CS4248 Assignment 3
A0184679H

1A. Declaration of Original Work: By entering my student ID below, I certify that I completed my assignment independently of all others (except where sanctioned during in-class sessions), obeying the class policy outlined in the introductory lecture. In particular, I am allowed to discuss the problems and solutions in this assignment, but have waited at least 30 minutes by doing other activities unrelated to class before attempting to complete or modify my answers as per the Pokemon Go rule.

Signed,
A0184679H

2. References: I give credit where credit is due. I understand need not (but am encouraged to) reference papers already mentioned in this assignment specification. I acknowledged that I used the following websites or contacts to complete this assignment:

- Rundinger, R., May, C. & Van Durme, B., Social bias in elicited natural language inferences, for choosing my academic paper as my central discussion
- Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., Kalai, A. T., Man is to computer programmer as woman is to homemaker? Debiasing word embeddings, to read up on the debiasing algorithm
- Larson, B., Gender as a variable in natural-language processing, to read up on ethical frameworks when using gender as a variable in NLP

The Stanford Natural Language Inference (SNLI) corpus is introduced by Bowman et al. in 2015, and the corpus is generated by presenting US crowdworkers with a caption describing a photo but without the actual photo itself, and the crowdworkers are required to write a new photo caption that either be a true description of the photo, a “might-be-true” description of the photo, or a false description of the photo (Rundinger et al, 2017), and the corpus consists of 570,000 such pair of original caption and rewrote caption datasets. The three rewritten categories are labelled with “entailment”, “neutral” and “contradiction” respectively. An academic paper by Rundinger et al. describe an experiment conducted by them to determine social bias of natural language inference in the SNLI dataset.

Rundinger used a pointwise mutual information (PMI) approach, and aggregate statistics method with labelling of unigrams and bigrams used to describe people to determine stereotypes in the corpus. For example, labels like “women” maps to words like “scarves”, “wemon”, “affection”; “men” maps to “supervisors”, “engineers”, “computers”; “girls” map to “bikinis”; “boys” maps to “homework”. The results showed that gender categories have the clearest stereotypical patterns, and other categories include age, race, nationality, and ethnicity (Rundinger et al., 2017). For example, “asians” is associated with food, although the association is not as strong as gender categories. Therefore, it is evident that discriminations and biases exist in the datasets from the research.

Biases have existed in languages for decades, and certain sides are bound to be harmed from these biases. For example, these biases exist in online algorithms of various user applications and systems, such as recommendation systems and advertising platforms (Bolukbasi et al., 2016). This will indirectly create confusion or even annoyance to application users. Furthermore, certain races are selected as the output of algorithms when predicting repeating offenders (Bolukbasi et. al, 2016). Biases in demographics also may cause implications such as languages used by minority groups on social media might not be processed by natural language tools that are trained on languages used by majority groups. Another unexpected harm caused by bias is in the generating of datasets such as the SNLI, where crowdworkers view their roles to be contradictory when attempting to generate datasets that are labelled with “contradiction” (Rundinger et al.,

2017). This is possibly caused by various factors such as crowdworkers' personal experiences, cultural English dialect, and socioeconomic statuses.

To mitigate the bias, several approaches have been proposed. Bolukbasi et al. developed a debiasing algorithm in word embeddings: a set of gender-neutral words that may cause bias such as gender bias, racial bias or religious bias is prepared. The second step, identifying gender subspace, determines the direction of the embedding that captures the bias, and the third step, neutralize and equalize, creates equal distances between neutral words and all words in the equality set (Bolukbasi et al., 2016). After applying the debiasing algorithm, gender stereotypes decreased from 19% to 6%.

Another way to mitigate this bias is the prevention or caution in using categories that are prone to bias as a classification label, such as gender. Using gender as a classification label may be an ethical issue, and frameworks and guidelines should be applied when using gender as a variable in NLP experiments (Larson, 2017). The four guidelines when using gender include making theories of gender explicit by offering a definition and some discussions about its concept; avoid using gender as a variable unless necessary; stating explicit methods of assigning gender categories to research participants; respect the difficulty of respondents in answering self-identification gender questions (Larson, 2017).

After comparing the scenario between Rundinger's academic research and my NLP's project, I discovered that both projects share a common ethical issue in NLP: exclusion and overgeneralization. My project on reference string parsing using regular expressions classifies authors' names, in particular, first and last names by labelling the last name as the family name and the first name as the given name. However, although this technique works on western names, I failed to recognize the difference in non-western names, for example, Chinese names may have their family name before their given name in the Eastern context and vice versa in the Western context.

As NLP techniques continue to develop nowadays, it is difficult to avoid bias in datasets due to overgeneralization and exclusion. Therefore, it is important for researchers and programmers who train their machine learning models to identify and mitigate these social biases such that the impact of biases in machine learning is kept to a minimum.