Algorithm Tips: A Resource for Algorithmic Accountability in Government

Daniel Trielli
Northwestern University
dtrielli@u.northwestern.edu

Jennifer A. Stark
University of Maryland
starkja@umd.edu

Nicholas Diakopoulos
Northwestern University
nad@northwestern.edu

ABSTRACT

Algorithms are coming to pervade both private and public systems as decision-making instruments. As a result, algorithmic accountability reporting has emerged as a suite of methods for investigating how algorithms are used to make consequential decisions, and where they can go wrong or otherwise cause impacts. This paper presents the Algorithm Tips project, which seeks to lower the cost of finding newsworthy leads about the use of algorithms in government by providing an easily searchable database. We detail the methods we used to create the database including several method refinements that may be useful to other computational journalists doing keyword monitoring on the web.

1 INTRODUCTION

As powerful and complex as they may be, algorithmic decision-making (ADM) systems can be as inherently faulty as their human creators. Whether by learning patterns from imperfect datasets [1] or by having inadvertent or even hard-wired discriminatory code, ADM systems are bound to reflect bias in one form or another. This can be problematic as government increasingly adopts ADM systems, as illustrated by the recent investigative series "Machine Bias" by ProPublica [2] While not the only agent of government accountability, the press has an important role as an independent check on the power governments wield via the algorithms they employ [3].

The topic of government algorithms presents itself as a new beat in journalism. However, there are institutional challenges that prevent journalists from exploring automated systems in government. The unknown characteristics, legal murkiness, and lack of standard language around algorithms introduce hassle costs that make the process of requesting such information from the government challenging, as shown by a survey of algorithm-related Freedom of Information Act requests [4]. But the challenge in covering algorithms as a beat begins even before that, at discovering the existence of algorithms and why, where, and how they are used throughout the government.

These challenges motivated the development of Algorithm Tips (www.algorithmtips.org), which we present in this paper. Algorithm Tips is an informational site and database meant as a first attempt to improve access to information about government algorithms. The initial goal is to develop and deploy a method for seeding a list of government algorithms, first in the United States federal jurisdiction, but eventually also extending to local, state, or international jurisdictions. This database provides a starting point to help orient journalists and researchers in pursuing algorithmic accountability investigations. The project is also geared towards being a repository of information about reporting on those algorithms by aggregating information, such as tutorials, examples, and methodological references that can help journalists and researchers overcome practical and technical challenges in those initiatives. Finally, there is the element of engaging a community of journalists and researchers through a system of volunteering in the project and inviting submissions of information. By aligning these three elements – data gathering, information sharing, and community building – Algorithm Tips seeks to be an initial step in facilitating algorithmic accountability as a journalistic beat.

The contribution of this work is not only the database and site, which hopefully will generate interest and a starting point for algorithmic accountability projects, but also a method for articulating and expanding the vocabulary of terms that can be used to search for and find government algorithms, which could inform future document-driven methods for identifying algorithms in society. Additionally, we develop and demonstrate a process for web monitoring; in this case, it was monitoring algorithms in government pages, but the method could be adapted to other topics and corpora that journalists may want to monitor.

2 RELATED WORK

2.1 Algorithmic Accountability

This work aims to be a practical yet conceptually grounded step towards improving journalistic coverage of algorithms in government. The need for such a beat has been discussed in our previous work and this project seeks to facilitate algorithmic accountability reporting by lowering the cost of finding leads that jumpstart investigations [3, 5].

The topic of algorithmic failure is increasingly subject to study and advocacy, particularly in regards to the possible effect it might have in amplifying discrimination and bias [6]. As the study of algorithmic impact on society advances, so does the discussion of methods through which algorithms can be audited, including auditing the actual code or analyzing results of the algorithms through scraping [7].

