Alexander Tyan

MACSS 35000

February 8, 2019

Assignment 4: Preliminary Data Analysis

**Introduction:**

How do online news article features explain and predict the bias of the article content? In this thesis, I intend to argue that news article topics (along with the source of publication) are associated with the extent to which news sources converge or diverge from each other in news content. I define this distance as *bias.* Mathematically, this *bias* is a cosine distance between vector forms of text contents­ of news articles and their corresponding sources (more on computing variable measures in the Data Engineering and Analysis Section).

I hypothesize that there is a statistically significant difference in the bias measure between different news sources’ articles, conditional on topics. That is, one would expect that for topic A the average cosine distance between texts of sources X and Y should be different than for topic B. The intuition is that, for example, CNN’s articles may be quite different from RT’s articles on topics such as “politics”, but not as different on less politicized topics like “crime.”

In terms of analysis methods, for these preliminary results, I use Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003), a probabilistic topic modelling technique to infer the most dominant topic for each observation (i.e. article). I use this method as outlined in Grimmer and Stewart (2013, 17–19). Prior Political Science research used topic modelling to infer predominant topics in Senate floor speeches (Quinn et al. 2010) and article topics in US news covering women in predominantly Muslim countries (Terman 2017). Hence, this is an appropriate method to engineer the topic variable for the data, the “predominant topic” for each news article observation.

To construct the independent variable, cosine distances, I vectorize texts using word2vec word embeddings, which represent each text string in its vector form as extracted from a vector space that represents latent semantic content of news articles (Mikolov et al. 2013). There are two reasons for using these pre-trained word embeddings. First, to train our own word embeddings, one would need a very large corpus (of news, in this case) (Chollet 2017). Since the article dataset for the purposes of this analysis has only 3327 observations, I resort to pretrained word embeddings. Second, the embeddings I use were trained on Google News. This is appropriate as the vector space of word embeddings is tied to the language context (Chollet 2017). Since the embeddings were trained on 100 billions words from Google News, I expect that these embeddings would be appropriate to this project’s context (Google 2013).

For the purposes of the preliminary analysis, I use Welch’s T-test for difference in means in cosine distances between pairs of news sources, conditional on topics. However, for the full version of the analysis, I anticipate experimenting with multiple linear regression and supervised machine learning methods to infer a function that maps article sources and topics onto values of cosine distances (more on that in the Further Steps and Limitations).

**Data Engineering and Analysis**

The dataset was gathered through a scraping process with RSS feeds and using Python’s newspaper3k package, as outlined in the Methods paper.

|  |  |
| --- | --- |
| Source | Count |
| BBC News | 797 |
| CNN | 939 |
| Fox News | 760 |
| RT | 831 |
| Total | 3327 |

Table 1: Source Article Counts

Table 1 summarizes total counts of articles per each source. Note that MSNBC source is missing. This is because the project could not discover a reliable RSS feed service for the website and data has not been gathered for this source using RSS feeds.[[1]](#footnote-1) For this analysis, the range of scraping dates is between January 31 and February 7, 2019.

After cleaning and word-preprocessing steps outlined in the methods paper, the next step in analysis consists of generating predominant topics for each of the articles, which necessitates building a topic model. Since this is an unsupervised algorithm, one does not know a priori the optimal number of topics that the texts would cluster round. One way to decide an appropriate number is the coherence measure; higher coherence measures demonstrate a better topic number (Röder, Both, and Hinneburg 2015). Figure 1 is a result of training 38 different LDA topic models, each associated with a number of topics ranging from 2 to 40.

