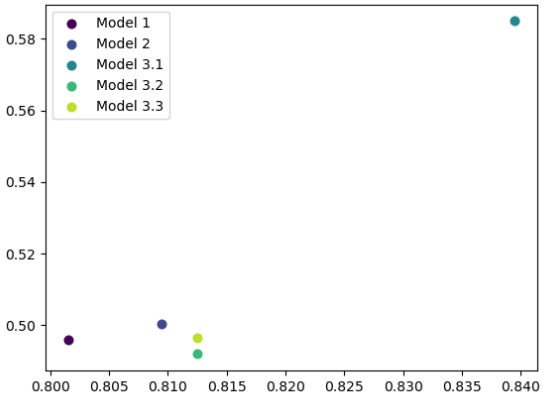
Assignment 5 Writeup

Murray Kang

Problem 1



	Subreddit Accuracy	Gender Accuracy
Model 1	0.8015	0.496
Model 2	0.8095	0.5005
Model 3.1	0.8395	0.585
Model 3.2	0.8125	0.492
Model 3.3	0.8125	0.4965

- Model 1: My baseline obfuscation model that replace the words in one gender text with random words in the other gender text.
- Model 2: I used the cosine similarity to determine the most semantically similar words by using the Google's trained Word2Vec word embeddings provided in the spec, and then I replaced the word in one gender text with the word in the other gender text with highest cosine similarity. For the word out of the vocabulary, I randomly chose one word in the other gender text to replace it. I chose cosine similarity since in the paper, Pinky and his team showed that cosine similarity is a metric that can represent semantic similarity with relative high accuracy.
- Model 3.1: For the word in one gender text, I randomly choose if it would be replaced. If so, and it is not a VOO, replace it with its most semantically similar word in the other gender text by cosine similarity; if it is a VOO, replace it with a random word in the other gender text.
- Model 3.2 : I used a threshold 0.5 to select the semantically similar word. For the word in one gender text and not a VOO, replace it with the most semantically similar word in the other gender text which cosine similarity is at least 0.5. For VOOs, ignore them.
- Model 3.3: This model is a small variant of the model 3.2. For the word in one gender text that does not have similar enough counterparts (maximum cosine similarity < 0.5) in the other gender text, I labeled them as 'NCW' (non-counterpart word). For the word in one gender text which is neither a VOO, nor an NCW, replace it with its most similar word in the other gender text. For VOO and NCW, replace it with a random word in the other gender text.

Examples: I chose an example of male reddit, and the shaded words are the ones in the male text.

Remember, his wife did not start this affair because he cheated. His wife did it of her own accord. There is hypocrisy here, but there is a major difference between a small fling and a conscious effort that lasted 5 years involving the people closest to you.

Model 1: Remember, his trio did frames shanghainese malone affair because he cheated. His tobanyaki did it olive her own accord. There mix-ins hypocrisy here, but there happily time septum sent molten photobooth small fling and migraine conscious treats annoyed lasted figs esther involving migraines people closest to you.

Model 2: Remember, his husband did did started last affair because he cheated. His husband did it knick her own accord. There was hypocrisy here, but there was leche huge really both anxious small fling and portabello conscious try it lasted 2 weeks involving however people closest to you.

Model 3.1: Remember, his husband did not start last affair because he cheated. His wife did it of her own accord. There is hypocrisy here, but there is a huge difference both thrilled small fling and ma'am conscious effort it lasted 2 weeks involving the people closest to you.

Model 3.2: Remember, his husband did did started last affair because he cheated. His husband did it focaccia her own accord. There was hypocrisy here, but there was grocery huge son salesgirl exhibit small fling and ck conscious faves it lasted 2 weeks involving prenatal people closest to you.

Model 3.3: Remember, his husband did did started last affair because he cheated. His husband did it curry her own accord. There was hypocrisy here, but there was rolled huge kibble cheese soap small fling and drizzled conscious steal it lasted 2 weeks involving our people closest to you.

Instead of treating gender as a binary variable, we can treat it as a classification task with multiple labels. However, it's better to think of gender as falling along a spectrum rather than being comprised of a series of discrete, clearly designated labels. Given such circumstances, the classification task to classify every author in the reddit into one specific label is impossible. The pro of doing so is we fully respect everyone’s opinion regarding their own gender. The con of doing so is we need to create many labels and if there are more and more authors we need to take into account, there is going to be nearly infinite labels needed to create, which is very hard to distinguish the difference for the machine.

Another way of treating gender is we can just use three labels to finish this task, female, male, and other (non-binary). Now it becomes a classification task for three targets. The pros are this is doable, and we can really build a model to classify and obfuscate the texts, and this method respect the non-binary people. The cons of doing so are it might be hard to find the words that are the representative of non-binary people or non-binary people are more likely to use and to say. In other words, the accuracy of such model for classifying non-binary people might be relatively inaccurate.

