The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

Proceedings of the Undergraduate Consortium Mentoring Program

New York, New York, USA

February 7-12, 2020

Quantifying the Effect of Unrepresentative Training Data on Fairness Intervention Performance

Jessica Dai[1], **Sarah M. Brown**[2]
Brown University
[1]jessica.dai@brown.edu
[2]sarah_m_brown@brown.edu

Abstract

While many fairness interventions exist to improve outcomes of machine learning algorithms, their performance is typically evaluated with the assumption that training and testing data are similarly representative of all relevant groups in the model. In this work, we conduct an empirical investigation of fairness intervention performance in situations where data from particular subgroups is systemically under- or over- represented in training data when compared to testing data. We find post intervention fairness scores vary with representation in often-unpredictable and dataset-specific ways.

Introduction

As machine learning algorithms are applied to ever more domains, the data used to train these models is also increasingly messy; fairness interventions aim to prevent algorithms from reproducing societal biases encoded in data. These interventions are generally evaluated under the assumption that the training data is well-representative of the data on which the model will be deployed. However, systemic disparities in group representation in training data is uniquely likely in domains where historical bias is prevalent.

Our main question is: how does the oversampling or undersampling of a particular group affect the performance of the post-intervention algorithm, in terms of overall accuracy and in terms of various measures of fairness? We resample existing datasets to simulate different proportions of demographic groups for training and testing, extending the previous work of Friedler et al. 2019 to evaluate the performance of those interventions on our artificially unrepresentative test-train splits of the data; this serves as our proxy for real-world unrepresentative data.

We find that changing the representation of a protected class in the training data affects the ultimate performance of fairness interventions in somewhat unpredictable ways. For the rest of this paper, we will use *representation effects* to describe the way in which changing representation affects fairness performance. In particular, our results are:

1. **Fairness-accuracy tradeoff.** Representation effects with regards to the fairness-accuracy tradeoff are inconsistent even within a specific dataset; in each of the datasets we analyzed, they differ depending on the algorithm and intervention being analyzed. The only generalization to be made is that representation effects for an *intervention* on a baseline algorithm follow the same pattern as representation effects on the baseline itself.
2. **Calibration-error rate tradeoff.** Representation effects with respect to the calibration-error rate tradeoff are also inconsistent across datasets, but representation effects of different algorithms *are* consistent within a single dataset.

Related work

Existing fairness interventions and metrics. A wide variety of intervention strategies exist. These include modifications or reweighting of the training data (Feldman et al. 2015) and modifications of the objective function (Kamishima et al. 2012; Zafar et al. 2017), among other approaches. At the same time, there are a variety of ways to quantify "fairness," including base rates and group-conditioned classification statistics (i.e. accuracy and true/false positive/negative rates for each group).

Class imbalance and distribution shift. While these are known problems in machine learning broadly, they are largely unconsidered in the context of fairness interventions; they typically assume variation in the distribution of the target variable, while we are interested in the distribution of the protected class. Scholarship in this area often suggests some method for oversampling (e.g. Fernández, García, and Herrera), albeit more nuanced than the experiments run here.

Existing empirical survey. Friedler et al. published an empirical comparison of several fairness interventions across multiple datasets with an open source framework for replication. They found that intervention performance is context-dependent—that is, varies across datasets—and empirically verified that many fairness metrics directly compete with one another. This survey also investigated the relationship between fairness and accuracy (which has often been characterized as a tradeoff), noting that stability in the context of fairness is much lower than accuracy.

Experimental setup

We preserve the experimental pipeline of Friedler et al.: For each dataset, we run standard algorithms (SVM, Naive Bayes, Logistic Regression) and several fairness intervention algorithms introduced by Feldman et al.; Kamishima et al.; Calders and Verwer, and Zafar et al.. For each run of each algorithm, we compute overall accuracy and a variety of fairness metrics. However, in each experiment, we replace the train-test splits—which were random in Friedler et al.'s original work—to simulate unrepresentative data.

