Curated News Whitepaper

This White Paper serves as a **Codebook** for our Curated News Dataset. It also serves to provide explanation and transparency.

Contents

Abstract 3

Introduction 4-5

Methodology 5-7

Listing 1. - Sources and Selections by Theme 6

Operationalization 7-8

Collection 8

Citation & Terms of Use 8-9

Abstract

This White Paper outlines the Curated News Dataset. It provides background, operationalization, and construction methodology. It contains information about the purpose and design of the Curated News service along with expressed written consent and terms of use should academics, journalists, and miscellaneous researchers like to explore and use our data for their own purposes and projects.

Introduction

Data can be hard to come by for researchers looking to evaluate an empirical world. Our world of big data has only served to increase our reliance on information. After looking at a range of news related data, we decided it was high time someone made a particularly transparent, public, and free-to-use dataset to improve our relationship to news. We believe problems of misinformation will only get worse as complex systems continue to complexify. Data needs lots of eyes on it in order for it to be used to achieve important outcomes for societies across the world.

Our dataset was originally crafted to study textual data. We wanted to determine if big effects come in small packages. We decided it would be prudent to start with news headlines. We plan to expand to pictures with an emphasis on symbology in the future.

Logically, news headlines are the most widely consumed information in the world. We started with this exploratory research question in mind: how effective are headlines at crafting narratives? Can they be used to determine public sentiment in the absence of polling? Interestingly, this is an elite argument because news is highly stratified in many societies. And yet, many societies with internet connections find it is difficult for elites to control informational states and people have short attention spans. Someone may read a headline or look at a picture but skip reading the news, favoring their imagined version of the truth, instead of what lays beyond the infinite scroll. We find much of human behavior is a time-effort-cost-benefit relationship.

We decided the best way to construct this dataset would be to rely on vetted statistical principles. We gathered a list of news sources and separated them into a plethora of fields, ranging professional journalism to corporate and university academia. Topics range from Technology to U.S. News to International News.

Our dataset is very much a researcher’s paradise and the insights are limitless because we are still collecting data in real-time to keep updating this dataset with new versions in waves every year. This Codebook is bare minimum, designed for researchers who just want the meat and potatoes of what we did and how we did it. As such, we will organize this codebook into three sections: methodological construction, variable operationalization, and method of collection. At the end, we will have a section not part of the Codebook proper that is dedicated to citation and terms of use.

Methodology

We collected a range of news sources, organized them by theme, and created a string vector in R. We then randomly sampled that vector without replacement. We decided not to cluster those themes when we randomly sampled them for two reasons that are both related to data collection limitations. First, we were relying on Really Simple Syndication (RSS) for our data collection methodology. If you scrape a page, you have to trust the person or organization scraping that page. With RSS, the entire ledger is publicly transparent by default and design so you don’t have to trust this dataset. Instead, you can trust the methodology. If you wanted to, you could download all of the data by going to an archive site. It would take a lot of work but it is still feasible. This makes our dataset verifiable - something missing in many academic datasets today. This was important to us. Second, scraping web pages can break over time. With the volume we were collecting, we needed a robust methodology to collect and ensure we could gather it all without breaks in data continuity. Once we set the data collection scheme, we were not going to change it no matter what. An experiment doesn’t work if you have to keep modifying the experiment. RSS was the best way to ensure collection consistency for us. At the time, we were just a fledgling startup without a lot of money and resources. Exclusive use of RSS certainly means this dataset is potentially biased. It is a small price to pay for transparency and consistency.

Because not every news organization uses RSS, sometimes favoring excludable information with paywalls or website tracking mechanisms, we couldn’t provide a robust methodology for sampling by clusters within an over-arching aggregate sample. Here is the complete list of news sources as originally constructed. We have highlighted the sources that were selected in our sample, organized them in the order of our dataset, put the ones that were not selected at the end, and organized them in their original themes so you can see where some news sources might be over-sampled when compared to their originally crafted categories.

Listing 1 – Sources and Selections by Theme

University-related Journalism:

MIT Technology Review, YaleNews, Stanford News, Harvard Business Review

U.S. Journalism

CNBC, Christian Science Monitor, The Hill, Huffington Post, Five Thirty Eight, The Economist, Washington Press, National Journal, National Review, Real Clear Politics, The Guardian, Wired, Gizmodo, The Washington Post, The Nation, Vox, Reason (attempted collection for an undisclosed side project but it failed), PBS News, Market Watch, NPR, Politico, The Atlantic, Tech Dirt, Geek Wire, Tech Crunch, Forbes

International Journalism

Global Slavery Index, Times of India, Channel News Asia, France24, BBC News, Al Monitor, CNN (World), Fox News (World), Al Jazeera, MSNBC (World News), European Union Newsroom (Business), IMF Blog

Government and/or Government-Related Journalism

CDC (Podcast), NOAA, Washington Times, Real Clear Investigations, DC Circuit Breaker, Library of Congress,

Academic Journalism

Cato Institute, RAND Corporation, Mises Institute, American Enterprise Institute, Center for American Progress, The Conversation, The Brookings Institution, Foreign Policy Research Institute, Hover Institute

Miscellaneous:

World Wildlife Foundation, History.com, Media Bias Fact Check, The Economist (Technology), New Atlas, Phys.org

As you can tell from the breakdown above, the distribution of sources selected based on our initial themes is pretty well distributed. Academic institutions might be overrepresented and international news might be under-represented. The academic over-sampling is likely due to the fact *we did not randomly select by our initial themes as clusters*. The international news under-representation is likely a result of our requirements for news sources that are both English and contain an adequate RSS feed.

