# Gender Obfuscation and Transfer

**Kapil Gupta  Zekun Zhao  Lia Gianfortone**

## 1. Introduction

As algorithms that classify attributes such as gender from textual data become increasingly accurate and prolific, obfuscation of such attributes is useful to preserve the privacy of authors of shared textual content. Sensitive material can be shared globally if it is possible to successfully hide the protected attributes of the data without changing the intent of the data, giving the author more control over the reception of their content.


*Figure 1.* The well-known gender difference in taboo word usage, strikingly evident in the word clouds for female vs. male Facebook posts (Schwartz et al., 2013)

There is an ethical responsibility on the part of the data scientist to commit to the removal of and/or accounting for bias from data, in algorithms, and in the allocation of resources downstream from the data analysis. A social media poster should not assume that their published data will be used responsibly or without bias and may thus wish to assert some control over NLP interpretations of their data. This project seeks to develop and study methods that are able to obfuscate an attribute (gender) without diminishing the meaning of the obfuscated text. These experiments define some initial steps in the development of a privacy tool that would allow a social media user to post potentially sensitive data that thwarts NLP gender detection.

## 2. Related Work

There is an ample selection of recent works in which researchers are trying to mitigate gender bias in natural language processing algorithms (NLP). A collection of these works is organized nicely in a literature review by UCSB and UCLA computer scientists (Sun et al., 2019). The latest state-of-the-art models aim to decouple the style from the content with a variety of different methods.

One such approach is outlined in "In Predicting Sales..." (Pryzant et al., 2017). The authors seek to identify the linguistic features that people can use to predict the target attribute. The first method they implement uses a convoluted neural network (CNN) to encode the text. Via an attentional bi-LSTM, the vector feature is extracted from text data. Then, using a series of feedforward neural networks to predict attributes. The most important step is the gradient reversal layer, which encourages the model to learn representations of the text which are unrelated to the certain label. The adversarial objective function is to control for confounding variables, and project the network's weights onto its activation functions to interpret the importance of each phrase towards each output class. The other method they utilize leverages residualization to control for confounds and performs interpretation by aggregating over learned word vectors. It first predicts target labels from other attributes as well as possible, and then seeks to fine-tune those predictions using a bag-of-words representation of the text. This two-stage prediction process implicitly controlled the words which used to explain other attributes. These methods inform one of our approaches (Section 3.2).

## 3. Methods

We used several models of textual replacement to obscure the gender of a text's author for a dataset[1] of Reddit text posts. The primary dataset (`test.csv`) consists of 2000 text posts, each annotated with the gender of the post's author ("op_gender") and the subreddit where the post was made ("subreddit"). The main text of the post is in the column "post_text". The contents of the provided data include:

- `classifier.py`: a simple word-based classifier that predicts the author's gender and the subreddit for a post
- `train.csv`: training data for the classifier test.csv
- `background.csv`: additional Reddit posts
- `female.txt`: a list of 3000 words commonly used by women
- `male.txt`: a list of 3000 words commonly used by men.

### 3.1. Baseline Model and Pre-trained Word Embeddings

The simplest of the implemented methods was directed by a Carnegie Mellon University (CMU) assignment on "Privacy and Obfuscation" (Yulia Tsvetkov, 2019 (accessed May 1, 2020)) that was based on "Obfuscating Gender in Social Media Writing" (Reddy & Knight, 2016).

The assignment prompted us to compare the words of each text post by the 3000-word-count corpora of male- and female-associated words and replace words associated with the non-target (original) gender with words that are randomly selected from the target (opposite) gender's common-words corpus.

Using the same corpora and textual data, for each text post we replaced words associated with the non-target gender with the word from the other gender's associated corpus that is determined to be most contextually or semantically similar. We performed this using the GloVe 200d (Pennington et al., 2014) and Word2vec (Church, 2017) pre-trained word vector embeddings and the WordNet (Fellbaum, 1998) lexical database of semantic relations.