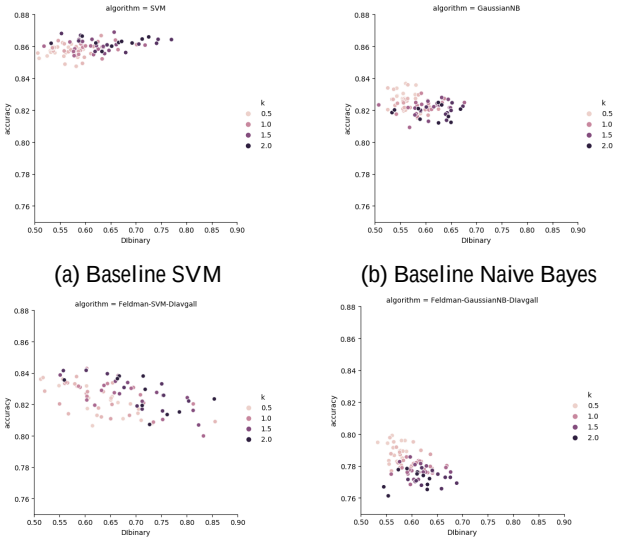
To simulate unrepresentativeness, we create train-test splits for each dataset that represent a variety of distribution shift or oversampling possibilities. We introduce a parameter $k = \frac{q}{r}$, where q is the proportion of the protected class in the training set and r is the proportion of the protected class in the testing set, so that for $k = 0.5$ the disadvantaged group is half as prevalent in the training set as it is in the testing set (underrepresented), and for $k = 2.0$ the protected class is twice as prevalent (overrepresented). We run a series of experiments each for a value of k in $\frac{1}{2}, \frac{4}{7}, \frac{2}{3}, \frac{4}{5}, 1, \frac{5}{4}, \frac{3}{2}, \frac{7}{4}, 2$. We use 80-20 test train splits in all experiments. Most new work on this project is my own; Dr. Sarah Brown advises this project, providing guidance on framing research questions and formulating new approaches.

Results & evaluation

For the following figures, each dot represents the statistic calculated for one run of the algorithm. A darker dot indicates a higher k.

Fairness-accuracy tradeoff. Representation effects in the context of the fairness-accuracy tradeoff are inconsistent not only across datasets but for algorithms within the same dataset as well. The Adult dataset (figure 1) is one example of this phenomenon: increasing training representation appears to increase both fairness and accuracy in SVM algorithms, but reduces accuracy with little impact on fairness in Naive Bayes algorithms. Interestingly, when fairness interventions are applied to the same baseline algorithms, the representation effects on the interventions follow the same general pattern as representation effects on the baseline algorithms. Here, fairness is measured through disparate impact (formally, $\frac{P(\hat{Y}=1,S=1)}{P(\hat{Y}=1,S=1)}$ where $\hat{Y}$ is the predicted label and $S = 1$ indicates the privileged demographic).

Calibration-error rate tradeoff. It is impossible to achieve equal true positive and negative calibration rates across groups, given unequal base rates (Kleinberg, Mullainathan, and Raghavan 2016). More formally, we compare the true positive rate (TPR) of the unprivileged demographic ($P(\hat{Y} = 1|Y = 1, S = 0)$ to the negative calibration of the same demographic ($P(Y = 1|\hat{Y} = 1, S = 0)$). Here, representation effects also differ across datasets, though different algorithms within the same dataset respond similarly to changes in representation, as illustrated in figure 2. While the general shape of the tradeoff follows the expected downward slope in each dataset and each algorithm, note that in the Adult dataset, representation appears to have little effect on TPR, while it tends to increase TPR in the Ricci dataset.



(a) Baseline SVM (b) Baseline Naive Bayes

(c) Feldman SVM (d) Feldman Naive Bayes

Figure 1: Fairness-accuracy tradeoff of a subset of algorithms run on the Adult dataset. Disparate impact is on the horizontal axis, while accuracy is on the vertical axis.