2.2 Computational Methods for Journalistic Monitoring

This work also aims to develop a hybrid method combining computational and human effort for collecting and monitoring documents for data journalism purposes. Discussions of computational or automated journalism frequently relate to the

generation of content [8]. However, there is also interest and attention in the field to how to monitor and analyze data (semi-) automatically in order to orient attention towards possible news stories. The benefits of adopting computational newsgathering processes are the increase of scale and improvement of productivity, as well as better accuracy and the ability to uncover stories that were hard to find [9]. Real-world examples have been very successful, such as a project that automatically monitored the status of books in the Philadelphia school districts and discovered discrepancies that generated news stories [10].

Newsrooms have been searching for solutions to collect and wrangle data automatically. Tools have been developed to allow journalists to monitor datasets that are updated regularly, in search for leads for information that can be quickly turned into stories [11]. Other software solutions include ways to acquire, translate, and extract information from unstructured data [12].

3 METHOD

The method for creating the Algorithm Tips database is described in the next subsections. The first phase is developing a method for searching links that might refer to government algorithms. Then the newsworthiness of those algorithms was evaluated. The last phase is to add contextual information to the algorithms that were found in order to create journalistically interesting leads for the database.

3.1 Searching and Classifying Links that Potentially Discussed Algorithms

A method of directed Google web searches, filtering, and data aggregation was used to create the database of newsworthy government algorithms. The premise was that references to ADM systems would be found in government documents stored on government websites.

The first challenge in finding those documents was to speculate on how algorithms are being referred to across various government agencies [13]. From our own experience in FOIAing algorithms from state agencies we know that the vocabulary used to describe algorithms internally may vary. For instance, an algorithm for predicting criminal recidivism might be referred to as a "risk assessment" by the relevant government agency. The first step in our process was, therefore, to create a list of algorithm-related terms to be used as the searches. This process was based on variations on terms that are associated with actions related to algorithms. Initially, 61 terms were listed, based on brainstorming and use of various thesauri to expand the term set. The full set of terms can be seen in Table 2, in the Appendix.

The next step was refining the search. Since the focus was on government links, the first decision was to limit the searches to United States government domains, which was achieved by using the "site:.gov" operator.

Initially, results included a high degree of noise due to research papers hosted in government websites that described their methodologies using many of the search terms. A majority of those research papers was hosted by the National Institutes of Health (NIH). While informative, initial assessments revealed that many of these references would not lead to algorithms necessarily being used by government agencies. Thus, they were filtered out of the searches by using another operator, "-site:.nih.gov"

As a result of that process, searches were constructed in the following format:

"[search term]" site:.gov -site:.nih.gov

The results for each search term were collected using the Linkclump Chrome plugin and exported to a spreadsheet. This raw data included the search term, the title of the webpage found, and the full link.

3.2 Finding Algorithms and Determining Newsworthiness

The links that were returned from each search were not always indicative of ADM systems. A preliminary analysis of the webpages showed many cases in which the selected search terms were being used to describe processes that were either not algorithmic or not used by government. Therefore, all the links had to be individually opened and read to determine the presence of an algorithm being used in a governmental decision process. This was determined by answering two questions: (1) Does this webpage have information about an algorithm? And (2) Is this process used or endorsed by the U.S. federal government? If both had affirmative answers, the link in the database would be considered to have an algorithm and would be considered in the next stage of tagging.

For the purposes of Algorithm Tips the definition of an algorithm is a set of rules to which data can be input and from which a result – a decision, a recommendation, a score – is obtained. That means that the algorithm can be computational (a computer software or spreadsheet) or not computational (a weighted score card or flowchart). Algorithms can also be in active use in government operations or can be endorsed by the government to assist third-party decisions.