### 3.2. Adversarial and Residual Networks

Instead of using the existing corpora of gender associated words, we used both adversarial and residualization networks to get lists of words based on data from different domains. In the experiment, we set training step as 800 and batch size as 2 to calculate scores for 5000 common tokens, since we assume that the usage of common words can reflect the gender difference easily.

[1] http://demo.clab.cs.cmu.edu/ethical_nlp2019/homeworks/hw4/hw4.html

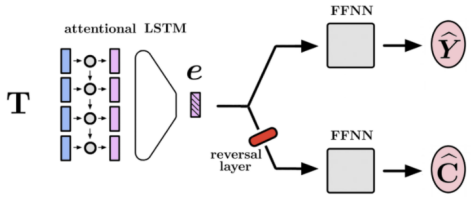


*Figure 2.* Adversarial network (Pryzant et al., 2018)

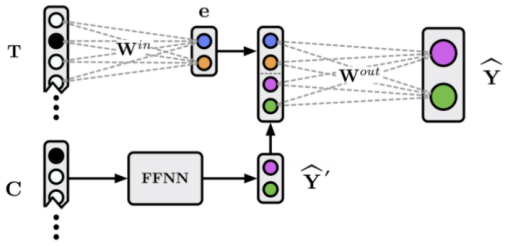


*Figure 3.* Residualization network (Pryzant et al., 2018)

### 3.3. Style Transfer

In this method, instead of replacing words from a processed lexicon, we transferred the style of text from current gender to the other gender. The sentences were rephrased by the model by first translating(encoding) them into french, and while back-translation, using a decoder that has the other gender's style. The style encoders and decoders were trained on a large corpus of gender specific text. The method uses a Variational Autoencoder (Kingma & Max, 2014) for language translation and Cross-aligned Autoencoders (Shen et al., 2017) for performing back-translation(using encoder) and style-transfer(using decoder). We used the pre-trained language translation model (trained on more than 1 million sentences) to run on the input data and the trained style transfer model along with back-translation model to get the text with different style.

The idea behind language translation before performing the style transfer is that it has been observed in previous works that performing translation and then back-translation leads to loss of style in the sentence while preserving context. This is essential because removing previous style from the sentence is important before adding a new style to make the sentence coherent.

## 4. Evaluation

The evaluation metric is a classifier script provided by CMU. The script uses logistic regression to classify texts by the predicted gender of the author, as well as a specified target

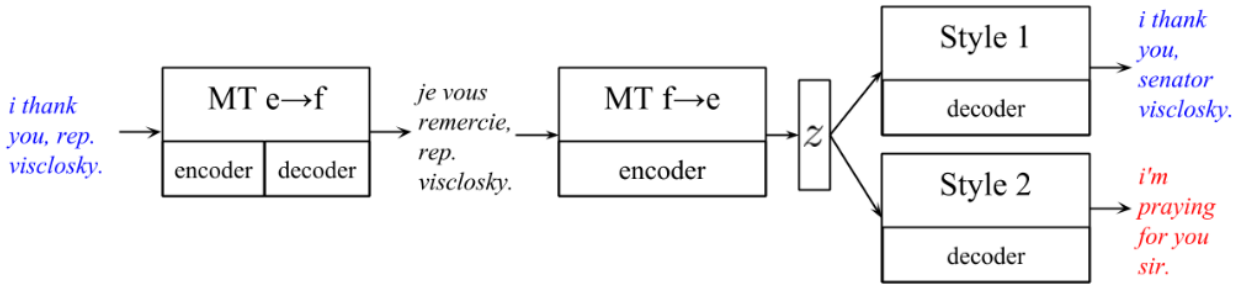


*Figure 4.* "Style transfer pipeline: to rephrase a sentence and reduce its stylistic characteristics, the sentence is back-translated. Then, separate style-specific generators are used for style transfer" (Prabhumoye et al., 2018)

| | Gender ACC | Subreddit ACC |
|---|---|---|
| Original classification | 0.656 | 0.8385 |
| Base random model | 0.455 | 0.709 |
| Glove 200d | 0.525 | 0.818 |
| Word2Vec | 0.5245 | 0.8075 |
| WordNet | 0.5605 | 0.812 |