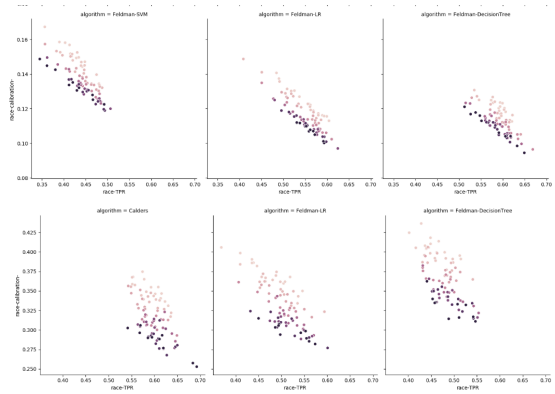


Figure 2: Calibration-error rate tradeoff in Adult (top) and ProPublica (bottom) datasets. TPR is on the horizontal axis.

Discussion

These empirical results further illustrate the importance of context and domain awareness when considering the "fairness" of an algorithm. In particular, the somewhat unpredictable representation effects across datasets and algorithms suggest a need for a rethinking of approaches to fairness interventions; while (over)representation may sometimes be helpful, it is clear that datasets contain some intrinsic properties that affect observed fairness.

In future work, we hope to develop a model which provides a theoretical explanation for our results; this also aids us in commenting on the interpretation of "fairness results," as well as arriving at a framework for understanding *a priori* when overrepresentation in training data may be helpful.

References

Calders, T., and Verwer, S. 2010. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21(2):277–292.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268. ACM.

Fernández, A.; García, S.; and Herrera, F. 2011. Addressing the classification with imbalanced data: open problems and new challenges on class distribution. In *International Conference on Hybrid Artificial Intelligence Systems*, 1–10. Springer.

Friedler, S. A.; Scheidegger, C.; Venkatasubramanian, S.; Choudhary, S.; Hamilton, E. P.; and Roth, D. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 329–338. ACM.

Kamishima, T.; Akaho, S.; Asoh, H.; and Sakuma, J. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 35–50. Springer.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Zafar, M. B.; Valera, I.; Rogriguez, M. G.; and Gummadi, K. P. 2017. Fairness Constraints: Mechanisms for Fair Classification. In Singh, A., and Zhu, J., eds., *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, 962–970. PMLR.

Activity Recognition Using Deep Convolutional Networks for Classification in the SHL Recognition Challenge

Michael Sloma

Abstract

Project Goals

Personal Contributions

Previous Work

Prospective Next Steps

Results & Learning Opportunities

References

Application of Contributions

Harmful Feedback Loops in Unequitable Criminal Justice AI Models

Pazia Luz Bermudez-Silverman
Brown University, Department of Computer Science, Data Science Initiative
78 Arnold St. Fl 2
Providence, RI 02906
pazia_bermudez-silverman@brown.edu

Abstract

My research focuses on the harm that AI models currently cause and the larger-scale potential harm that they could cause if nothing is done to stop them now. In particular, I am focusing on AI systems used in criminal justice, including predictive policing and recidivism algorithms. My work synthesizes previous analyses of this topic and steps to make change in this area, including auditing these systems, spreading awareness and putting pressure on those using them.

Introduction

As many researchers have recently shown, AI systems used by law enforcement and the public to make decisions that directly impact people's lives perpetuate human biases surrounding race, gender, language, skin color, and a variety of intersections of these identities (Albright 2019; Buolamwini and Gebru 2018; Lum and Isaac 2016). While these biases already existed in our society long before AI and modern technology, AI algorithms and models reinforce them at an unprecedented scale. In addition, these models' feedback loops strengthen such biases by perpetuating harm to communities already at risk (O'Neil 2016). We see these algorithms and their harmful feedback loops in areas such as education, criminal justice and housing, but this paper will focus on criminal justice algorithmic models and their effects on lower-income communities of color.

Background

Feedback loops and proxy attributes are essential for understanding the scale of harm and influence AI models have on this society, especially in the criminal justice system.

Feedback Loops

Feedback is essential to the accuracy of AI algorithms and models. Without it, a model will never know how well or how poorly it is performing and thus, it will never get better. However, depending on which feedback is given to a model, that model will change and behave in particular ways according to that feedback. This is a feedback loop.

