Douglas Hummel-Price
PPOL_628: Text-As-Data: Computational Linguistics
May 7, 2020
Matthew Denny

Acknowledging the Gravity of the Pandemic Before the President Does:

Comparing US Department of Education Press Releases Before and After Five COVID-19-Related Milestones

Executive Summary

In order to examine the Department of Education's response to the coronavirus pandemic, I scraped, examined, and preprocessed a corpus of 376 press releases. Using five distinct milestones related to the virus as dichotomous variables, I ran five supervised learning models on the corpus, attempting to predict whether a given press release is pre- or post-milestone. If the Department noticeably altered the language or content of their press releases, at least one of these milestones should serve as a good dividing line between pre-COVID and subsequent releases. The comparative performance of the models can be used to evaluate when the Department began responding to the pandemic in its press releases and compare that the timing of various positions taken and statements made by President Trump.

Coronavirus Timeline Milestones

To examine whether and when the DeVos Administration modified their verbiage in response to the COVID-19 outbreak, I used the following dates and associated events as milestones to use as the dividing line between pre- and post-pandemic press releases.

- Milestone A: January 5th – WHO reports a "pneumonia of unknown cause" in Wuhan, China.
- Milestone B: February 2nd – Trump's executive order banning travel from China goes into effect.
- Milestone C: February 28th – WHO raises global risk of coronavirus from "high" to "very high".
- Milestone D: March 16th – Trump issues orders recommending behavioral changes, including schooling from home whenever possible.
- Milestone E: April 7th – Trump criticizes the WHO's handling of the crisis, foreshadowing funding halt on April 14th.

Note: all dates are in the year 2020 and come from an NPR article dated April 15th, 2020.[1]

[1] Tamara Keith and Malaka Gharib, "A Timeline Of Coronavirus Comments From President Trump And WHO," NPR (NPR, April 15, 2020), https://www.npr.org/sections/goatsandsoda/2020/04/15/835011346/a-timeline-of-coronavirus-comments-from-president-trump-and-who.

Milestone A: January 5th – WHO Reports a "pneumonia of unknown cause" in Wuhan, China.[2]

This date represents the earliest that anyone without privileged knowledge could have known that the virus might become a big problem. Since the administration did not meaningfully respond to the proclamation, this milestone serves as a good benchmark. Subsequent models should outperform this model.

Milestone B: February 2nd – Trump's executive order banning travel from China goes into effect.[3]

The China travel ban represents the first major action taken by the Trump administration in response to the pandemic. It serves as the earliest meaningful action by his Administration.

Milestone C: February 28th – WHO raises global risk of coronavirus from "high" to "very high".[4]

The elevation of threat level by the WHO serves as a secondary measure of how seriously the global community takes the threat. It represents the last point at which a presidential administration could still act and not be accused of ignoring the international community.

Milestone D: March 16th – Trump issues orders recommending behavioral changes, including schooling from home whenever possible.[5]

This action by the president represents an escalation in Federal coronavirus response. It directly refers to schooling, so it represents the point at which we should expect some language changes in the press briefings, if they coincide with presidential action.

Milestone E: April 7th – Trump criticizes the WHO's handling of the crisis, foreshadowing funding halt on April 14th.[6]

Since any analysis is artificially capped by the date it is performed and relies on at least some instances of both pre- and post-milestone press releases, this date represents the latest date that could reasonably be thought to provide any level of accuracy.

---

[2] "Pneumonia of Unknown Cause – China," World Health Organization (World Health Organization, January 30, 2020), https://www.who.int/csr/don/05-january-2020-pneumonia-of-unkown-cause-china/en/.

[3] "Proclamation on Suspension of Entry as Immigrants and Nonimmigrants of Persons Who Pose a Risk of Transmitting 2019 Novel Coronavirus," The White House (The United States Government, January 31, 2020), https://www.whitehouse.gov/presidential-actions/proclamation-suspension-entry-immigrants-nonimmigrants-persons-pose-risk-transmitting-2019-novel-coronavirus/.