*Table 1.* results by using different embedding methods to find most similar word to replace target word

attribute. For the Reddit data, this target attribute is the subreddit that the text was posted to. If we are able to increase the FPR and FNR values and reduce the accuracy of the classifier model from 65.65% (current accuracy) to about 50% (equivalent to random guessing), then we are successful.

## 5. Results

### 5.1. Baseline Model and Pre-trained Word Embeddings

The gender and subreddit classification accuracies for the various models are listed in Table 1.

While the baseline model of random replacement reduced the accuracy of gender classification the most, it also reduced the accuracy of subreddit classification substantially. Random replacement of words also resulted in nonsense sentences that are neither semantically nor syntactically sound. As an example of this confuddling, here is an unmodified Reddit text post that was written by a male-identifying person,

> It really comes *down* to *the* circumstances *under* which *you* broke up and *your* relationship (if any) afterwards.

And here the text has been modified with random substitution of female-associated words,

> It really comes *us* to *said* circumstances *gelato* which *milk* broke up and *marshmallows* relationship (if any) afterwards.

The models that used text embeddings to inform word replacement reduced the accuracy of gender classification by approximately equal amounts, but the GloVe implementation outperformed the Word2vec one in preserving the accuracy of subreddit classification. The modified posts are more coherent than those from random substitution but the word substitutions often lack semantic similarity. The Word2vec modification of the Reddit text post introduced above is

> It really comes *back* to *however* circumstances *umbrella* which *we* broke up and *my* relationship (if any) afterwards.

The same post modified with GloVe becomes

> It really comes *back* to *section* circumstances *which* which *n't* broke up and *my* relationship (if any) afterwards.

Replacing gendered words with words associated with the opposite gender that are deemed most similar by WordNet did not diminish the accuracy of gender classification by as much as the previously discussed models but the subreddit classification accuracy was on par with that of the GloVe model. We still found that the modified texts lack coherency and do not maintain the original meaning of the text. Here is an example of one such modified text post,

> It really comes *hair* to *drs* circumstances *pineapple* which *obsession* broke up and *extractions* relationship (if any) afterwards.

Overall, we find that the GloVe 200d implementation had the best performance for this dataset of Reddit posts. Improvements upon this implementation would include further

pre-processing of the data or lexicons. Most word embeddings link words on association or relatedness rather than similarity or synonymity. We could evaluate (and perhaps improve) the degree to which textual meaning is preserved in the gender obfuscation implementations discussed in this section with the incorporation of SimLex-999 (Hill et al., 2015).

### 5.2. Adversarial and Residual Networks

We have several experiments that implement adversarial and residual networks, which have been summarized in Table 3. There are two interesting results we are interested:

- Outside data cannot be used directly on our test examples, the accuracy dose not change too much for both gender and subreddit attributes.
- Using the adversarial and residual networks trained on provided data, we can using a much smaller dataset to replace the words in a sentence but still have a good performance on mitigating the gender attribute.

There could be different reasons to explain the situation. We find that the most convincing one is that different domains could have different languages usage behavior. We could have a general list if we have data from all domains, but it is a challenge job to finished. However, it is more piratical that we can use our model to get a smaller list which is specific from different domains.

| Male | Female |
|---|---|
| UNK | before |
| low | dead |
| yes | guy |
| simply | theyre |
| public | male |
| spend | above |
| crazy | concerned |
| past | appropriate |
| pull | friend |
| private | times |
| damn | used |
| left | atmosphere |
| gun | boyfriend |
| enjoy | particularly |
| around | once |
| dick | form |
| new | hope |
| fan | insane |
| literally | definitely |
| likely | ever |

*Table 3.* The 20 common words associated with binary gender classifications trained from background dataset

### 5.3. Style Transfer

As discussed before, the method first translates from English to french and then back-translates it to English. The output is then used to add the required style of the other gender to the sentence. Following are a few examples of this method.