Cathy O'Neil characterizes one of the main components in her definition of a "Weapon of Math Destruction" (WMD) as having a "pernicious feedback loop" (O'Neil 2016) that contributes to the power and harm of the WMD. Such feedback loops occur when an algorithm either does not receive feedback on its output or receives feedback that is not accurate in some way. O'Neil cites that organizations using AI algorithms allow for the continuation and growth of these pernicious feedback loops because they look at the short term satisfaction of their consumers rather than the long term accuracy of their models (O'Neil 2016). As long as companies are making money or organizations are meeting their goals, it is much easier for them ignore the harm that these models are causing. Additionally, Virginia Eubanks cites this "feedback loop of injustice" (Eubanks 2018) as harming specifically our country's poor and working-class people, mostly people of color, through examples of automated algorithms used in welfare, child and family and homeless services.

Proxy Attributes

While most predictive policing and other AI models used in law enforcement such as recidivism algorithms used to determine sentence length and parole opportunities do not directly take into account sensitive attributes such as race, other attributes such as zip code, friends and family criminal history, and income act as proxies for such sensitive attributes (Adler et al. 2018). In this way, predictive policing and recidivism prediction algorithms do directly make decisions having to do with race and other sensitive attributes.

These proxy attributes can actually directly lead to pernicious feedback loops not only because they are easier to "game" or otherwise manipulate, but also because the use of such proxies might make the model calculate something other than what the designers/users think. This can lead to false data (false positives or negatives) that is then fed back into the model, making each new iteration of the model based on false data, further corrupting the feedback loop (O'Neil 2016). Eubanks exemplifies this in her description of how using proxies for child maltreatment cause higher racial biases in automated welfare services (Eubanks 2018).

Predictive Policing

The most common model used for predictive policing in the U.S. is PredPol. We will focus on how this software perpetuates racial and class-based stereotypes and harms lower-income communities of color particularly through its pernicious feedback loops.

One reason for skewed results that are biased toward lower-income neighborhoods populated mostly by people of color is the choice of including two different types of crimes. Either only "Part 1 crimes" which are more violent like homicide and assault or also "Part 2 crimes" which are less violent crimes/misdemeanors such as consumption and sale of small quantities of drugs (O'Neil 2016). "Part 2 crimes" are often associated with these types of neighborhoods in our society. By following these results, law enforcement will send more officers into those areas, who will then "catch more crime," feed that data back into the model, perpetuating this pernicious feedback loop and continuing to send officers to these communities instead of other, more affluent and white areas.

Feedback Loops in PredPol Software

Crime is observed in two ways: law enforcement directly sees "discovered" crime and the public alerts them to "reported" crime. "Discovered" crime is a part of the harmful feedback loop: the algorithm sends officers somewhere and then when they observe crime the predictions are confirmed. Predpol is trained on observed crime which is only a proxy for true crime rates. PredPol lacks feedback about areas with "lower" crime-rates according to the model by not sending officers there, contributing further to the model's belief that one region's crime rate is much higher than the other. Given two areas with very similar crime rates, the PredPol algorithm will *always* choose the area with the slightly higher crime rate because of the feedback loop (Ensign et al. 2017).

Auditing Practices and Further Interventions

Auditing has been used by researchers such as Raji and Buolamwini, Adler et al. and Sandvig et al. to evaluate the accuracy and abilities of AI models as well as potential harm they could cause by inaccuracy such as through pernicious feedback loops. Corporations are not likely to change the way they use these models if change does not contribute to one of the following areas: "economic benefits, employee satisfaction, competitive advantage, social pressure and recent legal developments" (Raji and Buolamwini 2019). This means that external pressure, namely public awareness and organizing is necessary for change.

Ensign et al. propose an "Improvement Policy" for the PredPol software which suggests a filtering of feedback given to the model. They recommend that the more police are sent to a given district, the less weight discovered incidents in that area should count in feedback data (Ensign et al. 2017). They conclude that most "discovered" crime should not be counted in feedback data. This may still miss some crimes, but it is a better proxy, especially for use in algorithms, because it is not directly influenced by the model's previous choices. This should create a more equitable outcome that does not continue to target impoverished neighborhoods and communities of color.

A Socio-Technical Analysis of Feedback loops

One key question I am exploring involves how these analyses fit into the traps defined by Selbst et al., particularly the Ripple Effect Trap, which essentially speaks to the concept of feedback loops.

