Bias Detection in SNLI Dataset

Artidoro Pagnoni
apagnoni@cs.cmu.edu

February 13, 2020

1 Background

We seek to evaluate the presence of bias in a dataset. The presence of bias or social stereotypes in the data can trickle down to the models that are trained on the data with further cascading effects. It is therefore important to build tools to identify these issues. In this work we calculate the pointwise mutual information (PMI) between n-gram frequencies in the dataset and inspect the results identifying the presence of stereotypes. We apply this approach to the SNLI dataset [Bowman et al., 2015]. Combining training, evaluation, and test sets, the data contains 1140304 sentences, among which 652505 are unique.

All the code and results are available at: `https://github.com/artidoro/dataset_bias_detector`.

2 Results

We preprocess the data by lowercasing, tokenizing using Spacy[1], removing stop words (those that were not identity words provided by Rudinger et al. [2017]), and removing low frequency words. The question of removing low frequency words is delicate and we provide results for varying thresholds. We report pointwise mutual information:

$$PMI(w_i, w_j) = \log_2 \frac{P(w_i, w_j)}{P(w_i), P(w_j)} = \log_2 \frac{N \cdot c(w_i, w_j)}{c(w_i), c(w_j)}$$

Where we empirically evaluate $P(w_i)$ by $c(w_i)/N$ with N being the number of tokens in the corpus.

We report a table with the top 8 words in terms of PMI for 10 identity words in the list from Rudinger et al. [2017]. These identity words are often related with social stereotypes as we see in the results. Note that we are reporting two sets of results for unigrams. We are reporting the results when setting the minimum number of occurrences of words considered to 10 (Table 1), and when setting it to 100 (Table 2). The identity words in the two tables are not the same as some of them were filtered out because they were too infrequent.

Some words reported in Table 1 are very rare and appear to be questionably related to the identity word with which they have high PMI. This can be seen for example with "woman" and "eyeshadow" or "bagged". This is because PMI will return a high value for words that are infrequent but whose appearance is highly correlated with the given identity word. In this example, "eyeshadow" appears 12 times in the dataset, and 9 of those times with "woman". This leads to the highest PMI value over all words in the dataset for identity word "woman". When doing an audit of the bias of the dataset, it might be more useful to focus on words that appear more frequently. Detecting biased relations between high frequency words indicates a stronger bias in the dataset.

[1] `https://spacy.io`

Identity Words	Highest PMI words							
woman	eyeshadow	bagged	mascara	sari	veil	headscarf	weaves	refreshment
lesbian	figurine	adopted	attracted	experience	lover	shares	films	lovers
man	mustachioed	lanyard	beared	pac	thong	sideburns	refueling	bratwurst
gay	pride	rights	marriage	activists	canadians	attendees	springsteen	peruse
latino	checker	wildflowers	coolers	sowing	prepping	communicate	dart	superhero
mongolian	bbq	decals	employed	shines	exchange	employees	arrow	aprons
russian	prime	toga	tundra	spies	secretly	minister	womens	spy
muslims	terrorists	christians	channel	sponsored	celebrate	opening	1	phones
christians	praising	gospel	lord	muslims	impressed	pork	villagers	lobster

Table 1: Words with highest PMI for unigrams. Minimum number of occurrences is 10.

Identity Words	Highest PMI words							
woman	headscarf	purse	blouse	skirt	heels	gown	loom	attractive
women	bikinis	skirts	dresses	tops	cellphones	scarves	derby	knitting
man	mustache	bearded	balding	beard	tuxedo	shaves	turban	briefcase
men	guitars	suits	vests	caps	hats	microphones	laptops	shirts
boy	scouts	pinata	teenage	little	legos	trunks	pumpkin	pajamas
girl	pigtails	little	dolls	doll	leotard	boyfriend	pink	gymnastics
african	american	descent	village	hut	caucasian	traditional	pots	garb
chinese	dragon	year	written	buffet	menu	traditional	checkers	celebration
muslim	praying	turban	stall	knees	garb	peers	bringing	rows

Table 2: Words with highest PMI for unigrams. Minimum number of occurrences is 100.

