Bias as a Distinct Factor in Human Ratings of Machine Labeling

Eric P. S. Baumer
ericpsb@lehigh.edu
Lehigh University
Bethlehem, PA, USA

Amin Hosseiny Marani
amh418@lehigh.edu
Lehigh University
Bethlehem, PA, USA

ABSTRACT

This paper uses an excerpt from a larger analysis to argue that human assessments of machine labeling can reveal bias as a distinct measure separate from other perceptions of label quality. Human subjects were asked to assess the quality of automatically generated labels for a trained topic model. Quality assessments were gathered using 15 distinct self-report questions. Exploratory factor analysis identified a distinct "bias" factor. This point is likely relevant for a wide variety of machine labeling tasks.

KEYWORDS

Natural language processing, bias, survey

ACM Reference Format:
Eric P. S. Baumer and Amin Hosseiny Marani. 2020. Bias as a Distinct Factor in Human Ratings of Machine Labeling. In *Proceedings of CHI Workshop on Human-Centered Approaches to Fair and Responsible AI*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

CHI Workshop on Human-Centered Approaches to Fair and Responsible AI, April 26, 2020, Honolulu, HI, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

	Factors		
Item	Suitable	Perspectival	Nonsensical
Arbitrary	-0.02	-0.05	**0.87**
Biased	-0.08	**0.90**	-0.01
Coherent	0.77	0.04	-0.08
Confusing	-0.15	0.02	**0.69**
Expected	0.86	0.05	-0.05
Insightful	0.82	0.04	0.01
Meaningful	**0.92**	-0.03	0.02
Offensive	0.30	0.42	0.47
Sensible	**0.91**	-0.08	-0.03
Specificity	0.19	**0.57**	-0.01
Unpredictable	-0.31	0.13	0.49
Cum. Variance	0.40	0.55	0.69

Table 1: Loadings for each of the three factors on each of the 11 retained items from the original survey instrument. Values in bold indicate highest two loadings for each factor. Values in gray indicate which loadings fall below the 0.5 cut off. The bottom row indicates the cumulative proportion of variance in the original data accounted for by the three factors.

INTRODUCTION

Machine classification has, among other things, labeled images of African-Americans as "apes" [9], suggested an association between gay men and sex offenders [1], implied connections between Hispanic names and drug trafficking [18], and assessed African-Americans as being as higher risk of recidivism [2]. Such cases of bias go beyond the emerging metrics of fairness in AI [7], which often focus on error rates across different (protected) classes.

Such results can also influence (perceptions of) classifier performance. On one hand, some false positives (such as those above) may impact human perceptions of performance more than others. On the other, perceptions of a system's accuracy may be orthogonal to perceptions of its biases [19].

This paper considers these issues in one specific context: automatically generated labels for a trained topic model. In topic modeling [4, 5], a "topic," or latent theme within a corpus of documents, is often described as a probability distribution over a vocabulary words. Each topic is typically labelled with the "top N" high probability words for the topic. However, such labels can vary from informative to confusing. Thus, a variety of techniques attempt to automatically generate human-readable labels for topics [e.g., 3, 11, 14].

We use multi-item, multi-dimensional human assessments of these labels to show that the perceived bias of a label is a distinct component of perceptions about label quality.

TOPIC LABELING: SETTING AND RESULTS

This paper uses illustrative examples from a corpus of blog posts written by parents with children on the autism spectrum. These blogs were chosen in collaboration with a sociologists who has expertise in disability studies. The corpus spans from 2003 to 2015 and includes over 30,000 posts. A topic model [5] was trained using the R wrapper [15] for MALLET [13]. The model used 50 topics, based in part on coherence scores [10, 12], and in part on input from our collaborating sociologist. Labels were then generated for each topic using four different methods [6, 14].

Crowdworkers were asked to rate the quality of these labels. Workers were given a brief description of the corpus, shown excerpts from representative documents for one topic, and asked to describe the common theme of those documents in their own words. Workers then rated each label along 15 different criteria, for example, "Sensible: This label makes sense for these documents;" "Biased: The label indicates a limited perspective that favors one aspect or group;" or "Arbitrary: This label has no relation to these documents." Using exploratory factor analysis (EFA) revealed three distinct underlying dimensions within participants' ratings (see table 1 for factor loadings). The full analysis is, at the time of writing, under preparation in a separate submission.

This paper focuses on the second factor, which loads on the items Biased and Specificity. The wording for these two questions was "Biased: The label indicates a limited perspective that favors one

Label	Suitable	Perspectival
vaccines, david, offit, kirby, mccarthy, jenny	66.8	56.0
offit, kirby, mccarthy, jenny, vaccines, vaccine	54.1	55.5
doesn't, idea, david, offit, kirby, isn't	43.4	38.7
Respectful Insolance, Andrew Wakefield, Speak Dreams [...]	39.3	50.6

Table 2: Example labels and mean Suitable and Perspectival factor values from a topic related to autism and vaccines.

Label	Suitable	Perspectival
kids, play, ipad, toys, use, game	69.5	53.3
ipad, app, kids, apps, purple, toys	71.0	46.4
use, play, using, speech, ipad, stuff	58.1	38.8
Radiator Springs, Lightning McQueen, Brilliant Basics, [...]	21.0	37.4

Table 3: Example labels and mean Suitable and Perspectival factor values from a topic related to toys, games, and websites for children with special needs.

aspect or group," and "Specificity: People from a particular social group would agree with this labelling, while others would disagree." Intuitively, we can interpret this factor as *Perspectival*, capturing when a label indicates a limited, biased perspective of a topic.