If an algorithm was present, the next step was determining if the algorithms had the potential for being interesting to journalists – was it "newsworthy". For that, the guidelines came from literature revolving around traditional journalism. Harcup and O'Neill [14] propose "an updated set of contemporary news values that, in various combinations, seem to be identifiable within published news stories." They write that "potential news stories must generally satisfy one and preferably more" of 15 criteria. Out of those, five were deemed to be particularly relevant to algorithms, including:

- **Bad news**: Stories with particularly negative overtones such as death, injury, defeat and loss (of a job, for example).
- **Conflict**: Stories concerning conflict such as controversies, arguments, splits, strikes, fights, insurrections and warfare.
- **Surprise**: Stories that have an element of surprise, contrast and/or the unusual about them.

- **The power elite**: Stories concerning powerful individuals, organizations, institutions or corporations.
- **Magnitude**: Stories perceived as significant in the large numbers of people or amounts of money involved or involving a degree of extreme behavior or extreme occurrence with potential for impact.

Based on those criteria, five questions were formulated to evaluate potential newsworthiness of each algorithm. If an algorithm included at least one positive response to any of these questions, it was included in the database. The five questions are:

- Can this algorithm have a negative impact if used inappropriately?
- Can this algorithm raise controversy if adopted?
- Is the application of this algorithm surprising?
- Does this algorithm privilege or harm a specific subset of people?
- Does the algorithm have the potential of affecting a large population or section of the economy?

3.3 Enriching Leads with Metadata

The mere listing of the names or links to the algorithms would not suffice to provide journalists and researchers a set of useful leads. Additional information can provide context to help users decide whether investigating a given algorithm may be worth their time and effort. Therefore, the database was enriched with relevant metadata, such as information coming from the links or documentation found via the web search. The choice for the additional information fields was motivated by the possible angles to research and reporting.

To get the government agency name the domain name was extracted from the full link. Then a dataset of all .gov domains, compiled by Mill and Brooks [15], was used as reference to match the links to their respective United States government agencies. The other metadata fields listed in Table 1 were then manually added to each entry in the database. The presentation of these fields on the website was aimed to facilitate rapid assessment, with the name, what the algorithm does and why it is important at the top (providing journalistic context) followed by government context such as agency, topic, and who created it.

Table 1: Fields in the Algorithm Tips database

Field	Description
Name	The name of the algorithm (trademarked name or working name) or, in the absence of a clear name, a short description
Description	General description of what algorithm does
Why it is important	Possible impacts of the algorithm as they relate to the newsworthiness factors
Topic	General field or domain
Jurisdiction	Name of the country, state, or local government associated with deploying the algorithm
Government level	Whether that jurisdiction is federal, state, or local
Agency	Name of the government agency that uses or recommends it
Proprietary	Whether or not that algorithm was developed by a contractor and is owned by a company
Creator/author/vendor	Name of company or agency that created that algorithm
Date	Month and year in which the algorithm was launched or updated
Adoption Stage	Whether it is directly used by the government ("Active use"), developed or shared by the government ("Endorsement for use"), or is being studied or evaluated for use but not adopted yet ("Potential use")
Computational	Whether the algorithm is implemented in software as opposed to a non-software calculation such as a flowchart
Link	URL that links to algorithm documentation showing initial evidence of existence.

4 RESULTS

The initial web search using algorithmic-related terms yielded 5,337 links in total, with 300 showing up more than once across different search queries. The search was conducted between November 4 and 9, 2016. Of the 5,337 links, 2,908 were from federal agencies, the focus of this research. The remaining 2,429 belong to state and local governments. From those 2,908 federal links, 228 pointed to algorithms that was either directly used or endorsed by the federal government. There are fewer algorithms in the final database (159) because some links pointed to the same algorithms. The top ten federal agencies where we identified newsworthy algorithms are shown in Figure 1.