Generally, obfuscation can diminish some discrimination and stereotypes for a specific group of people. In terms of gender, in sociolinguistics, gender is known to be one of the most important social categories driving language choice. Obfuscation of gender can diminish the stereotypes and discrimination upon one gender. For example, if there is a female coach teaching driving on reddit, someone might think women drive worse than men, so that he or she might not be respectful to the author. In terms of ethicality, people might have some stereotypes upon some specific group of people. With obfuscation, readers are less likely to judge authors based on their ethicality and it can protect the minority from being judged or discriminated.

Extra Credit:

I used a model for style transfer in Prabhumoye et al. I firstly cleaned the data and preprocessed the data by doing lemmatization, stop-word removal, and filtering characters, punctuations and numbers. Then, I split the large dataset into training dataset, development set, and test set based on the gender with the proportion of 7:2:1. Then, as followed the steps in the paper, I used an encoder and decoder to translate the training set and development set from English to French. Then, I used the fixed encoder from French to English to encode this sentence in English without decoding. Next, I trained two separate decoders for male and female in English. I used a convolutional neural network classifier to predict the given style and to evaluate the error in the samples generated for the desired style. What's more, I used a bidirectional LSTM to build decoders which generate the sequence of tokens that are conditioned on the latent code z. I used a corpus translated to French by the machine translation system as the input to the encoder of the backtranslation model. The same encoder is used to encode sentences of both female and male. I also used global attention to aid generators.

Hyperparameters:

two-layer LSTM, input_size = 300, hidden_dimension = 500, maximum length of a sentence = 50, global attention vector dimension = 500, 100 filters of size 5 for the CNN classifier, the size of input to CNN = 302.

Results:

The subreddit accuracy on test set: 84.435%

The gender accuracy on test set: 52.243%

Problem 2

Summary

Hovy and Spruit's (2016) article, *The Social Impact of Natural Language Processing*, begins by offering an example of applying ethics in the medical field and explaining how data science affects humans directly. Privacy concerns are the primary ethical issues in adopting and using data sciences. However, digital management rights, policy-making procedures, and security matters are not specific to the NLP, causing privacy concerns (Hovy & Spruit, 2016). Therefore, the authors dismiss the discussion on privacy concerns and analyze the social impact attached to NLP. IT experts can now use language to predict individuals' traits; the continued expansion of NLP leads to an increase in the NLP's ethical implications. Language plays a significant role in creating demographic misrepresentation, which the article considers a form of exclusion. Hovy and Spruit (2016) stipulate that focusing on specific groups of people while excluding others risks the development of scientific knowledge. In addition, the authors state that the automatic interference of user attributes has led to the modeling effect of overgeneralization. The use of NLP promotes the increase in language for specified topics, thus leading to biasness (Hovy & Spruit 2016). The higher the exposure, the greater the biasness, which may further promote discrimination for the undiscussed. The dual usage ability of NLP is also addressed as a primary social effect (Hovy & Spruit 2016). NLP can detect fake reviews but can also be used to generate fake reviews. The double usage of NLP, in this case, raises the alarm on the need for people to be aware of the possible application of NLP for individual interests. The authors conclude by providing mitigation measures to the addressed social impacts. Among the identified solutions is developing complex research designs that can help reduce the impact of overgeneralizations (Hovy & Spruit 2016). Adopting bias control techniques was recommended

as the mitigation measure for exclusion. The community is called upon to help address the social impact attached to double usage. However, the authors desire that the paper will help in raising concerns on the ethical considerations required in NLP.

Personal Response

The article's contributions are crucial in developing NLP ethics. Therefore, it is easy to develop a code of ethics with the possible NLP's social impacts understanding. I concur with the authors' provided mitigation measures to offer ethical solutions. I also agree that NLP lacks adequate investigation on its social impact. The lack of appropriate guidelines in the developed countries to help control and develop emerging computer linguistics languages could have affected ethics development. In addition, I concur with the author on the public fear of privacy. Notably, the public is wary of the rumors of high political manipulation levels through AI and the increase in the rates of cyber-attacks. The user rights are also unclear with NLP. The authors recognize that one of the challenges attached to NLP is continued technological developments. All people living in the 21st century agree that technological developments create abrupt and vast changes to human geology. The profound changes affect ethics development as they require continued revisions of computer linguistics values and ethics to accommodate the new changes. It also takes time to understand such changes, leading to extensive delays. I believe that the authors should have provided more supporting evidence to prove the authenticity of the results of the research findings. However, offering examples of the social effects of NLP was a noble task as that strengthens the understanding for the general audience.

Reference

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black. Style transfer through

back-translation. In Proc. of ACL, 2018. URL https://aclanthology.org/P18-1080.

Sitikhu, P., & Pahi, K. (n.d.). A Comparison of Semantic Similarity Methods for Maximum Human Interpretability. Retrieved February 10, 2022, from https://arxiv.org/pdf/1910.09129

Hovy, D., & Spruit, S. L. (2016, August). 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).