Inspecting individual examples supports this interpretation. For example, Table 2 shows ratings of labels for a topic about autism and vaccines. Each of the four labels was generated by a different labeling technique. Crowdworkers rated the label "vaccines, david, offit, kirby, mccarthy, jenny" as the most Suitable. However, this label also receives the highest Perspectival ratings. The lowest Perspectival ratings are for the label "doesn't, idea, david, offit, kirby, isn't," does not mention vaccines.

Table 3 shows another example. This topic includes, in the words of one crowdworker, "documents [that] recommend toys for special needs children with disorders like autism." It is perhaps not surprising, then, that the labels "ipad, app, kids, apps, purple, toys" and "kids, play, ipad, toys, use, game" are rated most Suitable. At the same time, these two labels are also rated as having the most limited, biased perspective (i.e., the Perspectival factor). Inspecting the labels more closely, the fourth label comes from a technique that extracts noun bigrams as candidate labels. For this topic, this label includes only names of toys. The other label with a lower Perspectival rating, "use, play, using, speech, ipad, stuff," has two key differences. First, it places the term "ipad" lower in the list. Prior work has shown that relative ordering of words within a label can impact perceptions of label quality [17]. Second, it excludes the term "kids." Prior work has pointed to moral panics [8] around use of social media and other technologies, particularly by children [e.g., 16]. The implication from these labels, that kids are playing iPad games, may account for the high Perspectival ratings.

DISCUSSION

At the simplest level, these results show how human assessments of machine labeling can involve different dimensions of performance. To wit, the Perspectival factor, which indicates a biased perspective, has a weak correlation ($\rho = 0.15$) with the Suitable factor, which loads most strongly on the Sensible and Meaningful items.

Moreover, these results suggest a potential value in exploring human assessments of machine labeling results. Conducting similar analyses across a variety of different domains (text classification, image captioning, risk assessment, resume screening, etc.) may help identify more general latent factors in human assessment of machine labeling, both in terms of fairness and bias, as well as in more general terms of perceptions about performance.

ACKNOWLEDGMENTS

This material is based on work supported in part by the NSF under Grant No. IIS-1844901, and in part by a Seed Grant from Lehigh University's Data X Initiative.

REFERENCES

[1] Mike Ananny. 2011. The Curious Connection between Apps for Gay Men and Sex Offenders. *The Atlantic* (April 2011).

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *Pro Publica* (May 2016).

[3] Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2016. Automatic Labelling of Topics with Neural Embeddings. In *Proceedings of the International Conference on Computational Linguistics (COLING)*. Osaka, Japan, 953–963.

[4] David M. Blei. 2012. Probabilistic Topic Models. *Commun. ACM* 55, 4 (April 2012), 77–84. https://doi.org/10.1145/2133806.2133826

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

[6] Jason Chuang, Christopher D. Manning, and Jeffrey Heer. 2012. Termite: Visualization Techniques for Assessing Textual Topic Models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI)*. ACM, Capri Island, Italy, 74–77. https://doi.org/10.1145/2254556.2254572

[7] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv* (July 2018).

[8] Erich Goode and Nachman Ben-Yehuda. 1994. Moral Panics: Culture, Politics, and Social Construction. *Annual Review of Sociology* 20, 1 (1994), 149–171. https://doi.org/10.1146/annurev.so.20.080194.001053

[9] Alex Hern. 2015. Flickr Faces Complaints over 'offensive' Auto-Tagging for Photos. *The Guardian* (May 2015).

[10] Jey Han Lau and Timothy Baldwin. 2016. The Sensitivity of Topic Coherence Evaluation to Topic Cardinality. In *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Association for Computational Linguistics, San Diego, California, 483–487.

[11] Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. 2011. Automatic Labelling of Topic Models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, Portland, OR, 1536–1545.

[12] Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In *Proceedings of the Annual Meeting of the European Chapter of the Association for Computational Linguistics (EACL)*. Association for Computational Linguistics, Gothenburg, Sweden, 530–539.

[13] Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit.

[14] Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. 2007. Automatic Labeling of Multinomial Topic Models. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, San Jose, CA, 490–499.

[15] David Mimno. 2013. Mallet: A Wrapper around the Java Machine Learning Tool MALLET.

[16] Sarita Yardi Schoenebeck. 2014. Giving up Twitter for Lent: How and Why We Take Breaks from Social Media. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. Toronto, ON, 773–782. https://doi.org/10.1145/2556288.2556983

[17] Alison Smith, Tak Yeon Lee, Forough Poursabzi-Sangdeh, Jordan Boyd-Graber, Niklas Elmqvist, and Leah Findlater. 2017. Evaluating Visual Representations for Topic Understanding and Their Effects on Manually Generated Labels. *Transactions of the Association for Computational Linguistics* 5 (Jan. 2017), 1–15.

[18] Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan, IV, Mark D. M. Leiserson, and Adam Tauman Kalai. 2019. What Are the Biases in My Word Embedding?. In *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*. ACM, Honolulu, HI.

[19] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. ACM, Montréal, QC, 656:1–656:14. https://doi.org/10.1145/3173574.3174230