Social Bias in Elicited Natural Language Inferences

Rachel Rudinger*
Johns Hopkins University
rudinger@jhu.edu

Chandler May*
Johns Hopkins University
cjmay@jhu.edu

Benjamin Van Durme
Johns Hopkins University
vandurme@cs.jhu.edu

Abstract

We analyze the Stanford Natural Language Inference (SNLI) corpus in an investigation of bias and stereotyping in NLP data. The human-elicitation protocol employed in the construction of the SNLI makes it prone to amplifying bias and stereotypical associations, which we demonstrate statistically (using pointwise mutual information) and with qualitative examples.

1 Introduction

Since the statistical revolution in Artificial Intelligence (AI), it is standard in areas such as natural language processing and computer vision to train models on large amounts of empirical data. This "big data" approach popularly connotes objectivity; however, as a cultural, political, and economic phenomenon in addition to a technological one, big data carries subjective aspects (Crawford et al., 2014). The data mining process involves defining a target variable and evaluation criteria, collecting a dataset, selecting a manner in which to represent the data, and sometimes eliciting annotations: bias, whether or implicit or explicit, may be introduced in the performance of each of these tasks (Barocas and Selbst, 2016).

We focus on the problem of *overgeneralization*, in which a data mining model extrapolates excessively from observed patterns, leading to *bias confirmation* among the model's users (Hovy and Spruit, 2016). High-profile cases of overgeneralization in the public sphere abound (Crawford, 2013; Crawford, 2016; Barocas and Selbst, 2016).

Research on the measurement and correction of overgeneralization in NLP in particular is nascent. Stock word embeddings have been shown to exhibit gender bias, leading to proposed *debiasing* algorithms (Bolukbasi et al., 2016). Word embeddings have been shown to reproduce harmful implicit associations exhibited by human subjects in implicit association tests (Caliskan-Islam et al., 2016). Gender bias in sports journalism has been studied via language modeling, confirming that male athletes receive questions more focused on the game than female athletes (Fu et al., 2016). In guessing the gender, age, and education level of the authors of Tweets, crowdworkers found to exaggerate stereotypes (Carpenter et al., 2017).

A prerequisite to resolving the above issues is basic awareness among NLP researchers and practitioners of where systematic bias in datasets exists, and how it may arise. In service of this goal, we offer a case study of bias in the Stanford Natural Language Inference (SNLI) dataset. SNLI is a recent but popular NLP dataset for textual inference, the largest of its kind by two orders of magnitude, offering the potential to substantially advance research in Natural Language Understanding (NLU). We select this dataset because (1) we predict that natural language inference as a NLP task may be generally susceptible to emulating human cognitive biases like social stereotyping, and (2) we are interested in how eliciting written inferences from humans with minimal provided context may encourage stereotyped responses.

Using the statistical measure of pointwise mutual information along with qualitative examples, we demonstrate the existence of stereotypes of various forms in the elicited hypotheses of SNLI.

2 The SNLI Dataset

Bowman et al. (2015) introduce the Stanford Natural Language Inference corpus. The corpus was generated by presenting crowdworkers with

* denotes equal contribution.

Proceedings of the First Workshop on Ethics in Natural Language Processing, pages 74–79,
Valencia, Spain, April 4th, 2017.

a photo caption (but not the corresponding photo) from the Flickr30k corpus (Young et al., 2014) and instructing them to write a new alternate caption for the unseen photo under one of the following specifications: The new caption must either be [1] "definitely a true description of the photo," [2] "might be a true description of the photo," or [3] "definitely a false description of the photo." Thus, in the parlance of Natural Language Inference, the original caption and the newly elicited caption form a sentence pair consisting of a *premise* (the original caption) and a *hypothesis* (the newly elicited sentence). The pair is labeled with one of three entailment relation types (ENTAILMENT, NEUTRAL, or CONTRADICTION), corresponding to conditions [1–3] above. The dataset contains 570K such pairs in total.

Given the construction of this dataset, we identify two possible sources of social bias: **caption bias**,[1] already present in the premises from the Flickr30k corpus (van Miltenburg, 2016), and **(inference) elicitation bias**, resulting from the SNLI protocol of eliciting possible inferences from humans provided an image caption. Though we recognize these sources of bias may not be as tidy and independent as their names suggest, it is a useful conceptual shorthand: In this paper, we are primarily interested in detecting elicitation bias.

3 Methodology

We are ultimately concerned with the impact of a dataset's biases on the models and applications that are trained on it. To avoid dependence on a particular model or model family, we evaluate the SNLI dataset in a model-agnostic fashion using the pointwise mutual information (PMI) measure of association (Church and Hanks, 1990) and likelihood ratio tests of independence (Dunning, 1993) between lexical units.

