM estimates, we take the posterior means as the summary statistic. In Table 3 we show the raw correlation between the measures. In Appendix B, Figure 1 we plot each legislator's DW-NOMINATE score against their average predicted partisanship from text.

	DW-NOMINATE	MC3-GGUM
Avg. prob. text is Republican	0.096	0.103

Table 3: Correlations between ideology measures for ridge

	MC3-GGUM
Prob. text is Republican	2.274 (1.064)
Num.Obs.	1.062×10^3
R2	2.5×10^{-2}

Table 4: MC3-GGUM regressed on ridge predictions

Finally, in Table 4, we regress MC3-GGUM against the text measure with Congress fixed effects. If our ridge-based text measure is working well, we would expect it to correlate positively with the vote-based ideology measure. However overall, all three experiments show that the ridge-based partisanship measure is only weakly, positively correlated with voting-based ideology.

Neural results: The model we report results from is from a checkpoint at epoch .05, and still running. This model has a loss of 0.528, an evaluation loss of .506, an evaluation accuracy of 0.723. On the hold-out set, we calculate an accuracy of .691 with an AUC of .692. As seen in Table 5, this model performs on par with the baseline models, but we hope it learns more as it continues training (currently we have not enough reached half an epoch).

Category	Precision	Recall	Specificity
Democrat	0.731	0.632	0.659
Republican	0.659	0.753	0.731

Table 5: Classification accuracy statistics for fine-tuned RoBERTa

5 Evaluation

Topics classification: As mentioned in the results, we were hoping for higher topic classification accuracy. Although the fine-tuned model's precision and recall were better for most topics than our baseline model's, they were still lower than we expected.

Our fine-tuned RoBERTa model seems to perform better on segments that were chunked. However, is is unclear if certain topics are correlated with more text, or if perhaps more texts just signal the captions are better quality. We plan to investigate in the future. Furthermore, we would like to think of other ways of aggregating probabilities from our chunks, or look into models that take more tokens. Given our smaller sample of labeled data for topics, we will also run models with more epochs with evaluation at the end of each epoch. Finally, since caption label data formatting can be strange, we would like to gather and continue pre-training our model on some captions data, such as local news captions data.

In addition, we plan to apply our classification model to a prediction set – the cable news data. We would then hire coders to categorize these segments and compare their codings to our model's. We would like to see if there are certain categories with more disagreement, and if these align with the disagreement we saw originally while training our model.

Partisan classification: In the future, we would like to finish running this model. We would do a similar evaluation of comparing it to pre-existing measures of partisanship for congressmembers, and also have a hold-out set which we rank statements ourselves in terms of probability, and then study how that compares to the model. Then, we would like to apply it on the news segments and get estimates for the partisanship of various segments.

6 Conclusion

We find that our fine-tuned models perform better at classification than other models usually used for text analysis in political science, such as ridge, lasso, and elastic-net. However, these models still do not perform as well as we have hoped, as neural nets have been known to have evaluation accuracies of upwards in the 90s.

One of our primary limitations was computational resources and the length of our data, especially for our partisan model. We hope to think of better ways of accommodating longer documents in the future, including trying different models, fine-tuning parameters, pre-training our model on relevant text, and refining our chunking process. In the future, we will apply these models on unlabeled text.

References

2021. Cable news fact sheet. [Online; posted 13-July-2021].

Stefano DellaVigna and Ethan Kaplan. 2007. The fox news effect: Media bias and voting. *The Quarterly Journal of Economics*, 122(3):1187–1234.

JBrandon Duck-Mayr and Jacob Montgomery. 2023. Ends Against the Middle: Measuring Latent Traits when Opposites Respond the Same Way for Antithetical Reasons. *Political Analysis*, pages 1–20. Publisher: Cambridge University Press.

G Ray Funkhouser. 1973. The issues of the sixties: An exploratory study in the dynamics of public opinion. *Public Opinion Quarterly*, 37(1):62–75.