Some algorithms had a clear potential for journalistic interest. For instance, the Health Professional Shortage Area (HPSA) designation score is a set of scored metrics that determine if an area has a shortage of health care providers. A miscalculation in that algorithm could lead to negative impacts in areas with real shortages of health professionals that might not be included on the list. Another example is the Fraud Prevention System (FPS), which uses predictive modeling technology to prevent fraudulent Medicare payments. Errors could lead to high-impact losses for the government or providers. These examples should not be seen as anywhere approaching a finished journalistic product, but rather as leads to investigative stories. As previous research into computational journalism has pointed out, algorithms can surface data and ideas to be further explored by reporters. These are tools to supplement rather than substitute for efforts by reporters." [16]

Some search terms yielded more links than others, both in actual and proportional numbers. "Algorithm", "automated

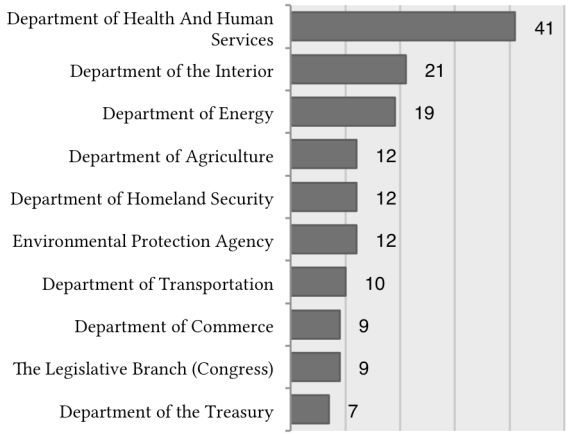

Figure 1: **Government agencies with the highest proportion of algorithms returned by the search.**

assessment", and "numerical rating" were the terms with most links returned (23, 16 and 14 respectively). High ratios of algorithms to links were yielded by the terms "grading calculation" (1 out of 2 links), "algorithm", (45%, or 9 out of 20 links), and "numerical rating" (25%, or 1 out of 4 links).

These results raised two questions, which in turn motivated some refinements to the methodology: (1) Are the selected search terms covering the entire corpus of algorithms or are important queries being overlooked? And (2) Is there a way to remove search terms that are less productive or which yield results that are already returned by more productive terms? The next section presents methods to deal with these two issues – term expansion and term reduction.

5 METHOD REFINEMENT

As the previous section showed, analyzing the results of the searching and tagging method raised two possibilities: first, that some useful terms were not used, since they were not in the initial list of brainstormed terms; and second, some terms that were used yielded few unique results and therefore could be discarded in future data collections. Methods to address these two questions are described in the next subsections.

5.1 Finding New Search Terms

In order to find overlooked terms that would also describe algorithms used by the government, we examined the frequency of terms in the initial set of documents that were deemed to refer to newsworthy algorithms. In this way we hoped to leverage our manual tagging efforts and suggest new terms. Using Python, natural language processing was applied to find the most frequent bigrams in those documents. A list of 94,189 bigrams was extracted, with 10,851 appearing more than once. A cursory look showed that most of them were not useful, whether they were too broad or just irrelevant as descriptors. For example, the three most common terms were “http www”, “long term”, and “additional information”. Therefore, a new evaluation had to take place to assess which of these bigrams might actually be productive.

The three authors each evaluated the 200 most frequent bigrams. Each researcher independently rated each bigram using a scale inspired by past research on lexicon expansion [17]: 1 for strongly algorithm-specific (the term is likely to appear more often on pages about algorithms); 2 for weakly algorithm-specific (the term could appear frequently on pages about algorithms); and 3 for not algorithm-specific (the term should not appear more often on pages about algorithms).

Terms that had an average of 1.3 or lower were selected for a new batch of web searches and link collections. Fifteen new terms fit that profile, and they are listed in the Table 3, in the Appendix. Using those terms in the search method described in section 3.1 yielded 2,250 results in total. Of those, 1,154 of those are related to federal agencies, and the rest are from local governments. And out of those, 43 links were already collected on the original list. This means that the lexicon expansion method yielded 1,111 new links with potential algorithms used by the federal government. We leave to future work the evaluation of these links for eventual inclusion in the database.