[4] Berkeley Lovelace, "WHO Raises Coronavirus Threat Assessment to Its Highest Level: 'Wake up. Get Ready. This Virus May Be on Its Way'," CNBC (CNBC, February 28, 2020), https://www.cnbc.com/2020/02/28/who-raises-risk-assessment-of-coronavirus-to-very-high-at-global-level.html.

[5] "Remarks by President Trump, Vice President Pence, and Members of the Coronavirus Task Force in Press Briefing," The White House (The United States Government, March 12, 2020), https://www.whitehouse.gov/briefings-statements/remarks-president-trump-vice-president-pence-members-coronavirus-task-force-press-briefing-3/.

[6] Donald J. Trump, "Tweet," Twitter (Twitter, April 7, 2020), https://twitter.com/realDonaldTrump/status/1247540701291638787?ref_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%5E1247540701291638787.

Wrangling the Corpus

My corpus consists of 376 press releases published by the Department of Education from the first day of the Trump Administration, January 21st, 2017, through May 6, 2020. The end date is dictated by the date of the analysis rather than an external, substantive reason.

The press releases each live on their own URL on the Department of Education website, so I wrote and executed a two-stage Python script in a Jupyter Notebook on a Windows machine to collect them.[7] The press releases are indexed on a series of thirty-seven pages that contain ten releases each, so the first stage iterates over the thirty-seven pages, snagging the URLs for the ten releases on each page to create a master list of URLs. The second stage then iterates over the master list, appending the following for each press release:

- Release/Publication Date
- Whether the Release Contains an Image
- Release Title/Headline
- Release Sub-title/Sub-headline
- Release Text

Before substantive preprocessing could be applied to the corpus, a variety of technical preprocessing steps that are not traditionally considered "text preprocessing" were applied to make the text usable in the appropriate human-and-machine-readable format. The structure of the underlying webpage code resulted in the text fields containing a variety of residual html elements, so these were searched and stripped. The date field was transformed to represent the number of days since the Trump inauguration. The resulting field was then used to construct dummy variables for each of the five milestones to be used in a supervised dichotomous prediction problem. Some residual JSON code within the entries was manually removed; this only appears on the most recent entries, so it may have something to do with the Department enacting compliance with schema and open-data reporting requirements as part of the Foundation for Evidence-Based Policymaking Act signed in 2019.

Preprocessing Decisions

Preprocessing decisions can have profound effects on subsequent analysis, especially within the context of a supervised learning problem. Since my corpus is relatively small and most documents fairly short, I was not concerned about the overall size of the term frequency-inverse document frequency that would result from my corpus. In general, I made decisions on the basis of giving preference to preserving meaning at the cost of larger vocabulary size and attempted to prophylactically deal with any context-specific issues that may arise with a corpus of documents that are all directly from a government press

[7] "Press Releases," U.S. Department of Education, accessed May 6, 2020, https://www.ed.gov/news/press-releases.

office and about education. Using the framework discussed by Dr. Matthew Denny, my decisions are as follows:[8]

- P – Punctuation: There seems no particularly good reason to retain most punctuation. The two exceptions to that are quotation marks and dollar signs. Some of the documents have multiple quotations while others do not, so I wanted to retain any informative value that the frequency of quotations might have. A number of the documents contain budgetary numbers that begin with dollar signs. Before stripping any dollar signs, I first applied the dollar_amount substitution described below.
- N – Numbers: The vast majority of the numbers in the corpus are either dollar amounts or years. To retain some information from the year numbers and dollar amounts, the following substitutions were made:

Number Description	Replaced With
Years 1700 – 2009	"older_year_number"
Years 2010 – 2020	"recent_year_number"
Years 2021 – 2059	"future_year_number"
${integer/float} million	"dollar_amount"
${integer/float} billion	"dollar_amount"
${integer/float} trillion	"dollar_amount"
${integer with 1-12 digits}	"dollar_amount"

