Gender Obfuscation Methods in Subreddit Classification

Athiya Deviyani
Carnegie Mellon University
adeviyan@cs.cmu.edu

Abstract

Applications such as targeted advertising have raised privacy concerns as social media posts may contain implicit features indicative of a user's personal information such as gender. The goal of this paper is to explore several methods to obfuscate an author's gender given a sequence of words. We will also explore its performance and the trade-offs between obfuscating an author's gender identity and preserving useful information in the data that might be useful for downstream natural language processing tasks.

1 Introduction

In the past few years, we have witnessed the exponential growth of social media content, and with it, accompanying metadata which might explicitly or implicitly surface user personal information such as name, gender, nationality, ethnicity, and more. This information may be used without the user's knowledge by revenue-generating companies, which consequently compromises the user's privacy. As noted by Hovy and Spruit (2016), user profiling can be used in ethically negative ways by adversarial parties. A study by Datta et al. (2014) shows that Google serves fewer ads for high-paying jobs to users profiled as female, entailing that sometimes personalized information contains inadvertent discrimination. Luu (2015) further suggests that some users would conceal their gender identity in order to avoid harassment in online forums. In this paper, we will evaluate several obfuscation strategies with the aim to 'confuse' a blackbox gender classifier. Additionally, we will evaluate the performance of a blackbox Subreddit classifier on the resulting obfuscated data. Finally, we will perform qualitative analysis on the obfuscated text to evaluate whether or not the obfuscation strategy generates fluent and meaningful sentences.

2 Methodology

2.1 Dataset

For the basic analysis, we are given a balanced dataset of 2000 Reddit posts with their corresponding Subreddit tags (FUNNY and RELATIONSHIPS) and the gender label of the original poster (M for man and W for woman). We are also given an accompanying list of male and female words, each containing nearly 3000 instances.

2.2 Obfuscation methodology

For our basic analysis, we will go through each Reddit post of the dataset and obfuscate them with respect to the words in the post. Given a word that is telling of the author's original gender, we would replace the word with a different word that belongs in the word list of the opposite gender, essentially flipping the gender of the original word. Below are the different word-replacement heuristics that we explored: replacing the word with a random word belonging in the word list of the opposite gender (RANDOM), removing any word that belongs in the gender words lists (REMOVAL), and replacing the word with the most similar word computed through the cosine similarity of the word vector representation using SpaCy's `en_core_web_lg` (SIM). We also experimented with replacement conditioned on similarity thresholds of above 0.5 (SIM50), 0.6 (SIM60) and 0.7 (SIM70).

2.3 Evaluation metrics

We are given a blackbox binary classifier which predicts the gender of the author given a Reddit post. 'Blackbox' entails that we have no knowledge of the model structure or what it learns, other than obtaining output labels given an input. We are also given a Subreddit classifier which operates similarly. The gender classifier has an accuracy of 64.95% on the original dataset while the Subreddit classifier has an accuracy of 85.85%.

A successful obfuscation method will have a gender classification performance very close to random (50%). This aligns with the adversarial objective in Generative Adversarial Networks (Goodfellow et al., 2014), where you want the generator to generate instances that achieves a discriminator accuracy of 50%. A low gender accuracy ($< 50\%$) entails that the gender classifier is able to retrieve the original gender label with an accuracy of more than 50% if we just flip the classifier output. On the other hand, the performance of the Subreddit classifier on our obfuscated data should not stray too far from its performance on the original data. The performance of the Subreddit classifier will also show whether or not obfuscating retains the general meaning of the text. We will also perform a qualitative analysis on the obfuscated sentences to further investigate their overall fluency.

3 Analysis of results

3.1 Quantitative analysis

Model	Gender acc. (%)	Subreddit acc. (%)
BASELINE	64.95	85.85
RANDOM	47.00	80.00
SIM	47.65	83.25
SIM50	48.05	83.35
SIM60	**51.15**	**84.30**
SIM70	66.85	85.15
REMOVAL	55.20	81.55

Table 1: Classification accuracies on the datasets obtained through basic obfuscation strategies.