> *Original (Female):* what a lovely wee shop !
>
> *French Translation :* ce guapos; un adorable magasin !
>
> *Style Transfer (Male):* what a nice place ?
>
> *Original (Male):* my wife ordered the tenderloin sliders .
>
> *French Translation :* ma femme a commandé les curseurs tenderloin .
>
> *Style Transfer (Female):* my husband ordered the tenderloin peeps .

The generated sentence's quality depends upon various factors such as length of the sentence, words used in the sentence as well as the language translation produced by the model. Since the model has been trained on short sentences with average length of 12 words, longer sentences do not perform well using this model. Similarly, if the sentence uses a lot of esoteric words, their translation can be imperfect, leading to lower accuracy in subreddit classification.

An example of a sentence that did not work as well is shown below.

> *Original (Female):* It really comes down to the circumstances under which you broke up and your relationship (if any) afterwards.
>
> *French Translation :* It tombe réellement dans les circonstances dans lesquelles vous avez brisé et votre relation (if any) suite à la suite de la suite.
>
> *Style Transfer (Female):* It's really nice in the area you have been back and your dedication (if ' s of the past time.

## 6. Future Directions

For our future work, we would like to introduce Abstract Meaning Representation (AMR) (Banarescu et al., 2013) into our model which is a semantic language representation using graph structure. It is intuitive that semantic representations can be useful for machine translation (Song

| MODEL AND NUMBER OF TRAINING DATA | GENDER ACC | TREND | SUBREDDIT ACC | BETTER? |
|---|---|---|---|---|
| ORIGINAL 6000 | 0.5505 | | 0.8230 | |
| REMOVE NON-ENGLISH TOKENS 6000- | 0.5485 | ↓ | 0.8205 | |
| OUTSIDE DATA TRAINED BY MODEL 1 6000 | 0.6420 | | 0.8270 | × |
| COMBINE AND FILTER 6000+ | 0.6015 | ↑ | 0.8275 | |
| ORIGINAL DATA TRAINED AND SEPERATED BY MODEL1 6000 | 0.5340 | ↓ | 0.7930 | |
| ORIGINAL DATA TRAINED AND SEPERATED BY MODEL2 6000 | 0.5350 | ↓ | 0.8000 | |
| REDUCED LIST SEPERATED BY MODEL2 3000 | 0.5500 | | 0.8015 | |
| REDUCED LIST SEPERATED BY MODEL2 2000 | 0.5470 | ↓ | 0.8030 | |
| REDUCED LIST SEPERATED BY MODEL2 1500 | 0.5370 | ↓ | 0.8045 | √ |
| REDUCED LIST SEPERATED BY MODEL2 1000 | 0.5600 | ↑ | 0.8165 | |
| REDUCED LIST SEPERATED BY MODEL2 400 | 0.5820 | ↑ | 0.8125 | |
| REDUCED LIST SEPERATED BY MODEL2 200 | 0.5945 | ↑ | 0.8240 | |

*Table 2.* EXPERIMENTS FROM MODEL1[RESIDUALIZATION] AND MODEL2[ADVERSARIAL].

et al., 2019), mainly because they can help in enforcing meaning preservation and handling data sparsity. We believe that incorporating AMR as additional knowledge can significantly improve a strong attention-based sequence to sequence model.

We identified a dataset of Enron emails have been labeled with the author's gender (Cukierski, 2016 (accessed June 1, 2020)) that is similar enough to the dataset we used in this project that is would be useful for performance comparison.