5.2 Removing Redundant Terms

While on one hand we can expand the set of terms by mining for related terms in documents already collected, on the other hand we also wanted to assess which terms were not productive or which provided results that were either wholly or partially redundant with other terms. The goal was to cull terms that were not productive so that future data collection passes would be more efficient.

To do this we devised an algorithm to compute the recall for each term (or set of terms) based on the number of newsworthy algorithms that term (or set of terms) led to. At each step of the algorithm the term that maximized recall was added to a final set of terms. Once the optimal set of terms had reached a recall of 1.0 this signaled that any remaining terms that were not in that set were redundant. In other words, the optimal set of terms had already uncovered all of the newsworthy algorithms and any additional terms would just lead to one of the algorithms that a previous term had already uncovered. In the end this method helped identify nine terms that were redundant, including: “automated simulation”, “automatic score”, “ranking equation”, “scoring equation”, “automatic rating”, “automatic scoring”, “grading method”, “calculating matrix” and “computing”. These terms, which yielded 327 of the 2,908 links analyzed (11%), could be removed from our list of search terms in future data collections in order to reduce manual tagging effort.

6 PRESENTING THE RESULTS ON A WEBSITE

Since the motivation of Algorithm Tips is to be a starting point for journalists and researchers in their algorithmic accountability projects, and considering the fact that these projects are varied by nature, the website on which the database

is presented was formulated to allow an exploratory experience through the database.

Design of the website, Algorithmtips.org, had two main criteria: first, it should have the database itself as the primary focus; and second, it should have a database that would be searchable by any term in any metadata field.

The result is a home page that has a relatively small area for contextualization and explanation of the project, and a larger area devoted entirely to listing entries from the database, with a search field to allow the user to explore it instantly. The results of the search are filtered on the home page itself.

The navigation bar includes links to methodology, further readings and a community engagement section. As previously discussed, since the last point is one of the motivators of the project, a special consideration was given to different means of contributing to the project. The first is an invitation to submit information about algorithms, via a Google Form. The second is an invitation to more involved volunteering, with the purpose of analyzing more links and attempt to discover more algorithms through the described methodology.

7 DISCUSSION

This work had two goals: first, to develop a system that allowed for monitoring of search engine results in search of specific documents on specific websites. And from that, came a new understanding of how the words used to describe a set of objects can vary. This project started with a list of 61 terms that were thought to be the most inclusive and productive in searching for algorithms. Through the process of methodologically testing, expanding, and reducing this list of terms, the project allowed for a new understanding of how lexicon building is essential for this type of work.

The other goal was to develop a database of government algorithms to provide a practical stepping-stone for journalists and researchers, and improve discussion of algorithmic accountability. In doing the qualitative analysis of the algorithms uncovered by this process, a new dimension of their variety and their journalistic potential was acquired.

Both objectives have potential for future exploration. The goal of developing a method for monitoring government websites based on search engines can be improved upon with new research on lexicon building, including crowdsourcing the development of key words related to specific target topics.

On the other track, the development of a database and a community around accountability of government algorithms can be expanded with the use of crowdsourcing to analyze more links in search of algorithms, and volunteers to replicate this database in other governmental jurisdictions and share information about existing algorithms. This collaboration and information sharing can strengthen and help solidify the coverage of government algorithms as a new beat in journalism.

ACKNOWLEDGMENTS

This work was partially supported through grants from the Rutgers Institute for Information Policy and Law (RIIPL) as well as from the Tow Center for Digital Journalism at Columbia University.

REFERENCES

[1] Aylin Caliskan-Islam, Joanna J Bryson, and Arvind Narayanan. 2016. Semantics derived automatically from language corpora necessarily contain human biases. arXiv1608.07187v2 [cs.AI] 30 Aug 2016 (2016), 1–14.

[2] Julia Angwin, Surya Mattu, Jeff Larson, Lauren Kirchner. 2016. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. (May 2016).