From a cursory observation of Table 1, we can observe that the basic obfuscation methods has not thrown away words that are crucial to the successful Subreddit classification, denoted by the relatively good accuracy range of around 80-85%. Nevertheless, they are all still below the baseline of 85.85%. Interestingly, we can see that almost all of the resulting datasets have achieved a gender classification accuracy of less than the baseline, with SIM60 achieving an accuracy 51.15% which is very close to random while still achieving an accuracy of 84.30%. The SIM70 method has a gender accuracy that is surprisingly higher than the baseline, indicating that a cosine similarity threshold of 0.7 is far too high to make any word replacements that consequently flips the gender label.

From the plot in Figure 1, we can see that there exists a weak positive correlation between the gender and Subreddit classification accuracy. This shows that the gendered words in the Reddit post also contain information regarding the corresponding Subreddit. To evaluate this hypothesis, we assessed the performance of the classifiers on a dataset where all the gendered words are omitted. The Subreddit classifier has dropped to 81.55%, which means that the gendered words have a nontrivial contribution to the Subreddit prediction. Intriguingly, the gender accuracy has not dropped to random, suggesting that the gender classifier has learned stylistic features to learn a user's gender other than the presence of a gendered word.

3.2 Qualitative analysis

From Table 2, we can see that the resulting text obtained through the random word replacement strategy yields the worst result in terms of fluency. The same could also be said about the text obtained through removing gendered words. For the remaining methods, it appears that the words that are most often changed are stop words such as 'this' and 'is', which when swapped out with another word makes the sentence less intelligible. When paying close attention to the gender label column, we can see that the similarity-based methods are more effective at flipping the gender labels, and thus successfully masking the user's actual gender. Even though the sentences generated by the similarity-based methods are slightly grammatically incorrect due to the swapping of stop words, they preserves the original sentiment of the sentence much better. Furthermore, as discussed in the quantitative analysis, a cosine similarity threshold of 0.7 is far too high to make meaningful word swaps, and from Table 2 we can see that the resulting obfuscated sentence is the most similar to the original text and that the gender label has not been swapped.

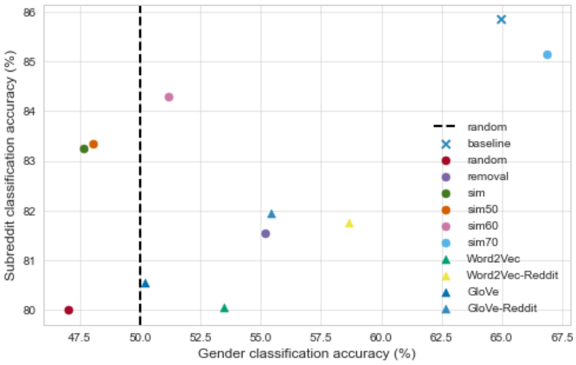

Figure 1: Gender classification accuracy against Subreddit classification accuracy on each obfuscated dataset

Model	Reddit post text	Gender label	Subreddit label
Baseline	This is the person who, by his accounts, abused the fuck out of him.	M	relationships
Basic			
Removal	Person who, his accounts, abused out him.	M	relationships
Random	Beau browse fluff person who, again his accounts, abused toilet cheesy out sauce him.	M	relationships
Sim	It which entire person who, made his accounts, abused entire lick out and him.	W	funny
Sim50	It which entire person who, by his accounts, abused entire lick out and him.	W	funny
Sim60	It which entire person who, by his accounts, abused entire lick out and him.	W	funny
Sim70	It which the person who, by his accounts, abused the fuck out of him.	M	relationships
Advanced			
GloVe	This is the person who, by his accounts, abused the bugger out of him.	W	relationships
GloVe-Reddit	This is the person who, by his account, abused the bugger out of him.	M	relationships
Word2Vec	This is the person who, by his accounts, abused the bugger out of him.	W	relationships
Word2Vec-Reddit	This is the person who, by his accounts, abused the fucking out of him.	M	relationships

Table 2: Resulting text generated by the various obfuscation methods and their corresponding predicted labels.

