Bias Audit of the Crowdsourced Stanford Natural Language Inference (SNLI) Dataset

Athiya Deviyani
adeviyan@cs.cmu.edu

Abstract

This paper will discuss the results of performing a bias audit of an Natural Language Processing (NLP) dataset that was produced through crowdsourcing. We will computationally measure and explore the social stereotypes that exist within the resulting dataset and discuss the potential ethical implications when using crowdsourced datasets to train models for other NLP tasks.

1 Introduction

Crowdsourcing is a popular method to generate natural language processing (NLP) datasets due to its low cost and scalability. Unfortunately, humans are especially vulnerable to subjectivity and bias. In this paper, we will perform a thorough bias audit of the Stanford Natural Language Inference (SNLI) corpus that was originally published by Bowman et al. (2015). The SNLI corpus contains 570K human-written English sentence pairs manually labeled for balanced classification with the labels entailment, contradiction, and neutral. The data collection was done through Amazon Mechanical Turk, where crowdworkers are asked to generate an alternate description (hypothesis) of a particular caption (premise) that has the entailment, contradiction, and neutral relationship. The captions were retrieved from the Flickr30k dataset van Miltenburg (2016).

To perform the audit, we will use a quantitative metric called the Pointwise Mutual Information (PMI). PMI is often used as a word-association measure within a corpus as it also takes into account whether or not their co-occurence is as expected due to their frequencies. We will then use PMI to measure which words within the SNLI corpus co-occur often with the identity labels used by Rudinger et al. (2017). From this, we will be able to see the identity-word associations that may exhibit and perpetuate harmful stereotypes. We will also extend the PMI analysis to higher-order n-grams for the identity labels to explore the distinction between the stereotypes of different subclasses within an identity group. Finally, we will present the quantitative results as well as qualitatively inspect several premise-hypothesis pairs which outwardly shows the presence of biases within the crowdworkers.

2 Experiments

2.1 Baseline

First and foremost, we have done basic preprocessing to rid the text of noise, such as removing duplicate entries, lowercasing the sentences and using NLTK's `RegexpTokenizer` to remove punctuation and tokenize the words. We have also removed any duplicate words as we are only interested in incrementing the count of a word based on its existence in the sentence.

For the baseline analysis, the quantitative metric that we will use to audit the datasets is the pointwise mutual information (PMI), which is formally defined as:

$$PMI(w_i, w_j) = log_2 \frac{N \cdot c(w_i, w_j)}{c(w_i)c(w_j)} \quad (1)$$

In the above formulation, N is the length of the corpus, $c(w_i)$ denotes the count of word w_i in a corpus, whereas $c(w_i, w_j)$ denotes the number of times words w_i and w_j occur together in a corpus. In our case, the premises and hypotheses are the corpora and each sentence is treated as an individual document. If the words occur or co-occur more than once within a document, the count will be denoted as 1. We have also removed words that occur less than 10 times to increase the efficiency of the experiments. Then, we calculated the PMI for each identity-word pair and plotted the results as a bar graph for evaluation.

2.2 Advanced analysis

For the advanced analysis, we have extended the baseline PMI analysis to higher-order n-grams, similar to the audit done by Rudinger et al. (2017). We will focus primarily on concatenating identity labels to obtain bigrams, such as "Asian man" or "Asian woman" to explore whether the stereotypes present in the unigrams extend to the different subgroups within a class. Furthermore, we have noticed from preliminary analysis that some of the highest PMIs of the identity labels of interest, such as "black" and "white", are extremely noisy as they are used in contexts other than race (e.g. "black shoes", "white shirt"). Therefore, we believe that using bigrams such as "black people" and "white men" will yield more fruitful results. To do this, we counted the co-occurence of an identity phrase-word pair. We have kept the experimental setup identical to the baseline.

3 Analysis of results

3.1 Word association analysis

A cursory glance at Table 1 can immediately point out several social biases that exist not only within the crowdsourced descriptions (hypothesis), but also in the original captions (premise). This is particularly worrying as the presence of implicit biases within a premise is more likely to elicit subjective hypotheses.

It is prominent that there exists a distinction between the perception of men and women. We can see that most related to the female identity are related to clothing (*bikinis*, *skirts*) and physical appearance (*beautiful*, *attractive*), which is not particularly apparent in its male counterpart. The nature of a woman's clothing is often extensively described (*skimpy*, *revealing*).

Additionally, stereotypes related to occupation also occur within the gender category. Women are often expected to complete household tasks (*sowing*) or do more hospitality-related (*barista*) and creative professions (*cheerleader*, *entertainer*). On the other hand, men often venture into more leadership (*chief*) and scientific (*surgeon*, *doctor*, *dentist*) roles. These professional stereotypes also extend to other identity labels such as race, where most *surgeons* and *captains* are Caucasian. This observation is also noted by Bolukbasi et al. (2016) during an attempt to de-bias word embeddings.