To solve this problem, we propose an investigation into whether such feedback loops contribute to the amplification of existing human biases or to the creation of new biases unique to such technology. Additionally, we hope to analyze the best ways to spread public awareness and influence companies and organizations to make changes in the way they use AI technologies. First analyses of these methods are discussed below and will be continued throughout this research.

Public Awareness

So, how do we combat these pernicious feedback loops and actually change the structure of the models and the way that organizations use them? The first step to making positive change in AI is spreading public awareness of the harm that AI systems currently cause and the misuse of them by law enforcement in these cases particularly. This can and should be through not only academic papers and articles, but through political activism on and off the web. As Safiya Noble has shown, the more people that understand what is currently happening and what we can possibly do to change it, the more pressure that is put on the companies, organizations and institutions that use these harmful models, which will encourage them to change the way they use the models and the way the models work in general (Noble 2018).

Next Steps and Timeline

I am currently working on a survey paper evaluating feedback loops and bias amplification through a socio-technical lens. Specifically, I will focus on the unique roles of AI researchers, practitioners, and activists in combating the harm caused by feedback loops. To investigate the question of how feedback loops amplify existing societal biases and/or create new unique biases, I am analyzing texts more relevant to my background in Africana Studies. This helps to provide societal context and background to this AI research.

Algorithms currently important in my research include PredPol (predictive policing software (Ensign et al. 2017)), COMPAS (the recidivism algorithm (Larson et al. 2016; Broussard 2018)), VI-SPDAT (used in automated services for the unhoused (Eubanks 2018)), as well as facial recognition software used by law enforcement (Garvie et al. 2016).

Following this academic version, I will translate and convey these findings into actionable insights accessible to all. Making my research available to many people will contribute to the public awareness I believe is necessary to combat the negative impacts of AI.

References

Adler, P.; Falk, C.; Friedler, S. A.; Nix, T.; Rybeck, G.; Scheidegger, C.; Smith, B.; and Venkatasubramanian, S. 2018. Auditing black-box models for indirect influence. *Knowledge and Information Systems* 54(1):95–122.

Albright, A. 2019. If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions.

Broussard, M. 2018. *Artificial unintelligence: how computers misunderstand the world*. MIT Press.

Buolamwini, J., and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91.

Ensign, D.; Friedler, S. A.; Neville, S.; Scheidegger, C.; and Venkatasubramanian, S. 2017. Runaway feedback loops in predictive policing. *arXiv preprint arXiv:1706.09847*.

Eubanks, V. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Garvie, C.; Privacy, G. U. C. o.; Technology; and Technology, G. U. L. C. C. o. P. . 2016. *The Perpetual Line-up: Unregulated Police Face Recognition in America*. Georgetown Law, Center on Privacy & Technology.

Larson, J.; Mattu, S.; Kirchner, L.; and Angwin, J. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)* 9.

Lum, K., and Isaac, W. 2016. To predict and serve? *Significance* 13(5):14–19.

Noble, S. U. 2018. *Algorithms of oppression: How search engines reinforce racism*. nyu Press.

O'Neil, C. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Raji, I. D., and Buolamwini, J. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *AAAI/ACM Conf. on AI Ethics and Society*, volume 1.

Sandvig, C.; Hamilton, K.; Karahalios, K.; and Langbort, C. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 22.

Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. ACM.

Search Tree Pruning for Progressive Neural Architecture Research Summary

Deanna Flynn
deannaflynn@gci.net
University of Alaska Anchorage
401 Winfield Circle
Anchorage, Alaska 99515

Abstract

A basic summary of the research in search tree pruning for progressive neural architecture search. An indiciation of the work which I contributed, as well as the advantages and possible ways the research can continue.

Summary of Research

We develop a neural architecture search algorithm that explores a search tree of neural networks. This work contrasts with cell-based networks ((Liu et al. 2017), (Liu, Simonyan, and Yang 2018)) and uses Levin search, progressively searching a tree of candidate network architectures (Schmidhuber 1997). The algorithm constructs the search tree using Depth First Search (DFS). Each node in the tree builds upon its parent's architecture with the addition of a single layer and a hyperparameter optimization search. Hyperparameters are trained greedily, inheriting values from parent nodes, as in the compositional kernel search of the Automated Statistician (Duvenaud et al. 2013).