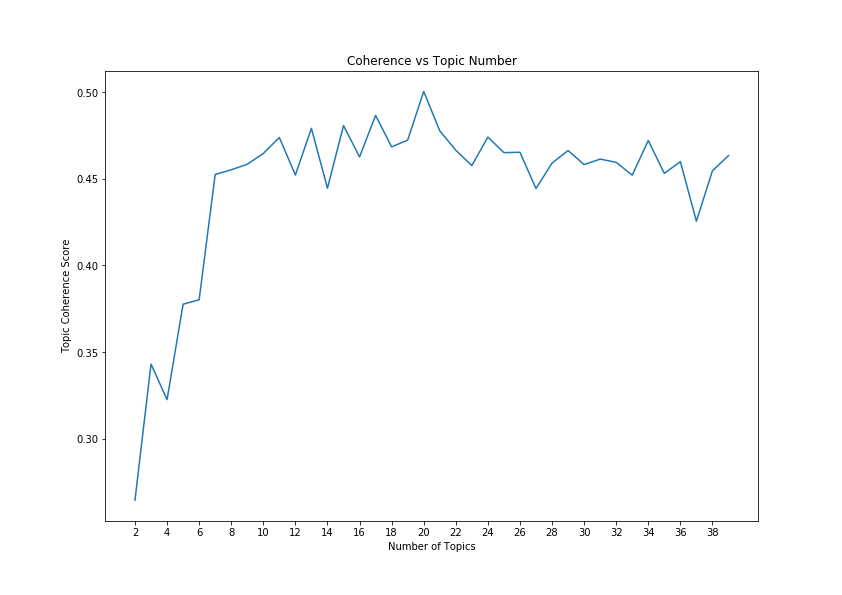


Figure 1: Coherence vs Number of Topics

Note that the coherence measure reaches its maximum and levels off at 20 topics. Thus, I choose the 20-topics LDA model for further analysis.

LDA assumes that each topic is composed words (lemmas), with each word responsible for a proportion of a topic. Appendix contains a table of the 10 of the 20 topics[[2]](#footnote-2), with the listing of each top lemmas associated with a given topic and the corresponding weights within the topic. Topic labels in the first column are researcher’s interpretations based on the lemmas within each topic and partial manual reading of news articles in which the topic is dominant. Note that some topics are more human-interpretable than others, topic labels that are less interpretable are appended with the question mark.

Once I assign a predominant to each article, we can see the distribution of topic for each source (Figure 2).

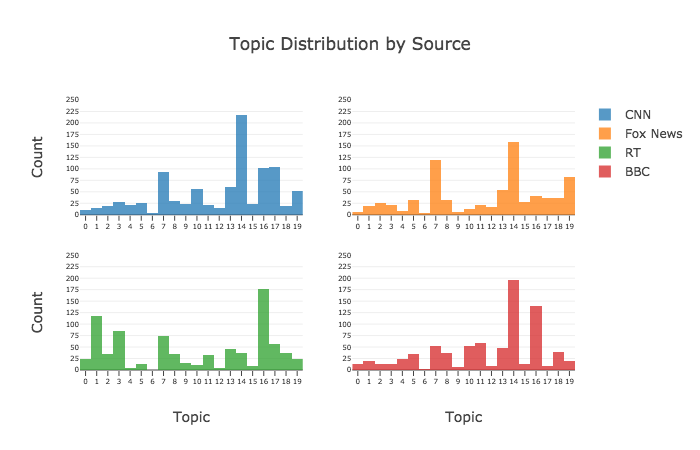


Figure 2: Topic Count Distribution per Source

Next variable for analysis is the cosine difference, the dependent variable. Cosine difference is . For computation purposes, I u­se cosine similarity, but the statistical results do not differ as statistical tests for difference in cosine distances (i.e. bias) are mathematically equivalent to tests in cosine similarities. The interpretation is that two “similar” news articles would have “little” bias, as defined by this study. The bias measure (cosine difference) between two news articles in their vector representations and is:

Where vector norms used in the denominator are L2-norms. Since there are 3327 article observations, I compute pairwise article cosine similarities. For the purposes of this preliminary analysis, let us look at CNN as the baseline standard for bias measure and Fox News and RT as compared to CNN and constrain the search to topics “Crime” and “US Politics.”

Table 2 reports summary statistics on cosine similarities with the data constrained to the conditions in the previous paragraphs.