4 Advanced analysis

4.1 Obfuscation methodology

For our advanced analysis, we will follow the obfuscation methodology introduced by Reddy and Knight (2016). We will use a larger dataset with a similar structure containing Reddit posts tagged with their respective Subreddit and gender labels. We will apply the obfuscation method and evaluate the classifiers on the same dataset used in the basic analysis. The methodology is described as follows:

For each word w_i in a text sequence $\mathbf{w}$ and target label y,

1. Compute the word association $Assoc(w_i, y)$ between w_i and y. The target label is the label that we want the sequence w to be classified as, where $y \in \{y_1, y_2\}$. Therefore, $Assoc(w_i, y)$ is the difference in normalized Pointwise Mutual Information between the two gender classes, more formally defined as:

$$nPMI(w_i, y_1) = \log \frac{P(w_i, y_1)}{P(w_i)P(y_1)} / -\log P(w_i, y_1)$$

$$Assoc(w_i, y_1) = nPMI(w_i, y_1) - nPMI(w_i, y_2)$$

2. If $Assoc(w_i, y) < 0$, this means that w_i is more associated to the opposite label, therefore we would want to replace it. We will build a vocabulary V, where each candidate replacement $v \in V$ satisfies:

$$SynSem(w_i, v) > \tau \wedge Assoc(v, y) > Assoc(w_i, y)$$
$$\wedge\, Assoc(v, y) > 0$$

$SynSem(w_i, v)$ denotes the syntactic and semantic similarity between w_i and v under the dependency parse-based Word2Vec by Levy and Goldberg (2014). Reddy and Knight (2016) argued that this ensures that the substitutions are also syntactically appropriate. τ denotes the similarity threshold, which we have found to be best at 0.7. We also experimented with an additional constraint that enforces the candidate word v to have a positive association with the true Subreddit tag, i.e. $Assoc(v, y) > 0$ where $y \in \{$FUNNY, RELATIONSHIPS$\}$. We hope that this additional constraint will choose substitutions that are not detrimental to the Subreddit classification.

3. Finally, we will compute the substitutability for each candidate word $v \in V$. Given a candidate substitution a and original word b, as well as the context C defined as the words to the left and to the right of b,

$$Subst(a, b, C) = \frac{SynSem(a, b) + \sum_{c \in C} Sem(a, c)}{|C| + 1}$$

The paper defines $Sem(a, c)$ as the cosine similarity between the regular window 5 skip-gram vectors (Mikolov et al., 2013) of a and c, which is the most commonly used Word2Vec embedding. Since the Reddit dataset contains highly colloquial text, we will also extend our evaluation to calculate the cosine similarity of a and c represented by GloVe word embeddings (Pennington et al., 2014) that were specifically trained on a Twitter corpus. We will then replace the word w_i with the candidate word v that has the highest $Subst$. We noticed that substituting stop words in a sentence affects the intelligibility of the sentence, therefore we will only replace w_i if it does not belong in NLTK's list of stop words.

All in all, we will have 4 advanced obfuscation methods: GLOVE wich uses GloVe-Twitter embeddings to calculate $Sem(a, c)$ and the GLOVE-REDDIT method that uses the positive Subreddit association constraint, as well as the corresponding WORD2VEC which uses the Word2Vec embeddings and WORD2VEC-REDDIT.

4.2 Quantitative analysis

Model	Gender acc. (%)	Subreddit acc. (%)
GLOVE	**50.20**	**80.55**
GLOVE-REDDIT	55.40	81.95
WORD2VEC	53.45	80.05
WORD2VEC-REDDIT	58.65	81.75

Table 3: Classification accuracies on the datasets obtained through advanced obfuscation strategies.