News sources in the above list were originally chosen to limit outliers. We used Media Bias Fact Check, a third-party independent fact checker, to ground our evaluation at the bottom. If Media Bias Fact Check had an evaluation on a news source, we used their evaluation if the source was Left-Center, Right-Center, or Least Biased. If it was outside those boundaries, we did not include that news source in our vector to be sampled.

Operationalization

We converted raw HTML data into three variables: the news title, the original news link, and the original publishing date. From there, we expanded our variable list by using the sentimentr package in R to create two more variables: titlewordcount and titlesentiment. Since news headlines are not particularly large amounts of textual material, we aggregated sentiment by group instead of separating it by sentence. From there, we created a categorical variable to help define positive or negative headline evaluations by overall sentiment levels. This variable is called titlesentimentoverall. If the sentiment for a given headline was positive, we labeled it positive. If it was negative, negative. If it was zero, meaning it contained a sentimentr evaluation of 0, it was labeled neutral. The Source and Topic variables are our best guess estimates and explanation for where the news comes from based on the topics it purveys. Sometimes we labeled academic journalism, like the Cato Institute or the RAND Corporation, as blogs because they refer to their own material as a blog and the content generated there is less topical and cohesive and more independent. Meaning, they write whatever they want whenever they want to – kind of like a blog. The President variable is a dichotomous and categorical variable indicating who was president at the time the article was originally published. Lastly, we created one more categorical variable to help give the data more meaning. This variable is a value judgment on the **overall** Leaning of the news organization according to U.S. ideological standards. An organization was liberal if they were rated Left-Center or below by Media Bias Fact Check. An organization was conservative if they were rated Right-Center or below by Media Bias Fact Check. An organization was neutral if they were a least biased source according to Media Bias Fact Check. We upgraded a source from Left-Center or Right-Center in a few instances. Namely, as an example, we rated CNBC as a neutral source where Media Bias Fact Check rates them as Left-Center bias. Our in-house methodology allows us to upgrade from the bottom most consensus evaluation. It does not allow us to downgrade past a bottom most evaluation. This keeps us grounded. It also gives us the flexibility to buck back against evaluations that are overly critical. Sometimes when you spend too much time next to the fire everything is still hot and evaluators are not immune to the effects of being too close to a problem. Feel free to modify these value judgments depending on your methodology or need within the confines of our terms of use.

Collection

Our data collection was done with bash scripting so we could automate the workflow while centralizing control to an onsite server. We utilized a crontab to run our script once everyday with a **wget** command using a Linux console. We began our collection on September 24, 2020 and started cultivating it into a dataset on August 19, 2021. We did not change or modify this collection scheme at any time and we are still collecting with this scheme in real-time. We retain logs on all get requests for each individual requests each time our server attempts to collect data. This allows us to determine when there was a failed collection event and the reason for that failure. All the data currently available in the dataset is still the original data. We have only transformed data into more variables. We have not pruned or modified it in any way to ensure maximum transparency. There are days missing but no more than approximately 3 days worth of news and no more than twice each month (oftentimes way less) – and never in consecutive months.

Citation & Terms of Use

Curated News and Matthew Benchimol maintain the exclusive rights for this dataset. Permission is hereby granted, free of charge, to any person obtaining a copy of this dataset and associated documentation files (the "Dataset"), to deal in the Dataset without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Dataset, and to permit persons to whom the Dataset is furnished to do so, subject to the following conditions:

If you are an academic using this dataset for research related purposes, you must cite this dataset as follows:

> Benchimol, Matthew. 2021. “Curated News – Dataset.”

The Codebook can be cited as follows:

> Benchimol, Matthew. 2021. “Curated News – Codebook.”

All other forms of use must publicly state where the data comes from and whether it diverges from the original, explicitly stating how and where it does so. Should you like to use one of the graphs from our Curated Analytics section available on our website, please use the graph as it is in its original form with whatever data transformations you have made within the panel.

If you are a developer who would like to gain access to our dataset for use within your platform using the static json file hosted on our website, you must request written permission from our founder and developer, Matthew Benchimol, here.

This dataset is provided “as is,” without warranty of any kind, express or implied, including but not limited to the warranties of merchantability, fitness for a particular purpose and non-infringement. In no event shall the authors or copyright holders be liable for any claim, damages or other liability, whether in an action of contract, tort or otherwise, arising from, out of or in connection with this dataset or the use or other dealings in the dataset.