Assignment 5
CSE 517: Natural Language Processing - University of Washington
Winter 2022

Solution 1.
In this problem, we are required to conceal the author's gender while retaining the useful information in the data. To solve this problem, I first loaded the original dataset and tested the classify.py model to evaluate how effectively it predicts the author's gender and the subreddit for a post without obfuscating the gender information. The findings showed that the gender classification accuracy using the original dataset was 0.6595, while the subreddit post accuracy is 0.8585. This means that the algorithm can accurately predict the gender 65 percent of the time, and our goal in this project is to minimize this accuracy while ensuring that the subreddit post prediction is accurate.

Model 1: The first model I constructed by making two dictionaries, one for male words and one for female words (data provided in the assignment). Then I wrote a function called 'anonymize,' which takes two arguments: words and gender. Next, I went through each term that appeared in a subreddit post by male gender and checked to see if it also appeared in the male dictionary. If the word occurred in the male dictionary, those words are randomly replaced from the female vocabulary using 'random.randint', and the result is kept appended for each post and recorded in a'res' list I constructed. Finally, I duplicated the original dataset and replaced the 'post text' with the 'res' I obtained. I did the same thing when the sub-reddit post was authored by a woman and replaced those terms with words from the male lexicon. For model 1, a small code snippet is provided below.

```
def anonymize(words, gender):
    res = []
    if gender == "M":
        for word in words:
            if word in dict_male:
                r_index = random.randint(1, 3000) - 1
                res.append(list_female[r_index])
            else:
                res.append(word)
    else:
        for word in words:
            if word in dict_female:
                r_index = random.randint(1, 3000) - 1
                res.append(list_male[r_index])
            else:
                res.append(word)
    return res
```

I did not preprocess the original dataset and used it exactly as is in my anonymize function. I was able to reduce the gender categorization accuracy to 0.289, however the sub-reddit post accuracy has also been reduced to 0.517. We achieved lower subreddit accuracy because the terms were chosen at random from the male and female dictionaries, making the post unclear. Here's an example of a post from a Model 1 post:
Qualitative example model 1*: [I tiramisu toxins how progr sick mm outdoors r sashmi, instagram ombr√É¬£√Ç¬© feeling fit o staurdays super u m guava r aback]*

Model 2: I used glove pre-trained embedding to find semantically similar words by measuring the cosine distance between each word in a subreddit post and the words that appeared in the male and female dictionary. I used glove because I already used it in my previous homework and wanted to minimize computation time.
In this model, I defined two functions: 'closest word in female' and 'closest word in male'. These two functions use glove to calculate the cosine distance between the word in the subreddit post and the word in the male and female dictionaries. Then calculating the shortest distance and the closest word that corresponds to that. I also defined the 'anonymize2' function, which is similar to Model 1 except that it replaces the word with the cosine distance (closest word) and uses the same words that are not in the dictionary. Sample code snippet shown below to find the closest word:

```
def closest_word_in_female(word):
    closest_word = ""
    min_distance = 100000
    for w in list_female:
        dis = torch.norm(glove[word] - glove[w])
        if(dis < min_distance):
            min_distance = dis
            closest_word = w
    return closest_word
```

With this model 2, as expected my accuracy for gender prediction drastically reduced to 0.082 and subreddit post prediction accuracy improved when compared to the previous model to 0.6525. Below is the qualitative example from model 2:
Qualitative example model 2: [*Well they keep the your r clothes on …so theres that]*

Model 3: In this model, I tried to retain the model the same as previously with a little alteration by considering the length of words in the sub-reddit post to be less than 2. While going through the dictionary of male and female words, I discovered a lot of words that were not useful, such as '.', '*', '?!', and so on, so I decided to delete them from my search and consider words that were longer than 2. Using this update, my total accuracy to forecast subreddit posts was dropped to 0.526 and gender prediction accuracy reached 50%, which was unexpected. As a result, I tried different ways to consider words that are less or equal to 7, the reason being that I thought words like "Male", "Women", "Female", "she", "wife", "husband", and so on would result in a lower level of accuracy for gender prediction but my results were not impressive.
As a result, I tried different ways to consider words that are less or equal to 7, the reason being that I thought words like "Male", "Women", "Female", "she", "wife", "husband", etc., but there was no significant difference in getting lower accuracy for gender prediction level and higher level of accuracy in case of predicting subreddit post.
I noted that the computation time for model 2 was rather long, and it improved slightly when I made changes to model 3. The following is a qualitative example from model 3:
Qualitative example model 3: [*Had he ever asked you to join him when he goes or have any plans to introduce you to them when they come?]*

Below is the table 1and scatter plot that shows gender classification accuracy and sub-reddit post accuracy for all four models (one with original dataset and rest for model 1,2 and 3).

In this assignment, we only looked at binary variables, but in reality, there are more than two gender variables to be addressed. Including, Other - Non-Binary, it is possible to accurately

represent the entire population in the real world. However, because I believe the population for non-binary gender would be small, this may have an impact on the outcome. We also need to specify the vocabulary that the non-binary population employs in order to conceal the information which may also require additional research.

