Using Baselines for Algorithm Audits

Jennifer A. Stark and Nicholas Diakopoulos
University of Maryland, College Park
{starkja, nad}@umd.edu

Abstract: Algorithm audits that assess bias often call for a suitable baseline to be defined. Bias becomes salient and newsworthy when put into contrast with a comparison, often an expectation derived by defining a baseline. This paper considers a baseline and three bias-rating methods for a study investigating bias in Google's image selection for presidential candidates' main search results page. Images of U.S. presidential candidates Hillary Clinton and Donald Trump were scraped from Google's image box on the main results page, and also from all results in Google Image search for the same time period, which served as the baseline. Using Google as its own baseline was sufficient to demonstrate that images on the main results page exhibit different distributions of news sources and sentiment, indicating that bias may have been introduced by the Google image selection algorithm.

Keywords: Algorithm Audits, Algorithmic Accountability, Baseline, Search Engines

Introduction

Reporting on bias in search engines is relatively new, and effective methodologies are still being established (Diakopoulos et al. 2018). One challenge is defining and obtaining baseline data appropriate to the specific study: assessments of bias must be put in context. The news value of an audit may depend on the deviation of some distribution with respect to that baseline. To date, studies have mostly not used baselines, opting instead to perform within-sample comparisons. For example, an exploratory study compared the emotions and age of women in news photographs with those of men, along with ratios of photographs of women and men across different news organisations (Kwak & An 2016) using data collected from the global-level news database GDELT (Global Data on Events, Location, and Tone), for a given time period. Another study compared images from specific queries between search engines Bing and Google, and between countries (Magno et al. 2016). Conversely, a study exploring stereotypical images of occupations used real-life counts from the Bureau of Labor Statistics as its baseline (Kay et al. 2015), enabling it to draw conclusions regarding stereotypes. While such comparisons are interesting and meaningful, isolating the role of a particular search algorithm in introducing bias into the most visible results requires that the comparison, or baseline, represent the universe of *potential* results surfaced by the search query studied.

We use this definition of baseline to investigate the visual framing of 2016 US presidential candidates Hillary Clinton and Donald Trump. Candidate images were collected from Google's main (non-personalized) search results page. Images were also collected from Google's Image search that served as our baseline. Search queries were "hillary clinton" and "donald trump".

We show that, compared with baseline images, the proportion of images of Clinton displayed on the main results page showing happiness was greater, while the proportion with a neutral expression was reduced. For Trump, percentages of happiness and anger were both increased compared with the baseline. Regarding political ideology, the cumulative proportion of left-leaning image sources was augmented in the image box for Clinton compared with the baseline, while the proportion from centrist sources was reduced. The proportion of Trump images from the centrist sources was reduced compared with baseline, while that from right-leaning sources was enhanced in the image box. The baselines of both candidates privileged left-leaning sources, overall. Together, this suggests that Google's main results page images are not curated in an ideologically balanced way.

Methods

We collected images from the main search result image box once per day from September 3rd to October 28th 2016 for the search queries "hillary clinton" and "donald trump", resulting in nine unique images for Clinton, and 11 for Trump. Baseline images were collected in early 2017 using Google Image search with advanced options specifying the same time period as for the image box collection, resulting in 353 images of Clinton, and 298 of Trump. Baseline images that were not of the queried candidate or that contained more than one face were removed. We

selected this baseline because Google was the universe we wanted to investigate; we assumed that images presented in the image box on the main results page would be gathered from Google's own indexed images, and therefore all potential images that *could* be presented in the image box would be found in Image Search. We acknowledge that images available in Google Images may themselves be biased in terms of selection, but evaluating the bias of Google Images itself (with respect to some other baseline) was beyond the scope of this study. All Google searches were non-personalized so that search history would not affect search results.

Emotion was determined for each unique image using the Microsoft emotion API set to a confidence threshold of ≥ 0.55 (considered 'likely') out of a total of 1.0 for an emotion to be declared as present. Sources were counted for each time the image occurred in a data-collection of the image box (i.e. once per day). All image sources were tagged with a bias rating based on a dataset aggregated, calculated, and provided by Allsides.com[1], combined with bias ratings from a study of news sharing on Facebook (Bakshy et al. 2015). Sources that remained unrated, or whose ratings from Allsides and the Facebook study disagreed, were assessed for bias by the first author according to the source's About pages, content of several articles not including that which the image linked to, and ratings from a third site MondoTimes.com. Code for all collection and analysis can be found on GitHub (https://github.com/comp-journalism/GoogleScraper, https://github.com/comp-journalism/Baseline_Problem_for_Algorithm_Audits).

Findings and Argument

We found that Clinton expressed emotions congruent with gender stereotypes for women in news photographs (Kwak & An 2016): that women smile more. However, this was not reflected in the baseline in which neutral expressions dominated instead (Figure 1 left). Trump had higher proportions of images showing happiness, anger, and slightly fewer showing no emotion compared with the baseline (Figure 1 right).