  The remaining numbers were dropped from the corpus.
- L – Lowercasing: The corpus contains a lot of proper nouns, e.g., “Department of Education”. Simply lowercasing would run the risk of losing information about capitalized entities that are important, but I did want to lowercase the corpus. I compensated for this by including n-grams.
- S – Stemming: Since all documents come from the same sanitized source, I generally expect the verb tenses to be consistent within my corpus. Beyond that, the reduction in vocabulary that would result from stemming is not worth the increased risk of misleading results or the increased difficult of human-interpretation of the final TF-IDF matrix. I did not stem my corpus.
- W – Stopword Removal: To remove terms that hold syntactic but not substantive value, I removed stopwords. I used the NLTK English stopword list in Python.
- 3 – n-gram inclusion: Since any government agency is likely to reference other agencies, which have multi-word names, n-gram inclusion is important. I incorporated n-grams of one to four words.
- I – Infrequently Used Terms: Since my corpus is relatively small, and I used a TF-IDF matrix, I set the document frequency floor to 3 and the term frequency floor 3. Relatedly, I set the document frequency ceiling to 363, which is the total number of documents, 376, minus the smallest number of documents that are after one of our milestones, 16, and then added 3 in case certain terms only start appearing in those post-milestone releases. I did not set a term

[8] Matthew J. Denny and Arthur Spirling, “Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It,” Political Analysis 26, no. 2 (2018): pp. 168-189, https://doi.org/10.1017/pan.2017.44.

frequency ceiling; since I removed stopwords, any remaining words that occur frequently hold information, even if that information does not help fitting a specific model.

- Other: Since President Trump, Secretary of Education Betsy DeVos, the United States, and the Department of Education occur frequently throughout the corpus, I have standardized these terms. Any likely combination of words referring to one of these concepts has been replaced by a single word for the concept, e.g., “US”, “U.S.”, and “United States” were replaced with “united_states”.

Preprocessing Results

After applying the above preprocessing steps except for the infrequently used terms preprocessing, I produced a document-term matrix that had 255,368 features, was 99.6% sparse, and was a 26.8Mb object in R. After trimming the dtm using the document frequency floor and ceiling and the term frequency floor, the dtm reduced down to 13,030 features, a sparsity of 97.9%, and a 2.4 Mb object in R. The reduction seemed to significantly reduce the size of the dtm while still retaining over 13,000 features to work with.

Hypothesis

While the type of modeling I performed does not directly lend itself to formal hypothesis testing, it is still worthwhile to speculate on the results of the models prior to running them. I hypothesized that models C and D will present the best AUC, based on the idea that the Trump administration was slow to recognize the gravity of the crisis and the fact that the more recent milestones will contain small sample sizes. The accuracy levels of all the models should be high due to the imbalance between the pre- and post-milestone sample sizes. I suspected that the models would identify very specific words that determine a post-milestone post, and the absence of those few words will result in a pre-milestone categorization.

Modeling

The goal of my analysis is to determine which of the five milestones provides the best dividing line as measured by the area under the receiver operating characteristic curve. I used the lasso logistic regression model from the glmnet package and the binary logistic model from XGBoost in R. I used an 80-20 training/test split using the createDataPartition function to ensure that my test and training sets are relatively balanced with regards to the dependent variable. Because of this, I used a distinct training-test split for each of the five models.

For the lasso models, I examined the accuracy at optimal cutoff, the AUC, and plots of the optimal lasso penalty graph and ROC curve. For the XGBoost models, I examined the accuracy at optimal cutoff, the AUC, and plots of the ROC curve and the top features by importance to the model.

Upon running the first few XGBoost models, I noticed that a few obvious terms dominated the feature importance: “COVID”, “Coronavirus”, and their variants. I decided then to run a second set of models, models A2-E2, in which these terms were removed. I noticed that there were still a few terms that seemed more explicitly COVID related, in particular references to the “national emergency” and boilerplate language recommending that the reader consult the CDC’s website for more information. I therefore decided to then run a third set of models, models A3-E3, in which these additional items were removed.

Results

All result tables and plots are found in the appendix. The highest AUC of all the models was XGBoost Model B, followed closely by XGBoost Model B2. Given that explicit COVID-19 references dominated the feature importances of the XGBoost models, it is neither surprising nor enlightening that the highest performing model contained these explicit references. Still, these results suggest that the Department of Education began making statements related to the pandemic earlier than Trump’s own statements expressed understanding of the gravity of the situation. The second and third iterations of a given milestone model generally performed worse than the first and second iterations, respectively. This is to be expected, given that the later models did not have the benefit of the obvious classifier tokens. I was surprised that when looking at the XGBoost models across milestones, the later milestones performed worse. This may be largely due to the smaller sample sizes of those target groups.