Table 2: CNN to Target Source Cosine Similarity Summary Statistics

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| topic | target source | count | mean | std | min | 25% | 50% | 75% | max |
| Crime | Fox News | 11160 | 0.829556933 | 0.05382656 | 0.546680808 | 0.799793571 | 0.835678518 | 0.86743091 | 0.983913898 |
| RT | 6789 | 0.839204574 | 0.048338039 | 0.622709155 | 0.811861634 | 0.844044447 | 0.872817934 | 0.970934093 |
| US Politics | Fox News | 3848 | 0.883838486 | 0.053207771 | 0.616143465 | 0.859033167 | 0.89343968 | 0.919283092 | 0.986397505 |
| RT | 11856 | 0.880364607 | 0.048124528 | 0.603044808 | 0.857095465 | 0.889096975 | 0.914100915 | 0.977500319 |

For instance, on the topic of “Crime,” there are 11160 cosine similarities between CNN and Fox News news articles, with an average similarity value of 0.82956 and standard deviation of 0.05384, with values ranging from 0.54668 and 0.98391. So, with constraints, there are a total of 33653 cosine similarity observations.

For this analysis, I would hypothesize that, on average, the cosine similarity between CNN and Fox News articles to be different from between Crime and US Politics topics. One would expect that, since US Politics should be a more polarizing topic. On the other hand, when writing on Crime, the two sources should be closer on average, so the mean difference (and similarity) should be significantly lower, or different from the US Politics. Similarly, average similarities between CNN and RT should be different conditional on the topic. So I perform two unequal variance t-tests, one between the two similarity means of CNN and Fox News and one between CNN and RT. More formally, the null hypothesis is that there is no difference between the two cosine similarity means within each target source. Alternative hypothesis is that there is a statistical difference. The assumption of normal distributions of sample values is acceptable due to a large sample size and equal variances are not assumed.

Table : Welch's 2 Sample T-Tests

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean Difference | Test Statistic | p-value |
| CNN vs Fox News | -0.05428 | -54.40851 | 0.0 |
| CNN vs RT | -0.04115 | -56.03709 | 0.0 |

Based on results in Table 3, we can reject the null hypotheses of no difference. There is a statistically significant difference within each set of means, for both t-tests (at any alpha significance threshold).

In terms of interpretation of the results, there is a confirmation of the intuition outlined above. However, as a counterpoint, one thing to consider is the practical interpretation of the small (albeit statistically significant) mean difference. Would such small mean differences between the two categories for CNN vs Fox News and CNN vs RT (-.05 and -0.04 respectively) have an impact in terms of a reader being able to tell the difference in content? Would a reader be able to tell the difference in similarity by reading articles between two topics? What does a value of -0.05 mean practically for a news reader? This is unclear based on the design of the study.

**Further steps and Limitations**

The analysis performed above is quite rudimentary in terms of statistical machinery and is also limited in that it is constrained to a small subset of the data (in both sources and topic categories). Therefore, a natural extension of this analysis will be expanding the data set to all topics and all news sources. This would mean including BBC and expanding to eighteen more topics. However, for more feasible statistical analysis, one can still keep the analysis constrained to one baseline source (as CNN in the preliminary analysis). Moreover, one more variable not considered above is the time of publication, a data point that I have also been extracting during data gathering.

The expansion of the analysis may naturally necessitate different statistical methods, because the use of many isolated t-tests may become insufficient.[[3]](#footnote-3) One way to do so is to use supervised machine learning to infer a function that takes in constructed dominant topics and news sources to predict the continuous dependent variable values of cosine similarities. A successful model should be able to capture this mapping, suggesting that topics and sources may play a nontrivial role in determining news bias.

Another extension of the current methodology is to account for the “dominance” of a topic. Currently, the project only accounts for what the dominant topic is for each article, but not to what extent that article is dominated by the dominant article. As of the writing of this, it is not yet clear how to incorporate this data into the statistical analysis.

One challenge is the interpretability of topics coming out of the LDA topic model. This is especially important since the hypotheses are based on the researcher’s interpretations of the topics. When topic interpretation is vague or unclear, as is the case with the “Europe/Religion” or “Miscellaneous” topic, it is unclear what hypotheses to posit for further analysis.

A final important extension needed for the project is a methods validation step. For instance, as Denny and Spirling (2018) suggest in application to text-as-data methods in Political Science, text preprocessing steps such as lemmatization and tokenization of words may have non-trivial consequences for the later analysis. The authors also provide a software tool to measure such sensitivity mathematically. Employing this method can contribute to validating this project’s methodology.

Bibliography

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. “Latent Dirichlet Allocation.” *Journal of Machine Learning Research* 3 (Jan): 993–1022.