[3] Nicholas Diakopoulos. 2015. Algorithmic Accountability : Journalistic investigation of computational power structures. Digit. Journal. 3, 3 (2015), 398–415.

[4] Katherine Fink. 2017. Opening the government's black boxes: freedom of information and algorithmic accountability. Information, Commun. Soc. 4462, July (2017), 1–19.

[5] James Hamilton. 2016. *Democracy's detectives: the economics of investigative journalism*, Cambridge, MA: Harvard University Press.

[6] Emily Shaw. 2015. Avoiding prejudice in data-based decisions. Sunlight Foundation. Retrieved from https://sunlightfoundation.com/2015/07/31/avoiding-prejudice-in-data-based-decisions/

[7] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing Algorithms: Research Methods for Detecting Discimination on Internet Platforms. In Data and Discrimination.

[8] Konstantin Nicholas Dörr. 2016. Mapping the field of Algorithmic Journalism. Digit. Journal. 4, 6 (2016), 700–722.

[9] Andreas Graefe. 2016. Guide to Automated Journalism. (2016).

[10] Meredith Broussard. 2015. Artificial Intelligence for Investigative Reporting. *Digit. Journal.* 3, 6 (2015), 814–831.

[11] BBC News Labs Shearer, Matt; Simon, Basile; Gieger, Clément. 2014. Datastringer: easy dataset monitoring for journalists. Proc. Computation + Journalism Symposium.

[12] Titus Plattner, Didier Orel, and Olivier Steiner. 2016. Flexible data scraping , multi-language indexing , entity extraction and taxonomies : Tadam , a Swiss tool to deal with huge amounts of unstructured data. Proc. Computation + Journalism Symposium.

[13] Nicholas Diakopoulos. 2016. We need to know the algorithms the government uses to make important decisions about us. The Conversation. https://theconversation.com/we-need-to-know-the-algorithms-the-government-uses-to-make-important-decisions-about-us-57869

[14] Tony Harcup and Deirdre O Neill. 2016. What is news? News values revisited (again). Journal. Stud. 9699, March (2016), 1–19.

[15] Eric Mill and Gray Brooks. 2014. A complete list of .gov domains. (2014).

[16] Fred Turner and James T. Hamilton. 2009. Accountability Through Algorithm: Developing the Field of Computational Journalism. A Cent. Adv. Study Behav. Sci. Summer Work. Rep. (2009), 1–22.

[17] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2014. CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises. Proc. 8th Int. Conf. Weblogs Soc. Media (2014), 376.

APPENDIX

Table 2: List of search terms used on initial algorithm web search (grayed terms were subsequently identified as unproductive)

Algorithm	Automatic ranking	Grading methodology	Rating method
Algorithmic	Automatic rating	Grading model	Rating methodology
Automated analysis	Automatic score	Numerical rating	Rating model
Automated assessment	Automatic scoring	Predictive Analytics	Scoring calculation
Automated calculation	Automatic sorting	Predictive modeling	Scoring equation
Automated filtering	Calculating matrix	Ranking calculation	Scoring formula
Automated grading	Calculating method	Ranking equation	Scoring matrix
Automated ranking	Calculating model	Ranking formula	Scoring method
Automated rating	Computation	Ranking matrix	Scoring model
Automated scoring	Computational	Ranking method	Statistical assessment
Automated simulation	Computing	Ranking methodology	Statistical methodology
Automated sorting	Grading calculation	Ranking model	Statistical model
Automatic assessment	Grading equation	Rating calculation	Statistical software
Automatic calculation	Grading formula	Rating equation	
Automatic filtering	Grading matrix	Rating formula	
Automatic grading	Grading method	Rating matrix	

Table 3: List of search terms extracted from bigrams and evaluated for relevancy

Assessment tool	Numerical score	Risk based
Automated system	Overall score	Risk ranking
Calculated using	Ranking system	Score based
Model used	Rating system	Scoring methodology
Monitoring system	Risk assessment	Software tool