Ethics for NLP: Spring 2022
Homework 4

TECHNISCHE UNIVERSITÄT DARMSTADT

Due until Thursday, 07.07.2022 at 11:59pm

Submission Guidelines for Homework

- This homework is worth 20 Points.
- Submit your obfuscated versions of the test.csv (name them accordingly) for each task together with the .ipynb notebook as one zip-archive using Moodle. Do not add other files.
- Name your submission file as: hw04.<matriculation_number>_lastname_firstname.zip
- Extra credit shall be given to well-structured submissions.
- In case of questions or remarks, please contact:
  - Sophia Lichtenberg, sophia.lichtenberg@stud.tu-darmstadt.de

1 Privacy

We encourage you to use the attached notebook as a template, as we have already put a lot of code in it. Also, this will give you a head start on the assignment. IMPORTANT NOTE: Please use Google Colab to run the notebook.

1.1 Goals

A major problem with utilizing web data as a source for NLP applications is the increasing concern for privacy, e.g., such as microtargeting. This homework is aimed at developing a method to obfuscate demographic features, in this case (binary) gender and to investigate the trade-off between obfuscating an users identity and preserving useful information.

1.2 Data Overview

The given dataset consists of Reddit posts („post_text“) which are annotated with the gender („op_gender“) of the user and the corresponding subreddit („subreddit“) category.

subreddit_classifier.pickle pretrained subreddit classifier

gender_classifier.pickle pretrained gender classifier

test.csv your primary test data

male.txt a list of words commonly used by men

female.txt a list of words commonly used by women

background.csv additional Reddit posts that you may optionally use for training an obfuscation model

1.3 Baseline

With the default test.csv, the classifier achieve an accuracy of 64.6% for determining the gender and an accuracy of 83.2% for identifying the subreddit of a post. The task is to obfuscate the data in the test.csv so that the classifier cannot predict the gender of the authors while still being able to correctly predict the subreddit of a post. Note that in this configuration we treat the provided classifier as an adversary. We assume that you have access to the test data and to a background corpus, but you do not know the details of the classifier nor the data on which the classifier was trained.

2 Obfuscation of the Test Dataset

2.1 Random Obfuscated Dataset (4P)

First, run a random experiment, by randomly swapping gender-specific words that appear in posts with a word from the respective list of words of the opposite gender.

- Write a function to read the female.txt and male.txt files.
- Tokenize the posts („post_text“) using NLTK. (0.5p)
- For each post, if written by a man („M“) and containing a token from the male.txt, replace that token with a random one from the female.txt . (1p)
- For each post, if written by a woman („W“) and containing a token from the female.txt, replace that token with a random one from the male.txt . (1p)
- Save the obfuscated version of the test.csv in a separate csv file (using pandas and make sure to name them accordingly). (0.5p)
- Run the given classifier again, report the accuracy and provide a brief commentary on the results compared to the baseline. (1p)

2.2 Similarity Obfuscated Dataset (4P)

In a second approach, refine the swap method. Instead of randomly selecting a word, use a similarity metric.

- Instead of the first method replace the tokens with semantically similar tokens from the other genders token list. For that you may choose any metric for identifying semantically similar words, but you have to justify your choice. (Recommend: using cosine distance between pre-trained word embeddings) (2p)
- Save the obfuscated version of the test.csv in a separate CSV file (using pandas and make sure to name them accordingly) (0.5p)
- Run the given classifier again, report the accuracy and provide a brief commentary on the results (compared to the baseline and your other results) (1p)
- The classifier's accuracy for predicting the gender should be below random guessing (50%) and for the subreddit prediction it should be above 80% (0.5p)

2.3 Your own Obfuscated Dataset (4P)

With this last approach, you can experiment by yourself how to obfuscate the posts.

- Some examples: 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? What if you use different word embeddings? (2p)
- Save the obfuscated version of the test.csv in a separate CSV file (using pandas and make sure to name them accordingly). (0.5p)
- Describe your modifications, report the accuracy and provide a brief commentary on the results compared to the baseline and your other results. (1.5p)

3 Advanced Obfuscated Model (5P)

Develop your own obfuscation model using the provided background.csv for training. Your ultimate goal should be to obfuscate text so that the classifier is unable to determine the gender of an user (no better than random guessing) without compromising the accuracy of the subreddit classification task. 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). In your report, include a description of your model and results.

- Develop your own classifier. (3p)
- Use only posts from the subreddits ”CasualConversation” and ”funny” (min. 1000 posts for each gender per subreddit). (0.5p)
- Use sklearn models (MLPClassifier, LogisticRegression, etc.).
- Use 90% for training and 10% for testing. (0.5p)
- In your report, include a description of your model and report the accuracy on the unmodified train data (your baseline here) as well as the modified train data and provide a brief commentary on the results. (1p)

4 Ethical Implications (3P)

Discuss the ethical implications of obfuscation and privacy based on the concepts covered in the lecture. Provide answers to the following points:

1. What are demographic features (name at least three) and explain shortly some of the privacy violation risks? (1p)
2. Explain the cultural and social implications and their effects? In this context discuss the information privacy paradox. You may refer to a recent example like the COVID-19 pandemic. (1.5p)
3. Name a at least three privacy preserving countermeasures. (0.5p)