Given categorical random variables W_1 and W_2 representing word occurrences in a corpus, for each word type (or bigram) w_1 in the range of W_1 and for each word type (or bigram) w_2 in the range of W_2, PMI is defined as

$$\text{PMI}(w_1, w_2) = \log \frac{P(W_1 = w_1, W_2 = w_2)}{P(W_1 = w_1)P(W_2 = w_2)}.$$

To compute PMI from corpus statistics, we plug in maximum-likelihood estimates of the joint and marginal probabilities:

$$\begin{aligned}\hat{P}(W_1 = w_1, W_2 = w_2) &= C(w_1, w_2)/C(*, *),\\ \hat{P}(W_1 = w_1) &= C(w_1, *)/C(*, *),\\ \hat{P}(W_2 = w_2) &= C(*, w_2)/C(*, *),\end{aligned}$$

where $C(w_1, w_2)$ represents the co-occurrence count of $W_1 = w_1$ and $W_2 = w_2$ in the corpus and $*$ denotes marginalization (summation) over the corresponding variable. We wish to focus on the bias introduced in the hypothesis elicitation process, so we count co-occurrences between words (or bigrams) w_1 in a premise and words (or bigrams) w_2 in a corresponding hypothesis.

For each pair of word types (or bigrams) w_1 and w_2, we can check the independence between the indicator variables $X_{w_1} = I_{\{W_1=w_1\}}$ and $Y_{w_2} = I_{\{W_2=w_2\}}$ with a likelihood ratio test. (Hereafter we omit subscripts w_1 and w_2 for ease of notation.) Denote the observed counts of X and Y over the corpus by $C'(x, y)$ for $x, y \in \{0, 1\}$.[2] The test statistic is

$$\Lambda(C') = \frac{\sum_{x,y} \left(\hat{P}(X = x)\hat{P}(Y = y)\right)^{C'(x,y)}}{\sum_{x,y} \hat{P}(X = x, Y = y)^{C'(x,y)}}.$$

where $\hat{P}$ is the maximum likelihood estimator (using C'), the summations range over $x, y \in \{0, 1\}$, and we have dropped the subscripts w_1 and w_2 for ease of notation. The quantity $-2 \log \Lambda(C')$ is χ^2-distributed with one degree of freedom, so we can use it to test rejection of the null hypothesis (independence between X and Y) for significance. That quantity is also equal to a factor of $2C'(*, *)$ times the mutual information between X and Y, and the PMI between W_1 and W_2 (on which X and Y are defined) is a (scaled) component of the mutual information. Noting this relationship between PMI (which we use to sort all candidate word pairs) and the likelihood ratio test statistic (which we use to test for independence of the top word

[1] Note that what we call caption bias may be due either to the Flickr30k caption writing procedure, or the underlying distribution of images themselves. Distilling these two sources of bias is outside the scope of this paper, as the SNLI corpus makes no direct use of the images themselves. Put another way, because SNLI annotators did not see images, the elicited hypotheses are independent of the Flickr images, conditioned on the premises.

[2] For example, note $C'(1, 1) = C(w_1, w_2)$ and $C'(1, 0) = C(w_1, *) - C(w_1, w_2)$; the other counts $C'(0, 1)$ and $C'(0, 0)$ can also be computed in this manner.

	GENDER		
woman	hairdresser‡ fairground grieving receptionist widow	**man**	rock-climbing videoing armband tatooes gent
women	actresses† husbands‡ womens‡ gossip‡ wemon‡	**men**	gypsies supervisors contractors mens‡ cds
girl	schoolgirl piata cindy pigtails‡ gril	**boy**	misbehaving see-saw timmy lad‡ sprained
girls	fifteen‡ slumber sking‡ jumprope† ballerinas‡	**boys**	giggle‡ youths‡ sons‡ brothers‡ skip
mother	kissed‡ parent‡ mom‡ feeds daughters	**father**	fathers‡ dad‡ sons† daughters plant
	AGE		
old	ferret‡ quilts‡ knits‡ grandpa‡ elderly‡	**young**	giggle cds youthful‡ tidal amusing
old woman	knits‡ grandmother‡ scarf† elderly‡ lady‡	**young woman**	salon† attractive blow blowing feeds
old man	ferret‡ grandpa‡ wrapping‡ grandfather‡ elderly‡	**young man**	boarder disabled rollerblades graduation skate‡
	RACE/ETHNICITY/NATIONALITY		
indian	indians‡ india‡ native‡ traditional‡ pouring†	**caucasian**	blond white‡ american asian blonde
indian woman	cooking† clothes lady using making	**american**	patriotic‡ canadian‡ americans‡ reenactment‡ america‡
indian man	food couple a‡ sleeping sitting	**american woman**	women‡ black white front her‡
asian	kimonos‡ asians‡ asain‡ oriental‡ chinatown‡	**american man**	speaking‡ money‡ black‡ white‡ music
asians	asian‡ food people‡ eating friends	**black woman**	african‡ american asian white‡ giving
asian woman	oriental‡ indian† chinese‡ listens† customers	**black man**	african‡ american white‡ roller face
asian man	shrimp† rice† chinese‡ businessman cooks†	**native american**	americans‡ music‡ dressed they woman
white woman	protesting‡ lady‡ looks women‡ was	**african american**	caucasian asian‡ speaking‡ black‡ white‡
white man	pancakes‡ caucasian‡ class black† concert	**african**	africans‡ africa‡ pots† receives† village†