The results of the advanced obfuscation methods are shown on Table 3. We can see that GLOVE has the closest gender classification accuracy to random (50.20%) even when compared with the basic obfuscation methods. However, the Subreddit classification accuracy has dropped to 80.55%. This phenomena is shared by all of the advanced obfuscation methods, where the Subreddit accuracies range around 80-82%. Additionally, even though methods that enforce the positive Subreddit association constraint yield a slightly higher Subreddit classification accuracy, the gender classification accuracy is 5% higher for both the WORD2VEC-REDDIT and GLOVE-REDDIT. This suggests the constraint is too strict and thus there were not enough words swapped for succsessful gender obfuscation. This also further strengthens the notion that there exists a trade-off between successfully obfuscating the user's gender and getting a high-performing Subreddit classifier. The plot in Figure 1 shows that the results of our methods follow the trends observed in the basic analysis.

4.3 Qualitative analysis

From Table 2, we can see that the sentences generated by the advanced obfuscation methods are more fluent. We believe that this is primarily due to the fact that we have not replaced the stop words in the sentence. Even though the total number of words flipped by the advanced methods are much fewer than the basic methods, GLOVE and WORD2VEC has successfully defeated the gender classifier. The methods that enforce the positive Subreddit association constraint has not swapped out enough gendered words to be able to mask the true gender of the user. Also, we can see that when converting a gender label of M to W, most of the obfuscation methods will substitute out any words that might indicate profanity. This suggests that the obfuscation methods learned that profanity is highly associated with male speech, which might propagate negative stereotypes. The tables provided in the appendix containing additional sample obfuscated sentences further corroborate our observations.

5 Discussion

The main bottleneck of the obfuscation methods explored in this work is the trade-off between gender obfuscation and Subreddit classification performance. Although obfuscating user data is much more important, having a dataset that yields a poor performance on a downstream text classification task would be useless in practice. The method with the best performance with respect to this trade-off is SIM60, with a gender classification accuracy of very close to random (51.15%) and a Subreddit classification accuracy that is within less than 2% away from the baseline (84.30%).

Additionally, we would want an obfuscation method that yield sentences that are fluent and meaningful, not just to a text classifier but also to humans. This is particularly useful when deploying a tool to users who wish to mask their gender when posting on social media, yet still create grammatically correct and intelligible writing. The sentences generated by the advanced obfuscation methods are much better at this, yet they prove to be less successful at Subreddit classification.

There are several ethical implications that are involved when considering deploying an obfuscation system for public use. Our obfuscation methods employs simple rule-based obfuscation strategies, yielding sentences which are easily distinguishable from natural human-written text. It is highly possible that the generated dataset is vulnerable to obfuscation detection models which allow adversarial actors to retrieve the true gender of the user.

Finally, our experiments are limited by the availability of datasets with more than binary gender labels based on the biological sex. This might not be an accurate representation of users on social media. Consequently, the dataset will not be representative of the linguistic style of a particular gender group, which in itself is ambiguous as we are making a very strict assumption that a particular gender group follows the same writing style.

Aside from that, making this assumption might also further enforce negative stereotypes. As mentioned previously, most of our obfuscation methods tend to associate profanity to male speech, which is shown by how they are changed to more 'polite' words in order to sound more female. It is highly likely that there are more stereotypical associations that have yet to be observed in a larger sample of obfuscated text.

References

Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2014. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *CoRR*, abs/1408.6491.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland. Association for Computational Linguistics.

Chi Luu. 2015. How to disappear completely: Linguistic anonymity on the internet.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Sravana Reddy and Kevin Knight. 2016. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 17–26, Austin, Texas. Association for Computational Linguistics.

Appendices

A Appendix

The tables below show three Reddit posts that are obfuscated using the basic and advanced gender obfuscation methods, to support the observations made in the basic and advanced analysis sections.