The aforementioned literature review (Sun et al., 2019) addressed several future directions in gender obfuscation that we perceive to be important. Addressing the non-binarism of gender is critically important. Extending obfuscation of attributes such as gender in texts written in non-English languages another essential direction. We would also like to implement the style transfer method presented in this paper with a language other than French. Lastly, continued interdisciplinary endeavors are needed to strengthen privacy protection algorithms. It is important that these interdisciplinary collaborations are made between not only technical fields, but with sociological ones as well in order to address latent bias in data and otherwise.

## 7. Acknowledgements

Thanks to Professor Lise Getoor for your advisement and insights on this project and, to TA Rishika Singh as well, for facilitating an imperative course with exemplary execution.

## 8. Team Member Contributions

We all worked on the initial assignment from CMU individually and discussed the results from Baseline Model and Pre-trained Word Embeddings.

Lia Gianfortone worked on finding relevant papers and datasets for the objective of gender obfuscation. She found the enron dataset as well as twitter dataset and worked on understanding which parameters could be used by the following two algorithms for context preservation. She also worked on the Pre-trained WordNet Embedding, editing the poster and the paper.

Zekun Zhao worked on the Pre-trained Glove Embedding and Adversarial and Residual Networks. Designed experiments for comparing results in table 2 and discussed the causes and future directions. He also used the text-to-speech website [2] to generate the audio for his part of the presentation.

Kapil Gupta worked on the Style Transfer method, where he trained the style-transfer decoder while using the language models present in the repository of the code. The language model needed to be retrained on reddit text and was too large to be trained on colab. So he setup the remote server available as University resource to train the model.

## References

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pp. 178–186, 2013.

Church, K. W. Word2vec. *Natural Language Engineering*, 23(1):155–162, 2017. doi: 10.1017/S1351324916000334.

Cukierski, W. *The Enron Email Dataset*, 2016 (accessed June 1, 2020). URL https://www.kaggle.com/wcukierski/enron-email-dataset.

Fellbaum, C. Wordnet: An electronic lexical database. *Cambridge, MA: MIT Press*, 1998.

[2] http://www.fromtexttospeech.com

Hill, F., Reichart, R., and Korhonen, A. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, 2015.

Kingma, H. and Max, W. Auto-encoding variational bayes. *International Conference on Learning Representations (ICLR)*, 2014.

Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

Prabhumoye, S., Tsvetkov, Y., Salakhutdinov, R., and Black, A. W. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 866–876, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1080. URL https://www.aclweb.org/anthology/P18-1080.

Pryzant, R., Chung, Y., and Jurafsky, D. Predicting sales from the language of product descriptions. 2017.

Pryzant, R., Basu, S., and Sone, K. Interpretable neural architectures for attributing an ad's performance to its writing style. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 125–135, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5415. URL https://www.aclweb.org/anthology/W18-5415.

Reddy, S. and Knight, K. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pp. 17–26, 2016.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., and Ungar, L. H. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE*, 8:1–16, 09 2013. doi: 10.1371/journal.pone.0073791. URL https://doi.org/10.1371/journal.pone.0073791.

Shen, T., Lei, T., Barzilay, R., and Jaakkola, T. Style transfer from non-parallel text bycross-alignment. *Neural Information Processing Systems (NeurIPS)*, 2017.

Song, L., Gildea, D., Zhang, Y., Wang, Z., and Su, J. Semantic neural machine translation using amr. *Transactions of the Association for Computational Linguistics*, 7:19–31, 2019.

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., and Wang, W. Y. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1630–1640, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1159. URL https://www.aclweb.org/anthology/P19-1159.

Yulia Tsvetkov, Alan W Black, A. F. *HW 4: Privacy and Obfuscation*, 2019 (accessed May 1, 2020). URL http://demo.clab.cs.cmu.edu/ethical_nlp2019/homeworks/hw4/hw4.html.