Table 1: Top five words in hypothesis by PMI with specified words in premise, filtered to co-occurrences with a unigram with count at least five. Queries in bold. Significance of a likelihood ratio test for independence denoted by † ($\alpha = 0.01$) and ‡ ($\alpha = 0.001$).

pairs), we control for the family-wise error rate using the Holm-Bonferroni procedure (Holm, 1979) on all candidate word pairs. The procedure is applied separately within each view of the corpus that we analyze: the all–inference-type view, ENTAILMENT-only view, NEUTRAL-only view, and CONTRADICTION-only view.

The U.S. Equal Employment Opportunity Commission (EEOC) characterizes discrimination by type, where types of discrimination include age, disability, national origin, pregnancy, race/color, religion, and sex.[3] To test for the existence of harmful stereotypes in the SNLI dataset we pick words and bigrams used to describe people labeled as belonging to each of these categories, such as *Asian* or *woman*, and list the top five or ten co-occurrences with each of those query terms in the SNLI dataset, sorted by PMI.[4] We omit co-occurrences with a count of less than five. We include both broad and specific query words; for example, we include adjectives describing nationalities as well as those describing regions and races. We also include query bigrams describing people labeled as belonging to more than one category, such as *Asian woman*. Due to space constraints, we report a subset of the top-five lists exhibiting harmful stereotypes. The code and query list used in our analysis are available online, facilitating further analysis of the complete results.[5]

[3] https://www.eeoc.gov/laws/types/

[4] We use the provided Stanford tokenization of the SNLI dataset, converting all words to lowercase before counting co-occurrences.

[5] https://github.com/cjmay/snli-ethics

Preliminary results contained many bigrams in the top-five lists that overlapped with the query—exactly or by lemma—along with a stop word. To mitigate this redundancy we filter the query results to unigrams before sorting and truncating.

4 Results

We analyze bias in the SNLI dataset using both PMI as a statistical measure of association (Sec. 4.1) and with demonstrative examples (Sec. 4.2).

4.1 Top Associated Terms by PMI

For each social identifier of interest (for example, "woman," "man," "Asian," "African American," etc.) we query for the top 5 or 10 unigrams in the dataset that share the highest PMI with the identifier. In Table 1, the results are broken down by gender-, age-, and race/ethnicity/nationality-based query terms, though some query terms combine more than one type of identifier (for example, gender and race). Table 2 shows the results for the same gender-based queries run over different portions of SNLI, as partitioned by entailment type (ENTAILMENT, NEUTRAL, and CONTRADICTION). As described in Sec. 3, the pairwise counts used to estimate PMI are between a word in the premise and a word in the hypothesis; thus, query terms correspond with SNLI premises, and the results of the query correspond with hypotheses. A discussion of these results follows in Sec. 5.

ENTAILMENT	**women**	scarves† ladies‡ womens‡ wemon‡ females‡ woman‡ affection dressing chat smile†
	men	mens‡ guys‡ guitars cowboys† remove dock dudes workers‡ computers‡ boxers
	girls	cheerleaders‡ females‡ girl‡ dancers children‡ smile practice dance‡ outfits laughing
	boys	males‡ children‡ boy‡ kids‡ four‡ fighting† exercise play‡ pose fun
NEUTRAL	**women**	actresses‡ gossip‡ husbands‡ womens‡ nuns† bridesmaids† gossiping‡ ladies‡ strippers purses
	men	lumberjacks mens‡ supervisors thieves‡ homosexual roofers reminisce† contractors groomsmen engineers‡
	girls	fifteen‡ slumber† gymnasts‡ cheerleading‡ bikinis† sisters‡ cheerleaders‡ daughters‡ selfies† teenage‡
	boys	skip† sons‡ brothers‡ twins‡ muddy trunks† males† league‡ cards recess†
CONTRADICTION	**women**	womens† wemon bikinis‡ ladies‡ towels females‡ politics dresses‡ discussing men‡
	men	dudes mens‡ motel‡ gossip surfboards wives caps sailors floors helmets
	girls	sking‡ boys‡ 50 brothers sisters dolls† pose opposite phones hopscotch
	boys	girls‡ sisters‡ sons bunk homework† males coats beds† guns professional