Table 1: Model and their prediction level		
Model	**Gender Classification Accuracy**	**Subreddit Accuracy**
Original Model	0.6495	0.8585
Model 1	0.289	0.517
Model 2	0.082	0.6525
Model 3	0.5	0.526

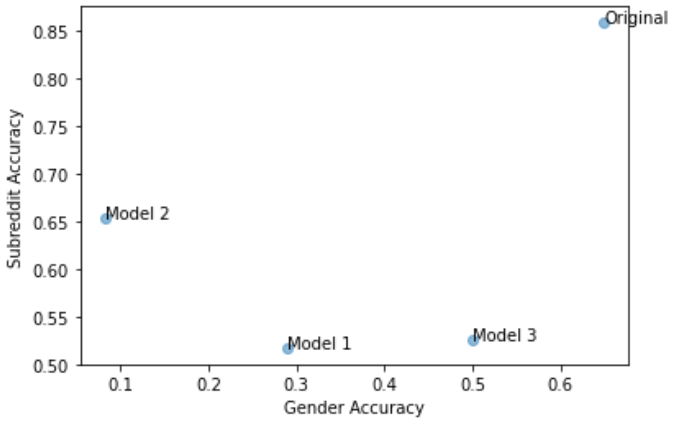

Fig 1: Scatter plot for four Models

In general, when employing artificial intelligence techniques for data prediction that involves humans, it is critical to address the issue of gender bias. I was reading this research study by Bolukbasi (2016), that uses algorithms to greatly eliminate gender bias in embeddings using crowd-worker evaluation as well as standard benchmarks [1]. To eliminate gender inequality, the embedding of gender-neutral words is modified in this analysis by eliminating their gender connection. This is accomplished by first capturing gender inequality in the direction of word embedding and then linearly separating the gender-neutral words from the gender definition words in the word embedding. The study has also shed light on how society is currently skewed, which has a significant impact on word embedding. As a result, rather than word embedding, one should attempt to debias society

Another research study that I read by Grag(2018), creates a context to explain how the temporal complexities of embedding serve to measure shifts in stereotypes and attitudes against women and ethnic minorities in the United States between the twentieth and twenty-first centuries [2]. To ensure that the bias in the embedding accurately reflects sociological trends, this analysis reveals patterns in the embeddings with quantifiable population trends in occupations involvement, as well as historical surveys of stereotypes. The study uses 100 years of text data to train word embeddings to quantify the biases for occupations and adjectives. This research makes use of the embedding bias to investigate historical shifts that would otherwise be difficult to measure. The study shows both gender and racial occupation biases in the embeddings which associate substantially with real occupation participation rates. It shed light on how particular biases diminish over time when other forms of stereotyping increase. It also illustrates adjective correlations in embeddings that offer details about how different groups of individuals are interpreted over time.

Solution 2.

Brief Summary:

Currently, research institutions utilize a set of specific rules to accept or reject experiments before conducting research. The Institutional Review Board (IRB) is the administrative authority that governs decisions involving human subjects, however NLP and other data sciences have typically not followed such restrictions. According to Dirk and Shannon the public outcry over Facebook's "emotional contagion" experiment shows that data science impacts human beings in real time, prompting us to examine the ethics of using NLP approaches in research [3]. Following are some key terms from the research paper [3] on technology ethics as mentioned below:

a) **Exclusion:** Overfitting and demographic bias may occur in any data collection. Most phycology research relies on western-educated, industrialized, wealthy, and democratic study participants. It is possible to address for demographic bias in data by using procedures that address overfitting or imbalanced data.
b) **Overgeneralization**: Overgeneralization is a modeling error. Using models that produce false positives may result in bias confirmation.
c) **Topic Overexposure:** The availability heuristic, a psychological phenomenon in which people assume that the things they know about are more important because they can remember them, can be triggered by overexposure. If research shows that the language of a given ethnic or racial group is more difficult to understand, that group may be perceived as difficult or odd.
d) **Topic under exposure:** NLP tends to focus on texts written in Indo-European languages, rather than languages spoken in Asia or Africa. There has been a lack of attention paid to typological diversity because of an abundance of resources focused on English-speaking users and developers.
e) **Dual Usage:** Even if we consider all the ethical implications of the research and do not intend to harm society, NLP may have unintended consequences. NLP developers should avoid obvious dual-use technologies in favor of ethical alternatives (e.g., work on health-care applications). While we cannot be held directly liable for the research's unanticipated outcomes. We can, however, identify the ethical implications of NLP through research, raise awareness, and educate people about the consequences.

Personal response:

NLP is about human language processing, and human language touches many parts of life; these areas also have an ethical dimension which are derived from the mutual relationships between language, society, and the individual. The 21st century has seen the ethical challenges raised by NLP, both in terms of the way research is performed (plagiarism, reproducibility, and transparency) and in terms of its influence on society [3] [4]. Consequently, it is important to address the implications of NLP on the way we conduct research.

Most earlier NLP applications concentrated on enhancing existing text that was not particularly identified with any one author. It was often published openly and had some temporal gap. As a result of these characteristics forming a barrier between the text and the author, the NLP research had little to no impact on the author's perspective. This is no longer the case since the proliferation of social media platforms and the widespread usage of these platforms by people has altered the societal effect of NLP applications. The language that individuals use reflects their human behavior and a powerful indicator of their distinct personalities [4]. On the social platform, individuals use language in both conscious and unconscious ways that may represent their identities as members of a group with a certain belief or set of characteristics. At times, the

language employed might be influenced by the context, time, circumstance, or location. These elements leave an imprint on the text that, when employed in NLP, may reveal various degrees of information about the author and the circumstance.

References:

[1] Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, *29*.

[2] Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635-E3644.

[3] Hovy, Dirk, and Shannon L. Spruit. "The social impact of natural language processing." In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 591-598. 2016.

[4] Fort, Karën, and Alain Couillault. "Yes, we care! results of the ethics and natural language processing surveys." In international Language Resources and Evaluation Conference (LREC) 2016. 2016.

[5] Natural Language Processing by Jacob Eisenstein

[6] McKinney, Wes. "Pandas, python data analysis library." URL http://pandas. pydata. org (2015).

[7] Link: https://www.cs.toronto.edu/~lczhang/321/

[9] Ed discussion board for A3 Assignment and discord group.