Model	**Reddit post text**	**Gender label**	**Subreddit label**
Baseline	Effort? Nevermind. Kidding, I’m pretty sure you guys can help me with my Russian studies.	M	funny
Basic			
Removed	Effort? Nevermind. Kidding, pretty sure help me with my Russian studies.	M	funny
Random	Effort? Nevermind. Kidding, much pretty sure flourless dogs mmm help me with my Russian studies.	M	funny
Sim	Effort? Nevermind. Kidding, I pretty sure sure know able help me with my Russian studies.	M	funny
Sim50	Effort? Nevermind. Kidding, I pretty sure sure know able help me with my Russian studies.	M	funny
Sim60	Effort? Nevermind. Kidding, I pretty sure sure know able help me with my Russian studies.	M	funny
Sim70	Effort? Nevermind. Kidding, I pretty sure sure guys able help me with my Russian studies.	M	funny
Advanced			
GloVe	Effort? Ugh. Kidding, guess pretty sure you guys can help me with my Polish studies.	W	funny
GloVe-Reddit	Effort? Ugh. Kidding, guess very sure you guys can help me with my Polish studies.	W	funny
Word2Vec	Effort? Whoa. Kidding, I’m pretty sure you fellas can help me with my Lithuanian studies.	M	funny
Word2Vec-Reddit	Effort? Ooh. Kidding, I’m pretty sure you fellas can help me with my Lithuanian studies.	W	funny

Table 4: Resulting text generated by the various obfuscation methods and their corresponding predicted labels.

Model	**Reddit post text**	**Gender label**	**Subreddit label**
Baseline	Wow youre a huge fucking cunt. I hope you die a horrible death.	M	funny
Basic			
Removed	Youre huge cunt. I die horrible death.	W	funny
Random	Bogi youre jarred huge megan cunt. I helpful weather die skewers horrible death.	W	funny
Sim	Omg youre kind huge freaking cunt. I glad sure die kind horrible death.	W	funny
Sim50	Omg youre kind huge freaking cunt. I glad sure die kind horrible death.	W	funny
Sim60	Omg youre kind huge freaking cunt. I glad sure die kind horrible death.	W	funny
Sim70	Omg youre kind huge fucking cunt. I glad sure die kind horrible death.	M	funny
Advanced			
GloVe	Haha youre a huge freaking slut. I well you die a horrible death.	W	funny
GloVe-Reddit	Haha youre a huge freaking slut. I well you die a horrible death.	W	funny
Word2Vec	Haha youre a huge freaking twat. I well you die a horrible death.	W	funny
Word2Vec-Reddit	Haha youre a huge fuck twat. I well you die a horrible death.	W	funny

Table 5: Resulting text generated by the various obfuscation methods and their corresponding predicted labels.

Model	**Reddit post text**	**Gender label**	**Subreddit label**
Baseline	Yes absolutely. Talking someone out of a terrible decision is absolutely right.	W	relationships
Basic			
Removed	Absolutely. talking out of a terrible decision is right.	W	funny
Random	Accustomed absolutely. Talking odor out of a terrible decision is yards right.	M	funny
Sim	Yeah absolutely. Talking somebody out of a terrible decision is certainly right.	M	funny
Sim50	Yeah absolutely. Talking somebody out of a terrible decision is certainly right.	M	funny
Sim60	Yeah absolutely. Talking somebody out of a terrible decision is certainly right.	M	funny
Sim70	Yeah absolutely. Talking somebody out of a terrible decision is certainly right.	M	funny
Advanced			
GloVe	Oh absolutely. Conversing somebody out of a atrocious decision is utterly there.	M	funny
GloVe-Reddit	Oh absolutely. Conversing somebody out of a atrocious decision is utterly there.	M	funny
Word2Vec	Yes absolutely. Conversing somebody out of a atrocious decision is unquestionably right.	M	funny
Word2Vec-Reddit	Yes absolutely. Conversing somebody out of a atrocious decision is unquestionably right.	M	funny

Table 6: Resulting text generated by the various obfuscation methods and their corresponding predicted labels.