Assignment 5

CSE 447 and 517: Natural Language Processing - University of Washington

Winter 2022

Please consult the course website for current information on the due date, the late policy, and any data you need to download for this assignment. This assignment is designed to advance your understanding of some ethical challenges that arise in NLP.

Submit: You will submit your writeup (a pdf) and your code (do not include data) via Gradescope. Instructions can be found here. Note that you will make two submissions: one for the pdf, one for the code.

1 Computational Ethics

This assignment is based on homework #3 from the course "Computational Ethics for NLP" at CMU, taught by Yulia Tsvetkov and Alan Black.

Goals Online data has become an essential source of training data for natural language processing and machine learning tools; however, the use of this type of data has raised concerns about privacy. Furthermore, the detection of demographic characteristics is a common component of micro-targeting. In this assignment, you will explore how to obfuscate demographic traits, specifically gender. The primary goals are (1) develop a method for obfuscating an author's gender and (2) explore the trade-off between obfuscating an author's identity and preserving useful information in the data.

Data: The data you need for this assignment is available at `https://nasmith.github.io/NLP-winter22/assets/data/A5.tgz`.

Overview The primary dataset we provide consists of posts from Reddit. Each post is 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:

- `classify.py`: a classifier that predicts the author's gender and the subreddit for a post (example run: `python classify.py --test_file dataset.csv`). Note that this file also uses the two provided pickle files.
- `dataset.csv`: your primary data
- `background.csv`: additional Reddit posts that you may optionally use for training an obfuscation model. We will also provide a larger version.
- `female.txt`: a list of words commonly used by women
- `male.txt`: a list of words commonly used by men

We note that this assignment uses a simple operationalization of gender as binary; we will return to this point below.

The provided classifier achieves an accuracy around 65% at identifying the gender of the poster and an accuracy around 85% at identifying a post's subreddit when tested over `dataset.csv`. Your goal in this assignment is to obfuscate the data in dataset.csv so that the provided classifier is unable to determine the gender of authors, while still being able to determine the subreddit of the post. Note that in this set-up, we treat the provided classifier as a fixed, blackbox adversary (please do not try to hack it). This assignment was largely inspired by Reddy and Knight [5], which may be a useful reference. Scenarios where this obfuscation model might be useful could be social media users who want to preserve their privacy by hiding their gender, without losing the meaning of their post. You could also imagine this is a dataset of health records or other sensitive information that needs to be anonymized before providing it to researchers.

Requirements First, build a baseline obfuscation model:

- For each post in `dataset.csv`, if the post was written by a man ("M") and it contains words from `male.txt`, replace these words with a random word from `female.txt`.
- Obfuscate posts written by women ("W") in the same way (i.e., by replacing words from `female.txt` with random words from `male.txt`).
- Test `classify.py` on your obfuscated data and report what happens to the two accuracy measurements discussed above.

Second, improve your obfuscation model:

- Instead of replacing words from `male.txt` with randomly chosen words from `female.txt`, choose a semantically similar word from `female.txt`. Do the same in reverse. You may use any metric you like for identifying semantically similar words, but you should explain why you chose it. We recommend starting with cosine distance between pretrained word embeddings (available, for example, here)
- Test `classify.py` on data obfuscated using your improved model and analyze the results. The classifier should perform close to random at identifying gender and should obtain at least 79% accuracy on classifying the subreddit.

Third, experiment with some basic modifications to your obfuscation models. For example, what if you randomly decide whether or not to replace words instead of replacing every lexicon word? What if you only replace words that have semantically similar enough counterparts?

Extra Credit: Advanced Obfuscation Develop your own obfuscation model. We provide `background.csv`, a large dataset of Reddit posts tagged with gender and subreddit information that you may use to train your obfuscation model. We also provide a larger version of the background corpus. Your goal should be to obfuscate text so that the classifier is unable to determine the gender of an author (no better than random guessing) without compromising the accuracy of the subreddit classification task. However, creative or thorough approaches will receive full credit, even if they do not significantly improve results. Some ideas you may consider:

- Develop your own lexicons using pointwise mutual information scores or log odds with a Dirichlet prior; see, e.g., Jurafsky et al. [2]

- Follow the procedure described by Reddy and Knight [5]
- Use an adversarial objective as described by Pryzant et al. [4] to train a model that is good at predicting subreddit classification but bad at predicting gender. The key idea in this approach is to design a model that does not encode information about protected attributes (in this case, gender).
- Use a model for style transfer, such as in Prabhumoye et al. [3]

In your report, include a description of your model and results, and clearly label it "Extra Credit."

Deliverables Submit a report that is no more than three pages long. It should include:

- A scatterplot where subreddit accuracy on `dataset.csv` is plotted on the x-axis, gender accuracy on the same data is plotted on the y-axis, and every model you built is a (labeled) point.
- The same information presented in a table (each row a system, one column for subreddit accuracy and one column for gender accuracy).
- One paragraph to describe each system you considered.
- Qualitative examples of text obfuscated with your models.
- As noted, this assignment assumes gender can be treated as a binary variable. Reflect briefly (a paragraph or so) on alternative ways one could formulate this problem, and the pros and cons of doing so.
- Discuss (in a paragraph or so) the ethical implications of obfuscation more generally, drawing from concepts you encountered in class, readings, or elsewhere.

2 Reading and Reflection

Read Hovy and Spruit [1]; write a brief summary of the paper's argument (one paragaph) and then write a personal response (one paragraph).

References

[1] Dirk Hovy and Shannon L. Spruit. The social impact of natural language processing. In *Proc. of ACL*, 2016. URL https://aclanthology.org/P16-2096.

[2] Dan Jurafsky, Victor Chahuneau, Bryan R. Routledge, and Noah A. Smith. Narrative framing of consumer sentiment in online restaurant reviews. *First Monday*, 19(4), April 2014. URL https://firstmonday.org/ojs/index.php/fm/article/view/4944.

[3] 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.

[4] Reid Pryzant, Young joo Chung, and Dan Jurafsky. Predicting sales from the language of product descriptions. In *Proc. of eCOM@SIGIR*, 2017. URL http://ceur-ws.org/Vol-2311/paper_3.pdf.

[5] Sravana Reddy and Kevin Knight. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, 2016. URL https://aclanthology.org/W16-5603.