We found that sources of baseline images were mostly left-leaning with lower proportions from centrist and right-leaning sources for both candidates (Figure 2). Further, a chi-square test showed that the distribution of political ideology for sources of images in the image box is significantly different from those in the baseline (Clinton: χ = 68.1, $p < 0.001$; Trump: χ = 36.4, $p < 0.001$). Specifically, the proportion of left-learning sources of Clinton image box images was augmented at the expense of centrist sources. For Trump, left-learning sources remained highly represented in the image box, while right-leaning representation was increased at the expense of centrist. Of note, right-leaning sources in the image box were highly partisan, including pro-Trump sub-reddit r/The_Donald, pro-Russia web news site Russia Insider, and an obscure pro-Trump YouTube channel[2]. Finally, not all sources of image box images were represented in the baseline. Some of these sources were identified in Google Image search results after extending the Google Image search time-frame further back in time, while others – including r/The_Donald and Russia Insider remained absent, indicating that curation of images for the image box may have different temporal dependencies than we assumed when initially collecting Google Image search images. Another possibility is that images in Google Image search do not represent all Google's indexed images.

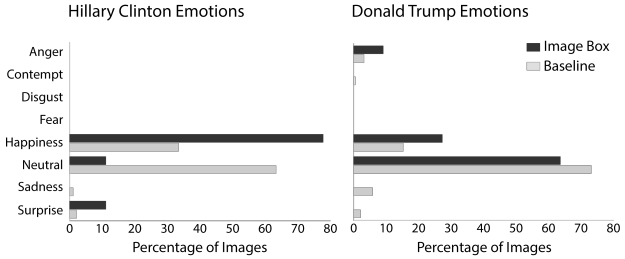

Figure 1 – Emotion disparities: Percentage of Clinton and Trump images from Image box, or Google Images (Baseline) that show listed emotions.

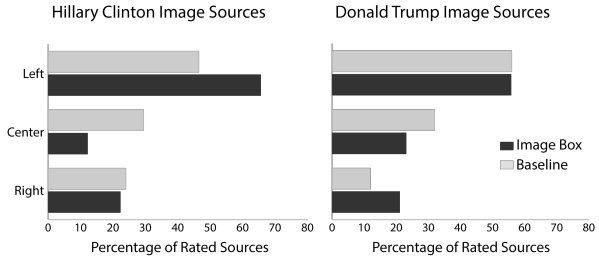

Figure 2 – Image Source Ideological Representation: Percentage of Clinton and Trump images from image box or Google Images baseline identified from each listed ideology.

[1] https://www.allsides.com/bias/about-bias
[2] https://www.youtube.com/watch?v=-dY77j6uBHI

Conclusions

We found that representation of candidates' emotions diverged from that seen in baseline images, whereby Clinton was predominantly happy, and Trump was happier and more angry compared with the baseline. We also found that the baseline images from Google Image search for Clinton and Trump were biased toward liberal sources. Sources of images on Google's main results page had a different political ideological distribution that suggests some degree of bias may have been introduced by the Google image selection algorithm. Finally, by incorporating a baseline into this study, we ensured that errors incurred by the emotion API resulting from its own training data are mitigated in part, since errors applied to the images of interest are also applied to the baseline images. Limitations include: a failure to disambiguate the influence of photographers' aesthetics and editors' selection criteria that may introduce biases into Google's image universe; we were not ultimately able to assign a bias rating to all sources, such as the stock image site Getty Images, foreign news outlets, and magazines with little to no news or political content (41 missing for Clinton; 23 for Trump), meaning that, were all sources rated, the final pattern of results may be different; and finally, that the chosen time-frame for the baseline to match that of the image-box image collection was insufficient to capture all the sources, and therefore did not represent the universe of potential images Google could select from to populate the main results page image box. We recommend that studies investigating bias in search engines consider and define baselines carefully and iteratively so as to ensure a reliable expectation to contrast with another sample.

Acknowledgements

This work has been supported by grants from the Tow Center for Digital Journalism and from the VisMedia project at the University of Bergen.

References

Bakshy, E., Messing, S. & Adamic, L.A., 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239), pp.1130–1132.

Diakopoulos, N. et al., 2018. I vote for – how search informs our choice of candidate. In M. Moore & D. Tambini, eds. *Digital Dominance: Implications and Risks.* Oxford University Press.

Kay, M., Matuszek, C. & Munson, S. a., 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, pp.3819–3828. Available at: http://dl.acm.org/citation.cfm?id=2702123.2702520.

Kwak, H. & An, J., 2016. Revealing the Hidden Patterns of News Photos: Analysis of Millions of News Photos Using GDELT and Deep Learning-based Vision APIs. In *The Workshops of the Tenth International AAAI Conference on Web and Social Media News and Public Opinion: Technical Report WS-16-18*. pp. 99–107.

Magno, G. et al., 2016. Stereotypes in Search Engine Results : Understanding The Role of Local and Global Factors. In *Workshop on Data and Algorithmic Transparency (DAT'16)*. Available at: http://datworkshop.org/papers/dat16-final35.pdf.