Matthew Gentzkow, Jesse M. Shapiro, and Matt Taddy. 2019. Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. *Econometrica*, 87(4):1307–1340.

David J. Hand and Robert J. Till. 2001. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, 45(2):171–186.

Shanto Iyengar and Donald R Kinder. 2010. *News that matters: Television and American opinion*. University of Chicago Press.

Eunji Kim. 2023. Entertaining beliefs in economic mobility. *American Journal of Political Science*, 67(1):39–54.

Jeffrey B. Lewis, Keith Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin, and Luke Sonnet. 2022. Voteview: Congressional Roll-Call Votes Database.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Gregory J. Martin and Ali Yurukoglu. 2017. Bias in Cable News: Persuasion and Polarization. *American Economic Review*, 107(9):2565–2599.

Amy Mitchell, Jeffrey Gottfried, Michael Barthel, and Elisa Shearer. 2016. The modern news consumer. [Online; posted 7-July-2016].

Andrew Peterson and Arthur Spirling. 2018. Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems. *Political Analysis*, 26(1):120–128.

Margaret E Roberts, Brandon M Stewart, and Edoardo M Airoldi. 2016. A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515):988–1003.

Eray Turkel, Anish Saha, Rhett Carson Owen, Gregory J Martin, and Shoshana Vasserman. 2021. A method for measuring investigative journalism in local newspapers. *Proceedings of the National Academy of Sciences*, 118(30):e2105155118.

David Manning White. 1950. The "gate keeper": A case study in the selection of news. *Journalism quarterly*, 27(4):383–390.

A Processing data details

A.1 Aligning the Internet Archives data to Vanderbilt TV News Archive

A.2 Final segment topics

Topic	N in training	Description	Sample titles
Abortion	390	Abortion	"abortion/arizona/the clinic patients"
LGBT	520	LGBTrights, gender identity	"gender transformation/bono interview (part i)"
Gender	628	Gender equality, sexual harrassement	"cosby / sex abuse"
Religion	709	Catholic church, sexual abuse	"pope/cuba visit, us visit"
Environment & Energy	948	Global warming, energy resources	"city of tomorrow (masdar city)" "colorado/wildlife refuge/environmental issues"
Health policy	949	Health care, Obamacare	"health care reform / obama plan"
Race	1259	Race relations, BLM, CRT	"dean racial slurs"
Immigration	1604	Immigration	"immigration/migrant family reunions"
Education	1657	Education policy, schools, teaching	"education/milwaukee, wisconsin/hope christian school"
Sports	3896	Sports	"horse racing/kentucky derby"
Guns	3908	Gun control, mass shootings	"las vegas, nevada / shooting massacre"
Economy	4175	Economy, welfare, wages	"business: target/security breach" "biden: gas prices"
Science & Health	4431	Medicine, technology	"social media: twitter & musk"
Crime & Police	5059	Crime, court cases, police brutality	"dallas, texas/ambush of police"
Covid	5170	Covid	"coronavirus/us/the variants, vaccine"
Elections	5274	Elections and campaigns	"campaign 2016/republicans", "campaign 2012/perry/a discussion"
Human interest	10784	Lifestyle, entertainment	"economy: holiday shopping" ""
Weather & Disasters	13699	Weather, natural disasters, accidents	"winter weather/storm" "arizona/wildfire"
Politics	17029	Non-electoral politics	"trump/session/"new york times" interview"
Foreign affairs	17319	Foreign affairs, security	"afghanistan war/kabul bombing" "egypt/presidential election"