We use two techniques to constrain the architecture search space. First, we constructed a transition graph to specify which layers can be inserted into the network, given preceding layers. The input and output layers are defined by the problem specification. Second, we prune children from the tree based on their performance relative to their parents' performance or we reach a maximum depth. The tree search is halted when no children out-perform their parents or we have reached the maximum depth.

The algorithm was tested on the CIFAR-10 and Fashion-MNIST image datasets ((Xiao, Rasul, and Vollgraf 2017), (Krizhevsky, Hinton, and others 2009)). After running our algorithm on a single Intel i7 8th generation CPU on the Fashion-MNIST dataset for four days, we generated 61 networks with one model achieving a benchmark performance of 91.9% accuracy. It is estimated our algorithm only traversed between a fifth and a third of the search tree. This result was acquired in less time than it took other benchmark models to train and test on the same dataset. Table 1 shows a comparison of other benchmark models on the Fashion-MNIST dataset.

When testing the CIFAR-10 dataset, our processing time was limited to 24 hours. We placed a depth limit on the search tree and trained on ten percent of the original dataset. In thirteen hours, on an Intel Xeon Gold 6154 CPU and NVIDIA Tesla V100-SXM2-32GB, the algorithm generated a network with an accuracy of 55.9%.

This algorithm is limited to finding feed-forward networks. Although contingent on the transition graph, the algorithm is simple to implement. However, by dealing with layers directly, it incorporates the macro-architecture search required in cell-based neural architecture search. The progressive nature of Levin search makes the algorithm relevant to resource-constrained individuals who need to find the simplest network that accomplishes their task.

What I Contributed

The research presented was intended for my summer internship at NASA working on a project called Deep Earth Learning, Training, and Analysis (DELTA). DELTA analyzes satellite images of rivers to predict flooding and uses the results to create flood maps for the United States Geological Survey (USGS). Also, work was beginning to examine images for identifying buildings which were damaged in natural disasters. For both aspects, a manually created neural network was used for the identification and learning process of each image.

The suggestions of the outcome for the research were predefined by my mentors who wanted to investigate having a neural network be automatically generated for a specific task. First, some form of search tree was to be created with every node representing a neural network. Next, each child node in the search tree would contain an additional network layer somewhere within the previous neural architecture. Finally, the tree was to have the capability of performing basic pruning.

The first problem which needed to be addressed was the creation of the neural networks. This included both what layers were used within the architecture and how to methodically insert the layers within pre-existing architectures. For simplicity purposes, only five layers were considered in the initial design: *Convolution, Flatten, Max Pooling, Dropout,* and *Dense*. As for determing what layers can be inserted, I constructed a transition graph through studying and testing multiple neural networks and carefully watching what

Various Perspectives of Studying Signals for Evaluating the Credibility of Online Sources of Information

Khonzodakhon Umarova
Department of Computer Science
Wellesley College
kumarova@wellesley.edu

Introduction

Misinformation, in its various forms and interpretations, has been part of societies for perhaps as long as actual information has: from gossip to myths to government issued propaganda. As technologies for mass communication evolved, so did the ways in which misinformation travels.

In the present day, people from all over the world query Google Search to access the information they need. Previous studies have shown that people automatically trust search results without realizing the work of algorithms behind the scenes in identifying and ranking “relevant” information (Pan et al. 2007). However, blind dependence on algorithms is problematic. Users who are disincentivized from critically evaluating information they see online are vulnerable to false, incomplete, or misleading results that slip through the net of algorithms. Therefore, it is integral for web users to develop web literacy skills that would empower them to evaluate web sources through so-called credibility signals.

CredLab (Credibility Lab) at Wellesley College, led by Computer Science Professor Eni Mustafaraj, is a research lab that studies web sources and credibility signals associated with them. As a member of CredLab, I have contributed on several branches of this multi-year project. The overarching goal of the project is to

1. Identify human-understandable signals that can be used to evaluate the credibility of a web source, and
2. Use AI tools and other computational methods to determine “values” of these signals for different web sources.