It is also noticeable that the crowdsourced descriptions enforce more stereotypes. The harmful biases indicates a view where positive traits are more commonly associated with the majority class while negative traits are associated with minority classes: Asians are *knowledgeable* and Caucasians are *handsome* while Africans are living in *poverty*, White people are *rich* while black people *rob*, Christians are *family*-oriented people while Muslims are *terrorists*. In some identity labels, both majority and minority classes are strongly associated with a negative stereotype, such as how the elderly are often *disabled*, teenagers are *stupid*, old men are *creepy* and most *scantily*-dressed young women are *attractive*.

3.2 Qualitative analysis

From Table 1, we have pinpointed some words in the list of top five PMIs within the hypotheses for some of the identity labels which may exhibit harmful stereotypes. In this section, we will evaluate them further by observing the corresponding premise. It is important to note that all of the premise-hypotheses pairs have the NEUTRAL label, which means they are supposed to serve as alternate descriptions that are *most likely to be true* about the given premise.

The following examples shows a strong gender bias within the crowdsourced annotations, as it denotes that a group of women "talking" indicates that they are "gossiping" yet a group of men "discuss". The high PMI values for both terms indicate that women is often talk about something trivial while men do not.

> PREMISE: Two older women are talking outside.
> HYP.: Two women are **gossiping** about their next door neighbors.

> PREMISE: Two gentlemen in green scrubs are talking to a third man on the street in front of an eatery.
> HYP.: Three men **discuss** medical business outside.

One of the most jarring terms that we found is how a seemingly mundane premise describing two Asians standing on the street elicited the word "prostitute". Additionally, a caption describing a lady dressed in black in a poverty stricken area corresponds to a hypothesis which contains the description of the lady being African, and reworded "poverty stricken area" as "ghetto".

> PREMISE: Asian friends standing in the street in busy part of town.
> HYP.: The Asian friends are **prostitutes**.

IDENTITY	PREMISE	HYPOTHESIS
women	bikinis, knits, headscarves, skirts, kimonos	husbands, bikinis, **gossip**, gossiping, sews
young woman*	hip, fancy, party, shopping, dancing	attractive, Caucasian, scantily, beautiful, husband
old woman*	weaves, knitting, blood, cane, comforting	robber, shading, obscene, sowing, knits
men	hell, cannon, suits, ladders, turbans	wives, cigars, **discuss**, gutters, harnesses
young man*	bust, business, suit, computer, instructing	thoughts, instructing, protests, braves, lectures
old man*	reclining, barren, wit, jetty, carves	coffin, **creepy**, grumpy, inappropriate, frown
female	cheerleader, leotard, acrobat, entertainer, dancer	tanned, barista, tummy, bra, revealing
male	Caucasian, trap, surgeon, dentist, chief	Invention, stripper, doctor, chief, agents
Black	poodle, graphic, fedora, trousers, beak	bets, labs, tights, Friday, ghetto
Black people*	dilapidated, shooting, fenced, gathered, fill	**rioting**, robbed, shades, **weapon**, crowds
White	footballer, sports, dribbling, undershirt, powder	rodent, frosting, bottoms, refreshments, hippy
White people*	horses, laughing, flags, building, speaking	sail, ponies, cruise, yacht, rich
Asian(s)	studying, schoolgirl, explains, wars, engage	**prostitutes**, skimpy, surgical, knowledge, elders
Asian woman*	fabrics, cloths, silk, beautifully, toned	breast, **sexy**, photoshoot, attractive, skimpy
Asian man*	diners, notes, badges, health, explains	blazer, **literature**, dinning, graduated, caring
African(s)	tribe, huts, villagers, warrior, Americans	**ghetto**, poverty, American, starving, weapons
Caucasian	handsome, surgeon, explains, captain, pub	spar, handsome, defending, taught, heavyset
Christian(s)	hell, ironing, Jesus, praying, church	hell, presents, praying, family, Christmas
Muslim(s)	desolate, hijab, Islamic, loud, sands	**terrorists**, body, celebrate, marching, riding
Muslim woman*	loud, desolate, headscarves, sands, desert	dead, cold, aloud, headed, market
Muslim man*	powder, son, khakis, scarf, mother	**weapon**, temple, son, bus, next
elderly	canes, handicapped, dentist, crutches, visit	graveyard, cavity, coma, disabled, retirement
teenagers	study, skips, partying, mingle, socializing	streaking, loiter, violence, **stupid**, rave

Table 1: Top five words in the SNLI corpus hypotheses by PMI for each identity label or phrase (*). Words that are to be analyzed further for the potential presence of social bias are emboldened.

PREMISE: A poverty stricken housing area with a lady in black with white headdress in the distance.
HYP.: An African woman is standing in a **ghetto** neighborhood waiting for her husband.