Table 2: Top-ten words in hypothesis by PMI with gender-related query words in premise, filtered to co-occurrences with a unigram with count of at least five, sorted by inference type (ENTAILMENT, NEUTRAL, or CONTRADICTION). Queries in bold. Significance of a likelihood ratio test for independence denoted by † ($\alpha = 0.01$) and ‡ ($\alpha = 0.001$).

4.2 Qualitative Examples

Some forms of bias in a dataset may only be detectable with aggregate statistics such as PMI. Other, more explicit forms of bias may be apparent from individual data points. Here we present some example sentence pairs from SNLI that outwardly exhibit harmful stereotypes (labeled HS) or the use of pejorative language or slurs (labeled PL).[6] Note that in these examples, the identifiable biases have been introduced as a result of the SNLI inference elicitation protocol, that is, they arise in the hypothesis.

PREMISE: An African American man looking at some butchered meat that is hanging from a rack outside a building.
HYPOTHESIS (CONTRA.): A black man is in jail [HS]

PREMISE: New sport is being played to show appreciation to the kids who can not walk.
HYPOTHESIS (ENTAIL.): People are playing a sport in honor of crippled people. [PL]

PREMISE: Several people, including a shirtless man and a woman in purple shorts which say "P.I.N.K." on the back, are walking through a crowded outdoor area.
HYPOTHESIS (ENTAIL.): The woman is wearing slutty shorts. [PL]

PREMISE: adult with red boots and purse walking down the street next to a brink wall.
HYPOTHESIS (NEUTR.): A whore looking for clients. [PL, HS]

PREMISE: Several Muslim worshipers march towards Mecca.
HYPOTHESIS (NEUTR.): The Muslims are terrorists. [HS]

PREMISE: A man dressed as a woman and other people stand around tables with checkered tablecloths and a ladder.
HYPOTHESIS (NEUTR.): The man is a transvestite. [PL]

[6]The authors recognize the partially subjective nature of applying these labels.

Explicit introduction of harmful stereotypes or pejorative language by crowdworkers (such as that presented here) is a form of elicitation bias; it may be a result of many factors, including the crowdworker's personal experiences, cultural identities, native English dialect, political ideology, socioeconomic status, anonymity (and hence relative impunity), and lack of awareness of their responses' potential impact. As one reviewer suggested, in the case of CONTRADICTION elicitation, some crowdworkers may even have "viewed their role as being not just contradictory, but outrageously so." While these explanations are speculative, the harmful language and stereotypes observed in these examples are not.

5 Discussion of Results

From the top associated terms by PMI, as reported in Tables 1 and 2, the clearest stereotypical patterns emerge for gender categories. Stereotypical associations evoked for women (but not men) include: expectations of emotional labor (*smile*, *kissed*), "pink collar" jobs (*hairdresser*), sexualization and emphasis on physical appearance (*bikinis*), talkativeness (*gossip*, *gossiping*), and being defined in relation to men (*men*, *husbands*). Conversely, stereotypical views of men are also evoked: performance of physical labor (*cowboys*, *workers*), and professionals in technical jobs (*computers*, *engineers*).

Gender-based stereotypes in the corpus cut across age, as well. Girls are associated with particular sports (*ballerinas*, *cheerleaders*, *cheerleading*, *dance*, *gymnasts*), games and toys (*jumprope*, *dolls*), outward appearances (*pigtails*, *bikinis*), and activities (*slumber [parties]*, *selfies*). Boys, meanwhile, are stereotyped as troublemakers (*fighting*) and active outdoors (*recess*, *league*, *play*).

Though gender stereotypes appear in all three entailment categories in Table 2, those under the NEUTRAL label appear especially strong. We hypothesize this is a result of the less constrained nature of eliciting inferences that are neither "definitely true" nor "definitely false": Eliciting inferences that merely "might be true" may actually encourage stereotyped responses. Formally, neutral inferences may or may not be true, so those expressing stereotypes could be assumed to have no negative impact on the downstream model. However, if the model assumes neutral inferences are equally likely to be true or false *a priori*, that assumption's impact may be greater on minority groups subject to harmful negative stereotypes.