The highest AUC of the lasso models came from models A3 and B3 – in fact, B2 was the lowest performing model of all the lassos, and the second lowest overall. Oddly, many of the second and third iterations of a model performed *better* than the first and second iterations, respectively. The lasso model included cross-fold validation that may resulted in overall lower performance across the lasso models, but a higher performance in the later runs. It’s worth noting that the most recent milestone, E, did not even return AUCs for two out of the three models, suggesting that they simply never categorized any documents as post-release.

Conclusion

Ultimately, these results do not allow for particularly strong conclusions, but they do serve as a jumping off point for deeper analysis. Since these models are quite sensitive to preprocessing, perhaps more time dedicated to tuning to these steps – particularly stop words and infrequent terms – might lead to more robust results. In general, the accuracies of these models are quite high, but this is misleading. The sample imbalance likely dictates here; with an order of magnitude more pre-milestone press releases than post-milestone, a classifier that simply classifies all press releases as pre-milestone will generally perform in the low 90s. It appears that the Department of Education *may* have indicated that they took the pandemic seriously before the president did, but the analyses here can only serve as starting point for deeper consideration.

Appendix: Tables and Plots

Table 1: XGBoost Model Results

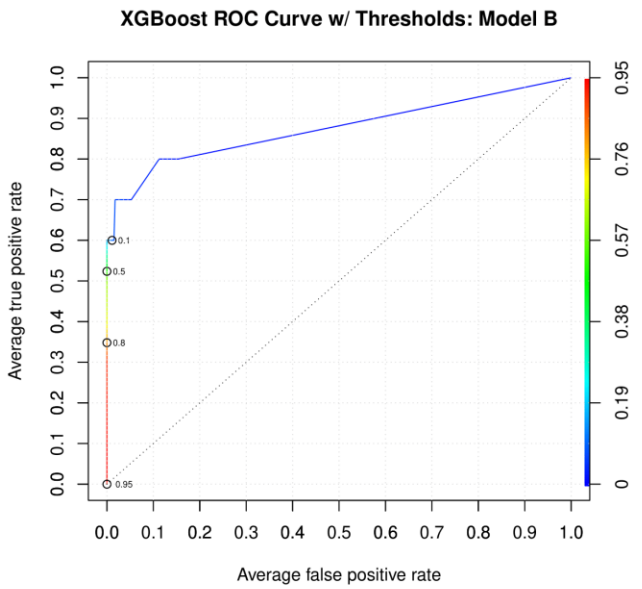
Model	AUC	Accuracy	Optimal Cutoff
XGBoost Model A	0.860	0.933	0.218
XGBoost Model A2	0.757	0.933	0.311
XGBoost Model A3	0.708	0.920	0.317
XGBoost Model B	0.882	0.947	0.165
XGBoost Model B2	0.862	0.960	0.076
XGBoost Model B3	0.768	0.947	0.082
XGBoost Model C	0.813	0.947	0.156
XGBoost Model C2	0.785	0.933	0.087
XGBoost Model C3	0.756	0.920	0.417
XGBoost Model D	0.799	0.947	0.037
XGBoost Model D2	0.798	0.933	0.258
XGBoost Model D3	0.801	0.933	0.201
XGBoost Model E	0.690	0.947	0.181
XGBoost Model E2	0.693	0.947	0.261
XGBoost Model E3	0.667	0.947	0.080