Chollet, Francois. 2017. *Deep Learning with Python*. 1 edition. Shelter Island, New York: Manning Publications.

Denny, Matthew J., and Arthur Spirling. 2018. “Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It.” *Political Analysis* 26 (2): 168–89. https://doi.org/10.1017/pan.2017.44.

Google. 2013. “Google Code Archive - Long-Term Storage for Google Code Project Hosting.” 2013. https://code.google.com/archive/p/word2vec/.

Grimmer, Justin, and Brandon M. Stewart. 2013. “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts.” *Political Analysis* 21 (3): 267–97. https://doi.org/10.1093/pan/mps028.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. “Efficient Estimation of Word Representations in Vector Space.” *arXiv:1301.3781 [Cs]*, January. http://arxiv.org/abs/1301.3781.

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. “How to Analyze Political Attention with Minimal Assumptions and Costs.” *American Journal of Political Science* 54 (1): 209–28. https://doi.org/10.1111/j.1540-5907.2009.00427.x.

Röder, Michael, Andreas Both, and Alexander Hinneburg. 2015. “Exploring the Space of Topic Coherence Measures.” In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408. ACM.

Terman, Rochelle. 2017. “Islamophobia and Media Portrayals of Muslim Women: A Computational Text Analysis of US News Coverage.” *International Studies Quarterly* 61 (3): 489–502. https://doi.org/10.1093/isq/sqx051.