As represented by top-k PMI lists, individual terms for race, ethnicity, and nationality appear to have less strongly stereotyped associations than gender terms, but some biased associations are still observed. Words associated with Asians in this dataset, for example, appear to center around food and eating; the problematic term "Oriental" is also highly associated (another example of pejorative language, as discussed in Sec. 4.2). For many race, ethnicity, and nationality descriptors, some of the top-5 results by PMI are terms for *other* races, ethnicities, or nationalities. This is in large part a result of an apparent SNLI annotator tactic for CONTRADICTION examples: If the race, ethnicity, or nationality of a person in the premise is specified, simply replace it with a different one.

6 Conclusion

We used a simple and interpretable association measure, namely pointwise mutual information, to test the SNLI corpus for elicitation bias, noting that bias at the level of word co-occurrences is likely to lead to overgeneralization in a large family of downstream models. We found evidence that the elicited hypotheses introduced substantial gender stereotypes as well as varying degrees of racial, religious, and age-based stereotypes. We caution that our results do not imply the latter stereotypes are not present: rather, the prominence of gender stereotypes may be due to the relatively visual expression of gender, and the absence of other stereotypes in our results may be due to sparsity. We also note that our analysis reflects our own experiences, beliefs, and biases, inevitably influencing our results.

Future work may find more comprehensive evidence of stereotypes, including stereotypes of intersectional identities, by merging the counts of semantically related terms (or, conversely, by decoupling the counts of homonyms). It could also be fruitful to infer dependency parses and compute co-occurrences between dependency paths rather than individual words to facilitate interpretation of the results (Lin and Pantel, 2001; Chambers and Jurafsky, 2009), if sparsity can be controlled.

We have focused on the identities and accompanying biases present in the SNLI dataset, in particular those created in the hypothesis elicitation process; one complement to our study would measure the demographic bias in the corpus. Correlations introduced at any level in the data collection process—including real-world correlations present in the population—are subject to scrutiny, as they may be both creations and creators of structural inequality.

As artificial intelligence absorbs the world's collective knowledge with increasing efficiency and comprehension, our collective knowledge is in turn shaped by the outputs of artificial intelligence. It is thus imperative that we understand how the bias pervading our society is encoded in artificial intelligence. This work constitutes a first step toward understanding and accounting for the social bias present in natural language inference.

Acknowledgments

We are grateful to our many reviewers who offered both candid and thoughtful feedback.

This material is based upon work supported by the JHU Human Language Technology Center of Excellence (HLTCOE), DARPA LORELEI, and the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1232825. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA, the NSF, or the U.S. Government.

References

Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *California Law Review*, 104(3):671–732.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016.

Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4349–4357.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September. Association for Computational Linguistics.

Aylin Caliskan-Islam, Joanna J. Bryson, and Arvind Narayanan. 2016. Semantics derived automatically from language corpora necessarily contain human biases. Preprint, arXiv:1608.07187.

Jordan Carpenter, Daniel Preotiuc-Pietro, Lucie Flekova, Salvatore Giorgi, Courtney Hagan, Margaret L. Kern, Anneke E. K. Buffone, Lyle Ungar, and Martin E. P. Seligman. 2017. Real men dont say "cute": Using automatic language analysis to isolate inaccurate aspects of stereotypes. to appear.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610, Suntec, Singapore, August. Association for Computational Linguistics.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms mutual information, and lexicography. *Computational Linguistics, Volume 16, Number 1, March 1990*, 16(1).

Kate Crawford, Kate Miltner, and Mary L. Gray. 2014. Critiquing big data: Politics, ethics, epistemology. *International Journal of Communication*, 8:1663–1672.

Kate Crawford. 2013. Think again: Big data. http://atfp.co/2k9jaBT. Accessed 2017-01-26.

Kate Crawford. 2016. Artificial intelligence's white guy problem. http://nyti.ms/2jVLJUh. Accessed 2017-01-22.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, Special Issue on Using Large Corpora: I*, 19(1).

Liye Fu, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Tie-breaker: Using language models to quantify gender bias in sports journalism. In *Proceedings of the IJCAI workshop on NLP meets Journalism*.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany, August. Association for Computational Linguistics.

Dekang Lin and Patrick Pantel. 2001. Dirt – discovery of inference rules from text. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 323–328. ACM.

Emiel van Miltenburg. 2016. Stereotyping and bias in the Flickr30K dataset. In Jens Edlund, Dirk Heylen, and Patrizia Paggio, editors, *Proceedings of the Workshop on Multimodal Corpora (MMC-2016)*, pages 1–4, May.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations. *Transactions of the Association of Computational Linguistics*, 2:67–78.