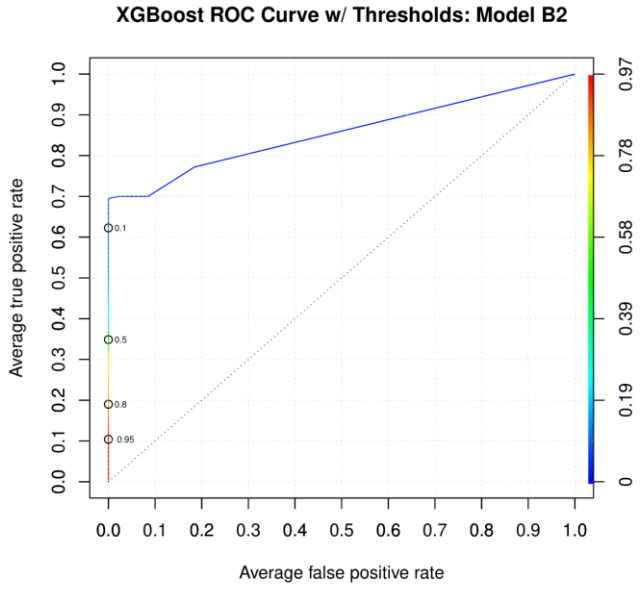
Table 2: Lasso Model Results

Model	AUC	Accuracy	Optimal Cutoff
Lasso Model A	0.723	0.920	0.236
Lasso Model A2	0.780	0.920	0.768
Lasso Model A3	0.808	0.933	0.530
Lasso Model B	0.742	0.933	0.069
Lasso Model B2	0.671	0.907	0.454
Lasso Model B3	0.808	0.907	0.583
Lasso Model C	0.764	0.920	0.569
Lasso Model C2	0.747	0.920	0.387
Lasso Model C3	0.690	0.920	0.698
Lasso Model D	0.754	0.920	0.662
Lasso Model D2	0.753	0.933	0.108
Lasso Model D3	0.705	0.920	0.800
Lasso Model E	N/A*	0.920	infinite
Lasso Model E2	N/A*	0.920	infinite
Lasso Model E3	0.655	0.933	0.447

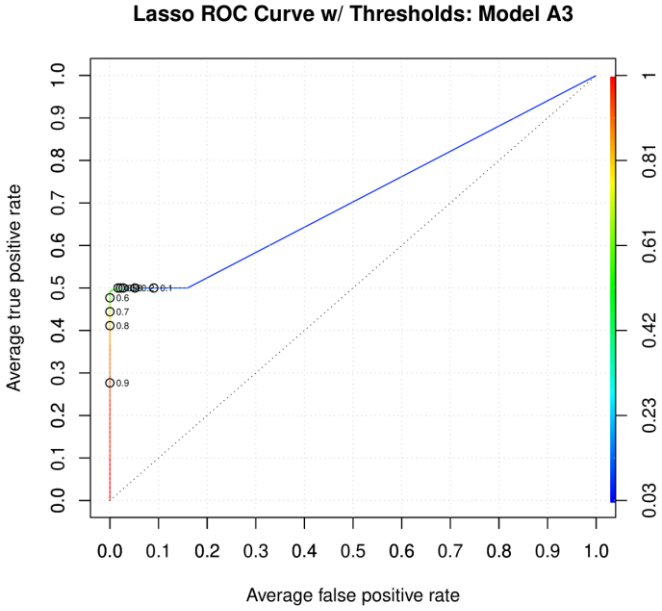
Best Performer | Worst Performer

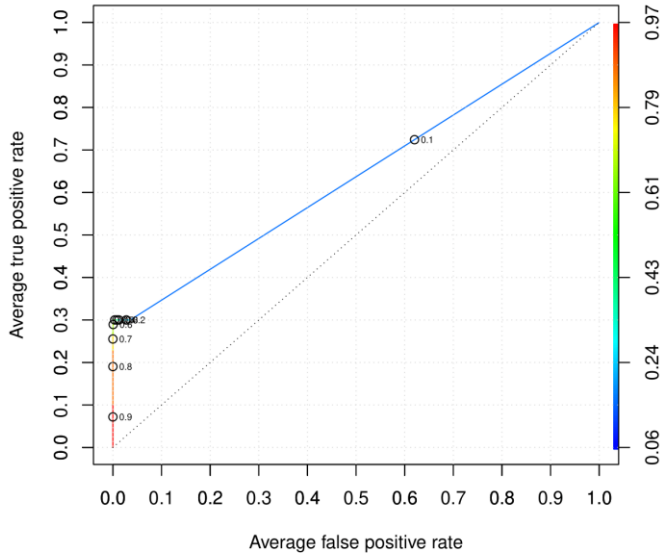
Plots 1-2: XGBoost Model Result ROC Curves for Highest Performing Models