Misinformation lives under many appearances and avors, and in my research at CredLab, I looked into three different types of misinformation:

Science/health/medicine, or so called “pseudoscience”;

Problematic information targeting women and other minority groups;

Information/disinformation related to political discourse.

In the following sections, I will explain each sub project in detail.

The Observatory of Pseudoscience

I was introduced to the CredLab through a Summer Research program sponsored by the Wellesley College Science Center in 2017. Motivated by news stories of problematic search results in featured snippets (Pick 2015), we investigated their prevalence for science and health claims, known as pseudoscience.

To do this systematically, I built a SERP Observatory – an automated infrastructure that allows to monitor Google SERPs (Search Engine results pages) for various queries, including important page elements (e.g. featured snippet). Using the observatory, I recorded and monitored over time results that appear on the rst pages of Google Search for a prede ned set of false scienti c search phrases. The queries were compiled based on the claims that were debunked by the Snopes.com fact-checking website in Science and Medical topics. Then, two students and I assigned labels to sources that appeared among the top search results as either ”reliable”, ”pseudoscience”(promoting alternative scienti c claims), ”dubious” (fake news and conspiracy theory sites), or ”misc” (ex. platforms). The ultimate goal was to understand what categories of sites prevail in search results and how the rankings uctuate over time.

The Observatory of Pseudoscience gave interesting insights into the kinds of websites that are ranked as most relevant by Google for false scienti c claims. In particular, we saw that sometimes the top ranks include alternative medicine and pseudoscience websites that promote and con- rm claims that are already debunked by fact-checkers.

At the same time, for the bigger goal of credibility research, the Observatory of Pseudoscience was a case study that served as the rst interaction with the information ecosystem and some of its actors. Through the process of inspecting and labelling web sources from SERPs, we also got ideas for potential credibility signals (e.g., the presence and kind of advertisements hosted on the website). Lastly, the SERP Observatory turned out to be an extremely useful tool for data gathering and parts of its infrastructure continue being used by the CredLab members for other aspects of Credibility Signals research (Lurie and Mustafaraj 2019).

Exemplifying Gender Bias with Word Embeddings

Objectivity is considered an important criteria to consider when assessing the credibility of web sources of information (Metzger 2007). Objectivity in this context encompasses not only the extent to which presented information constitutes facts as opposed to opinions but also the understanding of intentions and agenda of the author/publisher. Hence, the source's position on certain topic(s) (or in other words, its bias), if known, might be a valuable credibility signal. However, identifying the bias of the source/publisher is not trivial. First, there are many different forms of bias: political, racial, age, gender. Further, bias is often implicit and can be disguised under the use of coded language.

Bolukbasi et al. (2016) in their *Debiasing Word Embeddings* paper displayed the extent to which word embedding, in particular the word2vec model, may amplify biases present in the corpus. Inspired by this research, we decided to investigate the possibility of exemplifying bias of the source using word embedding. If one trains a word2vec model on the text data from the source, can it serve as a tool to display various implicit and explicit biases of the source? In order to answer this question, the focus was narrowed down to gender bias. For the proof of concept model of such tool, I collected data from three exemplar sources: a feminist blog, a website from manosphere (i.e. a network of websites that focus on a new vision of masculinity[1], including movements like men's rights and Men Going Their Own Way), and Wikipedia's featured articles as control ("neutral") set. I trained word vector representations using Tomas Mikolov's word2vec with the continuous bag-of-words architecture for each of the sources (Mikolov et al. 2013). Experimenting with the three models showed that source's gender bias can be deduced from the words, whose vector representation is close to vectors of the keywords that are associated with gender (feminist, men, girl, etc.). Finally, in a poster at AAAI FLAIRS 2018 conference, I presented the proposed design of the tool as well as it usage scenarios. Even though only explored through the gender bias perspective, with appropriate indicator keywords, it can be extended to other forms of bias.

An important contribution of this research is that such tool would empower users to decide which aspect of bias to focus on. In addition, with the current surge in demand for transparent AI, the output produced by such tool (i.e. the words that are "close" to the bias-indicator keywords) can be used as an explanation for decision made about the bias signal.