Appendix

|  |  |  |
| --- | --- | --- |
| Topic | Lemma | Weight |
| Healthcare | health | 0.021356275 |
| drug | 0.017169613 |
| patient | 0.014966376 |
| hospital | 0.014058208 |
| doctor | 0.013041544 |
| medical | 0.012904156 |
| death | 0.011811129 |
| child | 0.011142593 |
| die | 0.009480785 |
| accord | 0.009249739 |
| case | 0.007865488 |
| marijuana | 0.007396302 |
| state | 0.006826073 |
| disease | 0.006820308 |
| care | 0.006795365 |
| virus | 0.006617869 |
| treatment | 0.006525188 |
| physician | 0.005527902 |
| report | 0.005455626 |
| colorado | 0.005275759 |
| Virginia Scandal | woman | 0.03458775 |
| white | 0.01401055 |
| northam | 0.013049183 |
| racist | 0.012259012 |
| blackface | 0.012191094 |
| state | 0.011485968 |
| abortion | 0.010237206 |
| statement | 0.009424654 |
| page | 0.008444632 |
| black | 0.008390697 |
| law | 0.008121205 |
| racism | 0.007457194 |
| female | 0.007386289 |
| photo | 0.007292737 |
| resign | 0.007125538 |
| allegation | 0.007048612 |
| call | 0.006940115 |
| man | 0.006812302 |
| male | 0.006213304 |
| fox\_new | 0.00565377 |
| Internet | company | 0.032780383 |
| facebook | 0.02512524 |
| user | 0.02503532 |
| datum | 0.024067476 |
| use | 0.019326303 |
| app | 0.012956737 |
| apple | 0.008173591 |
| location | 0.00787725 |
| bezos | 0.007579328 |
| device | 0.007099449 |
| account | 0.0068419 |
| information | 0.006777159 |
| content | 0.006568655 |
| online | 0.006307459 |
| ad | 0.006261959 |
| amazon | 0.006236916 |
| platform | 0.006192211 |
| technology | 0.006115128 |
| access | 0.006042211 |
| service | 0.006024787 |
| Film Entertainment | film | 0.029596165 |
| star | 0.014588127 |
| show | 0.014189566 |
| actor | 0.007148785 |
| award | 0.0070799 |
| movie | 0.006924432 |
| viewer | 0.006377359 |
| character | 0.006114045 |
| director | 0.006041962 |
| gold | 0.006004583 |
| screen | 0.005792552 |
| singer | 0.0055738 |
| art | 0.005544589 |
| series | 0.005529801 |
| jackson | 0.005524482 |
| tv | 0.005436325 |
| feature | 0.005275256 |
| documentary | 0.005188048 |
| oscar | 0.005068018 |
| audience | 0.005006113 |
| Europe/Religion (?) | percent | 0.022427117 |
| france | 0.012742646 |
| migrant | 0.012226461 |
| agent | 0.011093353 |
| italian | 0.010313642 |
| indian | 0.009704245 |
| christian | 0.008950181 |
| brown | 0.008346878 |
| church | 0.007707134 |
| client | 0.006923214 |
| muslim | 0.00682775 |
| pope | 0.006379222 |
| religious | 0.006205661 |
| local | 0.006123465 |
| india | 0.006082493 |
| french | 0.005540685 |
| melenchon | 0.00534454 |
| rap | 0.005307347 |
| rally | 0.004812846 |
| yellow\_v | 0.004779587 |
| Social Media | post | 0.018759077 |
| show | 0.011776809 |
| write | 0.011317118 |
| twitter | 0.011096339 |
| tweet | 0.010678786 |
| photo | 0.010215716 |
| social\_media | 0.009319095 |
| picture | 0.008589178 |
| claim | 0.008389646 |
| appear | 0.007832815 |
| black | 0.00766507 |
| comment | 0.007629905 |
| book | 0.007580692 |
| story | 0.007560819 |
| news | 0.006640774 |
| call | 0.00648081 |
| medium | 0.005961065 |
| share | 0.005669491 |
| message | 0.005569206 |
| white | 0.005273922 |
| Crime | say | 0.03543368 |
| police | 0.017489148 |
| report | 0.01704276 |
| investigation | 0.010634722 |
| tell | 0.009418309 |
| case | 0.008900867 |
| man | 0.008822541 |
| court | 0.008685392 |
| charge | 0.008662362 |
| accord | 0.008612257 |
| officer | 0.00778858 |
| statement | 0.007241639 |
| incident | 0.007031478 |
| arrest | 0.006884985 |
| release | 0.006214471 |
| claim | 0.006078404 |
| kill | 0.005821042 |
| authority | 0.005523318 |
| official | 0.005438603 |
| department | 0.005374842 |
| Military Technology | new | 0.015843192 |
| use | 0.007960635 |
| system | 0.007286145 |
| could | 0.006790177 |
| build | 0.005663232 |
| large | 0.005628427 |
| make | 0.005056441 |
| say | 0.004959554 |
| need | 0.004699855 |
| test | 0.004691333 |
| missile | 0.004619563 |
| time | 0.004604614 |
| provide | 0.004378165 |
| work | 0.004273396 |
| technology | 0.004106806 |
| project | 0.003706465 |
| create | 0.003641293 |
| part | 0.003542974 |
| small | 0.003276763 |
| year | 0.00319213 |
| Miscellaneous (?) | say | 0.042327397 |
| not | 0.023021588 |
| get | 0.016169745 |
| go | 0.016026348 |
| people | 0.013929803 |
| do | 0.013088849 |
| time | 0.010979071 |
| know | 0.010479162 |
| be | 0.01028548 |
| tell | 0.010213814 |
| s | 0.010200537 |
| make | 0.00943996 |
| think | 0.009119026 |
| take | 0.009084662 |
| want | 0.008767445 |
| see | 0.008444416 |
| would | 0.008057352 |
| work | 0.007879037 |
| come | 0.006913245 |
| thing | 0.006682945 |
| Government | say | 0.040766444 |
| would | 0.018493077 |
| could | 0.008005157 |
| make | 0.007651093 |
| deal | 0.007466237 |
| take | 0.006904717 |
| plan | 0.00637016 |
| add | 0.006093674 |
| government | 0.005916797 |
| year | 0.005739184 |
| come | 0.005267058 |
| issue | 0.005138557 |
| time | 0.00488225 |
| decision | 0.004870927 |
| change | 0.004532074 |
| thursday | 0.004511569 |
| however | 0.004392966 |
| support | 0.004349074 |
| tell | 0.004121621 |
| leave | 0.003997585 |

1. However, the project is collecting the data from the website using non-RSS scraping, though this may render this data not comparable to data from other sources for which RSS feeds were used. [↑](#footnote-ref-1)
2. There is a technical issue printing out all 20 topics’ probabilities that is yet to be resolved. [↑](#footnote-ref-2)
3. Though I would need further consultation on what other more sophisticated statistical techniques may be needed. [↑](#footnote-ref-3)