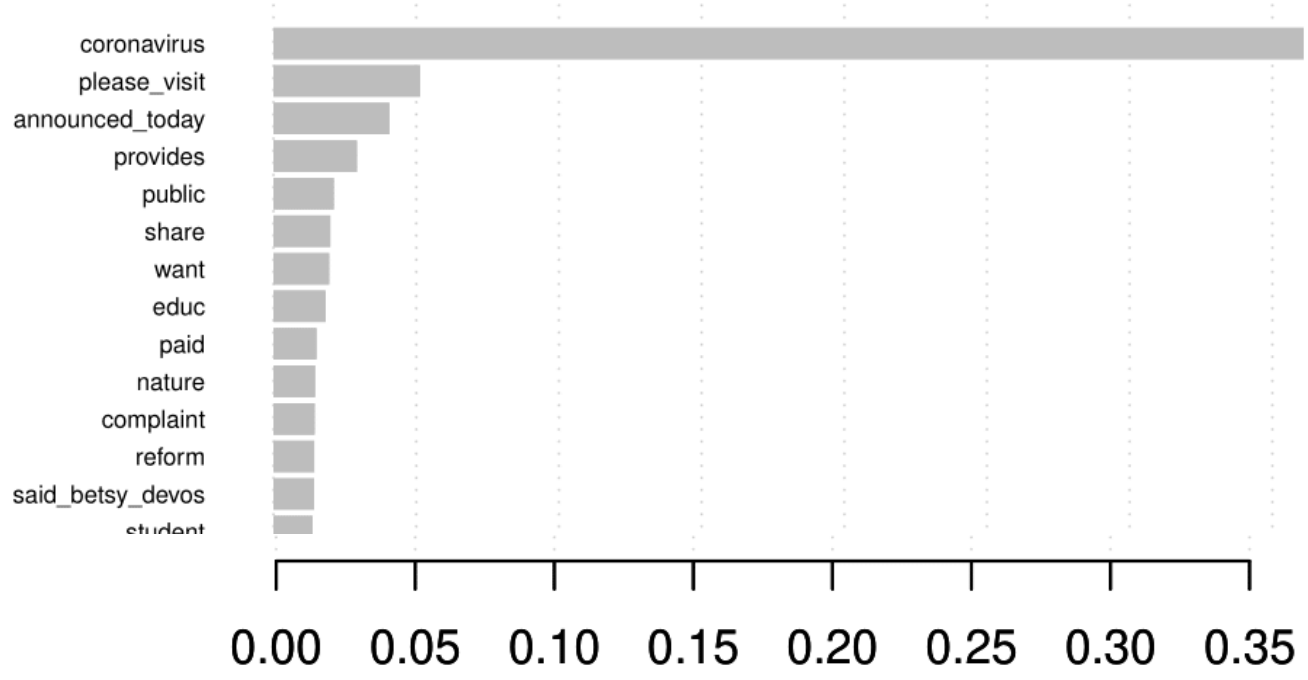


Plots 3-4: Lasso Model Result ROC Curves for Highest Performing Models

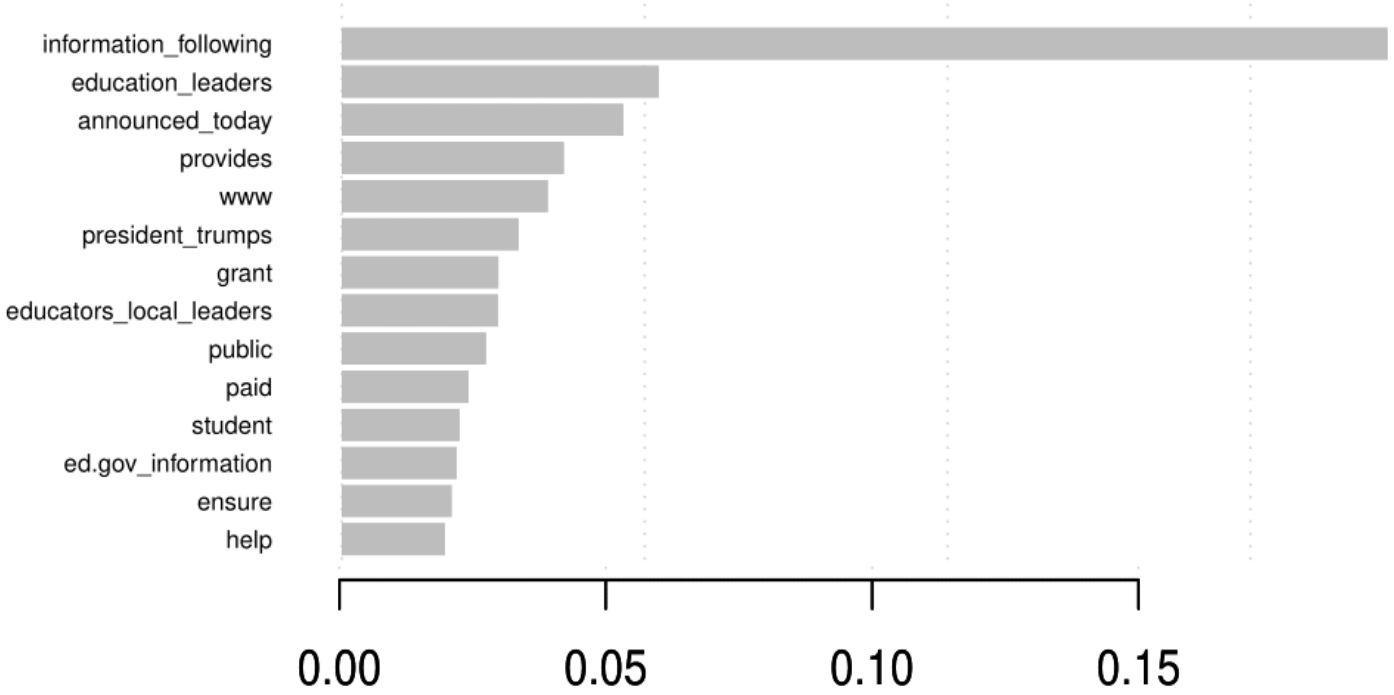




Plot 5: Binary Feature Importances for XGBoost Model A



Plot 6: Binary Feature Importances for XGBoost Model A3



Bibliography

Denny, Matthew J., and Arthur Spirling. “Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It.” Political Analysis 26, no. 2 (2018): 168–89. https://doi.org/10.1017/pan.2017.44.

Keith, Tamara, and Malaka Gharib. “A Timeline Of Coronavirus Comments From President Trump And WHO.” NPR. NPR, April 15, 2020. https://www.npr.org/sections/goatsandsoda/2020/04/15/835011346/a-timeline-of-coronavirus-comments-from-president-trump-and-who.

Lovelace, Berkeley. “WHO Raises Coronavirus Threat Assessment to Its Highest Level: 'Wake up. Get Ready. This Virus May Be on Its Way'.” CNBC. CNBC, February 28, 2020. https://www.cnbc.com/2020/02/28/who-raises-risk-assessment-of-coronavirus-to-very-high-at-global-level.html.

“Pneumonia of Unknown Cause – China.” World Health Organization. World Health Organization, January 30, 2020. https://www.who.int/csr/don/05-january-2020-pneumonia-of-unkown-cause-china/en/.

“Press Releases.” U.S. Department of Education. Accessed May 6, 2020. https://www.ed.gov/news/press-releases.

“Proclamation on Suspension of Entry as Immigrants and Nonimmigrants of Persons Who Pose a Risk of Transmitting 2019 Novel Coronavirus.” The White House. The United States Government, January 31, 2020. https://www.whitehouse.gov/presidential-actions/proclamation-suspension-entry-immigrants-nonimmigrants-persons-pose-risk-transmitting-2019-novel-coronavirus/.

“Remarks by President Trump, Vice President Pence, and Members of the Coronavirus Task Force in Press Briefing.” The White House. The United States Government, March 12, 2020. https://www.whitehouse.gov/briefings-statements/remarks-president-trump-vice-president-pence-members-coronavirus-task-force-press-briefing-3/.

Trump, Donald J. “Tweet.” Twitter. Twitter, April 7, 2020. https://twitter.com/realDonaldTrump/status/1247540701291638787?ref_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%5E1247540701291638787.