Political Bias of News Sources

Another signal to evaluate credibility of a news source is its reputation: how it is perceived by others. There are many factors that establish reputation, but one of the first steps recommended when assessing the reputation of an unfamiliar source is to "google it" – a strategy also known as lateral reading (Caulfield 2017). User studies conducted by the CredLab indicate that the Knowledge Panel (KP)–the box on the right-hand side of SERP–play an important role in this assessment (Lurie and Mustafaraj 2018). In particular, references to political bias have been identified as particularly helpful to users (Rothshild, Lurie, and Mustafaraj 2019).

However, these reference are often extracted from the first paragraphs of corresponding Wikipedia pages. Observing SERPs over time (with the SERP Observatory I built) led to the discovery of frequent changes in the portrayal of news sources in KPs (knowledge panels). It turns out, these changes are often associated with repeated addition and removal of political labels (such as "alt-right", "far-left", etc.). In order to understand this phenomenon, we did an investigation of Wikipedia revisions. By obtaining all revisions for Wikipeda pages of 1300 news source, I used Google's diff-match-patch library to study changes in the text. The results indicate that Wikipedia pages for sources that are perceived as strongly biased (both on the right and left sides of political spectrum) often experience intense "political labelling" edit wars (Umarova and Mustafaraj 2019).

The importance of this finding is that demand for "political bias" label as a credibility signal turned Wikipedia space into the arena for the new kind of edit wars. Hence, when studying this information ecosystem, one needs to be aware of different actors and consequences of various attempts to polish or tarnish a new source's reputation.

Broader Impact

Research that I have engaged in over the past three years at CredLab contributes to the overall long-term goal of building AI that works on behalf of the people. In order to achieve this, the AI should not only support information tasks, but also increase transparency and trust in the information ecosystem. In a way, this echoes Tim Berners-Lee's vision of Semantic Web and intelligent agents (Berners-Lee et al. 2001) that work together with humans. Findings about signals that we make as part of various branches of this project, useful techniques that we identify for obtaining "values" of important signals, and data collection/parsing tools that we build move us closer towards the goal of creating a system that augments search results pages with useful signal information. Presenting such information before a user would empower them to make informed decisions about the information they consume daily.

References

Berners-Lee, T.; Hendler, J.; Lassila, O.; et al. 2001. The semantic web. *Scientific american* 284(5):28–37.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, 4349–4357.

Caulfield, M. 2017. *Web literacy for student fact-checkers*. Michael Arthur Caulfield.

Lurie, E., and Mustafaraj, E. 2018. Investigating the effects of google's search engine result page in evaluating the credibility of online news sources. In *Proceedings of the 10th ACM Conference on Web Science*, 107–116. ACM.

[1] https://rationalwiki.org/wiki/Manosphere

Lurie, E., and Mustafaraj, E. 2019. Opening up the black box: Auditing google's top stories algorithm. In *Proceedings of the... International Florida Artificial Intelligence Research Society Conference*, volume 32.

Metzger, M. J. 2007. Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology* 58(13):2078–2091.

Mikolov, T.; Chen, K.; Corrado, G.; Dean, J.; Sutskever, L.; and Zweig, G. 2013. word2vec. *URL https://code. google. com/p/word2vec*.

Pan, B.; Hembrooke, H.; Joachims, T.; Lorigo, L.; Gay, G.; and Granka, L. 2007. In google we trust: Users' decisions on rank, position, and relevance. *Journal of computer-mediated communication* 12(3):801–823.

Pick, R. 2015. Go ahead, ask google 'what happened to the dinosaurs'. *Vice. URL https://www.vice.com/en_us/article/pga4wg/go-ahead-ask-google-what-happened-to-the-dinosaurs*.

Rothshild, A.; Lurie, E.; and Mustafaraj, E. 2019. How the interplay of google and wikipedia affects perceptions of online news sources. In *Computation+ Journalism Symposium*.

Umarova, K., and Mustafaraj, E. 2019. How partisanship and perceived political bias affect wikipedia entries of news sources. In *Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA*, 1248–1253.