3 Stereotypical Associations

We see strong evidence that the associations made in this dataset are stereotypical. We see in particular a strong **gender** bias with terms stereotypically relating to the female gender scoring high PMI with the the identity words. For example, words with high PMI with "woman", "women", "girl":

- Clothing: "bikinis", "skirt", "heels", "dress"
- Activity: "knitting", "gymnastics"
- Object: "purse", "doll"
- Appearance: "little", "attractive"

The same stereotypical associations can be seen along the dimension of **religion**. For example, the strongest association in terms of PMI for "muslims" is "terrorists". We also observe such biased associations with respect to **socio-economic** classes. We see that for both "african" and "muslim" the words "hut" and "stall" have high PMI. These words indicate the presence of stereotypical economic inequalities in the training examples involving these social groups. See the Appendix for examples of sentences where these stereotypical associations occur.

The presence of such biased associations is a significant issue in a dataset for natural language inference. The overall goal of the task is to achieve deeper semantic understanding and to verify the logical relation between sentences. The presence of social bias, or stereotypical associations, can lead the model to learn those biases instead of the true semantic or logical relation between the sentences. The priors learned by the model about the world described in this dataset might play a significant role in determining whether two sentence follow logically.

4 Mitigating the Bias

In general, we observe bias in slightly stronger bias in the hypothesis that are crowd sourced compared to the premises, and mostly in the neutral and contradicting examples. The original data, which was also crowdsourced, was created by asking to describe an image. The task of the crowdworker is then more deterministic with less room for social biases from the worker to creep in the data. The presence of bias is likely to come from the images themselves, but that can be controlled more easily by the designer of the experiment.

Regarding SNLI, the following part of the instructions to the crowd workers was particularly problematic in my opinion: "Using only the caption and what you know about the world". Saying that the worker should use their knowledge of the world already primes them towards using their stereotypical priors about what is generally assumed to be true. It fails to convey that the crowd worker should only use common sense and logic beyond the given text.

To mitigate this problem, and to improve the quality of the dataset it is necessary to shift the challenge towards real semantics and logic instead of allowing shortcuts that rely on social biases. As previously done in the visual captioning domain, it could be useful to request in a large portion of the data collection (possibly half) to use describe a non stereotypical setting. The task could be phrased in an adversarial manner inviting the worker to make improbable examples, that are logically correct and plausible, but very unlikely based on gender, social, and economic stereotypes. Ideally a second pass of workers would review the examples to ensure that it is truly more difficult. In this case I believe that giving more context could actually lead to less stereotypical responses.

5 Advanced Analysis

For the advanced analysis we generalize to bigrams. Results are summarized in Table 3. We see very stereotypical associations arising at the bigram level, even more so than at the unigram level. This partly due to the fact that combining identity words allows to restrict the domain to social groups that tend to suffer from strong biases. We see here "hispanic man" is associated with "tired looking", and "making pot". Similarly, "african boy" is associated with "sits dirt" surfacing the stereotype of poverty in Africa.

Identity Words	**Highest PMI biagrams**				
hispanic man	tired looking	his smartphone	making pot	painting large	large basket
hispanic women	moving legs	women moving	cart grocery	women bright	women children
african boy	sits dirt	young african	round object	inside old	chair inside
african woman	multiple bags	near hut	traditional african	looked her	children hang

Table 3: Words with highest PMI for bigrams. Minimum number of occurrences is 10.

6 Implementation Details

The tool that I implemented calculates pointwise mutual information (PMI) between n-gram frequencies in a dataset in jsonl format. To optimize performance, the single and joint occurrence counts of n-grams can be stored after preprocessing. When querying the PMI between an n-gram and the n-grams in the dataset the preprocessed counts can be loaded without needed retraining. This optimizes the time of querying the PMI for a large number of n-grams.
The implementation also uses multiprocessing to handle large datasets. With my machine (6 cores) single and joint occurrence counts for both unigrams and biagrams on the entire SNLI dataset can be calculated in

less than one minute. Loading the previously elaborated counts and querying the PMI of a single word takes less than a second.

7 Appendix

Examples of sterotypical associations in the data:

- Several Muslim worshipers march towards Mecca. | Muslims are terrorists.
- Little boy playing with his toy truck. | A girl plays with a doll.
- Woman in pink shirt and black shorts, holding a blue shovel, shoveling white snow. | The woman is young and attractive.
- Two little girls wearing pink dresses and sandals | Two little girls wearing bright pink dresses and sandals

References

S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642, 2015. doi: 10.18653/v1/d15-1075. URL `https://doi.org/10.18653/v1/d15-1075`.

R. Rudinger, C. May, and B. V. Durme. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, EthNLP@EACL, Valencia, Spain, April 4, 2017*, pages 74–79, 2017. doi: 10.18653/v1/w17-1609. URL `https://doi.org/10.18653/v1/w17-1609`.