Table 6: Topic labels and sample titles

B More evaluation statistics

Category	precision	recall	specificity
abortion	0.429	0.692	0.999
covid	0.662	0.723	0.984
crime_police	0.595	0.633	0.980
economy	0.481	0.666	0.991
education	0.358	0.553	0.995
elections	0.574	0.742	0.988
environment_energy	0.258	0.674	0.999
foreign	0.709	0.667	0.927
gender	0.467	0.705	0.999
guns	0.628	0.707	0.989
health_policy	0.524	0.811	0.999
human_interest	0.732	0.527	0.918
immigration	0.429	0.648	0.996
lgbt	0.375	0.692	0.999
politics	0.637	0.533	0.885
race	0.169	0.469	0.998
religion_catholic	0.521	0.662	0.998
science_health	0.607	0.721	0.990
sports	0.554	0.748	0.993
weather_disaster	0.773	0.779	0.965

Table 7: Statistics by topic for best performing elastic net model

cat	precision	recall	specificity
1	0.553	0.773	0.952
2	0.887	0.562	0.972
3	0.836	0.494	0.991
4	0.813	0.531	0.983
5	0.617	0.479	0.994
6	0.829	0.580	0.997
7	0.333	0.080	0.994
8	0.760	0.384	0.996
9	0.885	0.333	0.997
10	0.824	0.394	0.997
11	0.545	0.164	0.996
12	0.735	0.646	0.980
13	0.850	0.633	0.982
14	0.760	0.794	0.957
15	0.904	0.702	0.979
16	0.891	0.519	0.997
17	0.842	0.607	0.983
18	0.827	0.813	0.969
19	0.544	0.801	0.975
20	0.829	0.617	0.986

Table 8: Statistics by topic for fine-tuned RoBERTa model

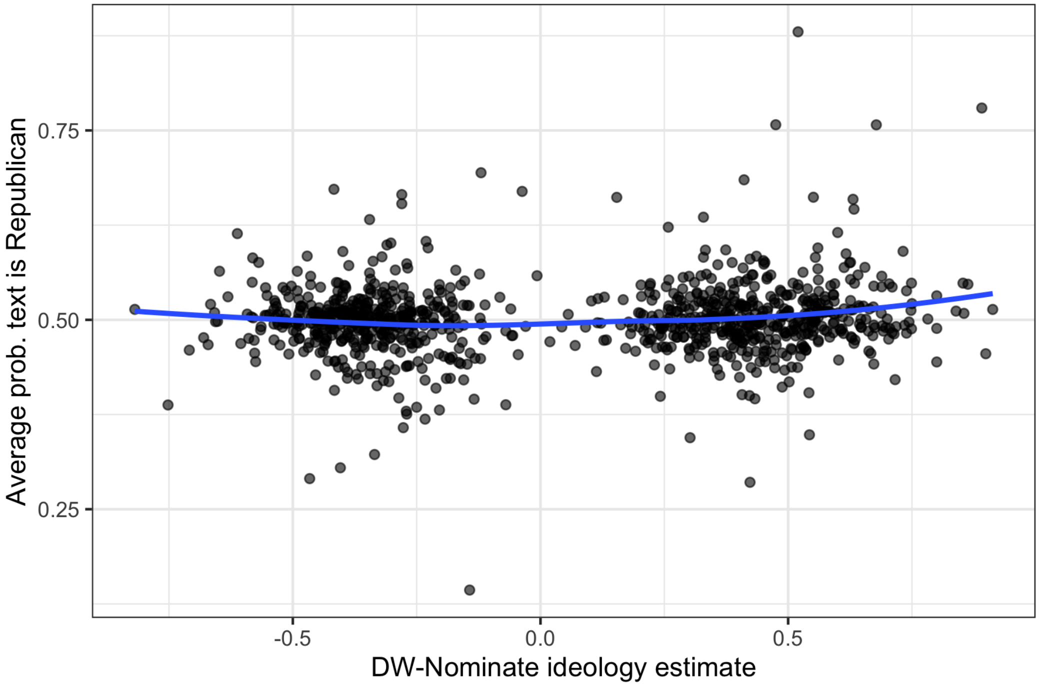

Figure 1: Scatterplot of DW-NOMINATE against ridge partisanship