Gender Classification in Modern Society

# Gabrielle Beinars DSC680-T301 Applied Data Science (2221-1) <https://github.com/gbeinars>

As I began to take a closer look at the data, I realized that a majority of it was actually restricted. This was a huge oversight from my first milestone, as I had focused heavily on the discussion points of my topic, rather than confirming that the data was functional to use. My original focus was on creating a gender classification prediction model, but with three gender identity groups. Many gender classification projects use the typical genders of male and female, but this isn’t always the case in today’s society. Unfortunately, there is extremely limited data on this due to a third gender option not in place just yet. I’ve since shifted my focus to using tweets to classify a gender as male or female. I do still plan to discuss the repercussions of only considering male and female, as well as ethical challenges that should be considered when focusing on gender identity.

Thus far, I have updated where appropriate (discussed below) and I’ve begun to clean the data with basic exploratory analysis. I anticipate that the data will need to be normalized (lower-case, symbols removed, etc), vectorized, encoded, where I can then feed to a neural network. I will explore different approaches and compare, in order to determine which model has the highest accuracy.

I’ve updated the domain from the original proposal to include Twitter, which contains unstructured data. Data is provided by Data for Everyone Library on CrowdFlower (linked above). Online social media can provide great insight and better understanding of current demographics. This can allow for improved add targeting, and simply a better understanding of today’s society. Below are additional sources for this project -

1. <https://www.kaggle.com/crowdflower/twitter-user-gender-classification> (dataset)
2. <https://glciampaglia.com/docs/papers/genderdetection_chasm.pdf> (related work)

Analyzing this data will provide insight and better understanding to current demographics. In addition to the following questions included in the original proposal, I’ve added a couple more to narrow my focus.

* Are there any significant factors that better predict gender identity?
* How well does the classification model predict gender identity?
* How well do words/tweets predict gender?
* Are there certain words that predict more strongly?