The following example shows how a premise that describes Muslim worshipers marching towards Mecca, a place of worship, elicits an alternate description that portray Muslims as terrorists. The harmful bias in this example is made obvious by the fact that there were no additional information depicting any form of violence.

PREMISE: Several Muslim worshipers march towards Mecca.
HYP.: The Muslims are **terrorists**.

The following example supports the stereotype that youths are clueless and unwise. The premise depicts five excited teenagers, yet the corresponding hypothesis makes an assumption that whatever they are excited about *must* be something stupid.

PREMISE: Three boy teenagers and two female teenagers are making excited facial expressions.
HYP: Five teenagers are excited about something **stupid**.

Next, we will look at some examples produced through obtaining the PMIs of identity *phrase*-word pairs from our advanced analysis. The following example suggests the violent nature of the identity phrase "black people" as the hypothesis adds the existence of a weapon while the original premise does not.

PREMISE: A crowd of black people are gathered and one person has a backpack on.
HYP.: A person has a **weapon** in a backpack in a crowd of black people.

The examples below show an interesting distinction between the phrases "Asian man" and "Asian woman". The premise depicting an "Asian man" pointing at an adult magazine (Playboy) sign describes the magazine as "Western literature". Conversely, the premise containing "Asian woman" posing on the floor implies that they are doing a sexy photoshoot. This shows a bias where Asian men are de-sexualized and depicted as scholarly while Asian women are overly sexualized.

PREMISE: Asian women sit on the floor and are working on a project while two of them smile at the camera.
HYP.: Asian woman at the sexy **photoshoot**.

PREMISE: Asian man pointing at a Playboy brand sign.
HYP.: An Asian man is pointing to his favorite piece of western **literature**.

We have observed that by exploring the various subclasses within an identity label (bigram of identity-identity) may either enforce or suppress a stereotype that was apparent in the unigram. In a previous example, we have observed a biased generalization that Muslims are terrorists. However in the example below, it appears that the generalization only extends to Muslim men. Contrarily,

Muslim women are often viewed as victims, with words such as "dead" having a high PMI.

> PREMISE: A man who has his face covered with a turban is carrying a weapon.
> HYP.: The Muslim man has a **weapon**.

The final example that we found worth evaluating is rather different than the preceding ones, as it has a label of CONTRADICTION. As noted in Rudinger et al. (2017), when a crowdworker is trying to come up with a sentence that is contradictory to the premise, they will come up with one that is outrageously so. This might lead to harmful language and stereotypes, which is observed in the example below:

> PREMISE: Group of colored people walking the street.
> HYP.: A group of black people are **rioting**.

3.3 Crowdsourcing set-up

The SNLI corpus was collected through Amazon Mechanical Turk. An image caption from the Flickr30k corpus collected by van Miltenburg (2016) was presented to the crowdworkers without the image itself, and they are asked to come up with alternate caption that is definitely true (ENTAILMENT), might be true (NEUTRAL), and definitely false (CONTRADICTION). They should only use the caption and what they know about the world to derive the alternate descriptions.

It is observed that the stereotypes are particularly noticeable in gender-related topics. For example, any premises that contain the word "woman" is shown to elicit additional physical descriptions such as "attractive" or "skimpy" that are not originally present. Furthermore, there are a lot of prominent biases across all race/ethnicity/nationality groups, with most non-Caucasian races suffering from negative stereotypes, such as how "African" is associated with poverty and "Black" with violence. This can be caused by an imbalanced racial distribution among the crowdworkers or by the lack of general knowledge about the minority races, such that they can only infer from stereotypes.

One way to potentially mitigate the generation of subjective hypotheses is to specifically instruct the crowdworkers not to include information that has no direct inference from the premise. For example, when given a premise "a crowd of black people", some crowdworkers might infer that a weapon is involved. However, by imposing the "direct inference" rule, the crowdworkers will instead come up with more mundane hypotheses such as "a crowd of black people standing."

Additionally, giving the crowdworkers the corresponding picture to the caption may reduce the production of unwarranted and harmful stereotypes, as it constrains the user to a particular visual space. For example, if a weapon is not seen in the picture, then it would be unreasonable for a crowdworker to come up with a hypothesis that indicates the presence of violence.

Finally, in a previous section, we have observed that social biases also exist within the premises. It is possible to perform a bias audit of just the premises themselves and identify captions that are potentially harmful. From this, the authors of the crowdsourcing experiments can remove the highly-biased captions or artificially generate new examples to combat the existing stereotypes. This would prevent inducing biased hypotheses from the crowdworkers.

4 Conclusion

From the results and discussion above, it is prominent that the SNLI corpus contains a certain level of social bias that may be harmful to particular groups. Hence, using the dataset to train models for other NLP tasks will most certainly propagate these biases. Thus, future work should aim to explore methods to mitigate biases obtained from various parts of the data collection pipeline, specifically during the collection of crowdsourced data.

References

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Emiel van Miltenburg. 2016. Stereotyping and bias in the flickr30k dataset.