**Twitter User Gender Classification**

Luke Barsotti

For my final project in my Data Science Fundamentals course, I chose to create a model to classify twitter users by their gender, based on a dataset provided by CrowdFlower on Kaggle. The original dataset contained tweets, combined with information about the profile of the twitter user (tweeter) and a gender assigned either manually, or by a classifier applied by CrowdFlower. My goal was to determine if I could use some combination of a user’s tweets and profile description to predict their gender (male, female, or brand).

The original model applied by CrowdFlower was only 60% accurate, so I did not want to trust that their gender classifications were correct. Therefore, I only used data points that had been manually classified and were 100% accurate. I also cut out any data points that whose gender was classified as “unknown.” This left me with a dataset of about 14,000 tweets, which were approximately 39% female, 34% male, and 27% brand.

I used the Naïve Bayes method to run 3 separate classifications using different combinations of tweet text and profile description. One used only the tweet text, the second used only the profile description text, and the third combined the two.

Of the three classifiers, the least accurate was the one that used only the text from the tweets. It came in with an accuracy score of approximately 57%. Next was the classifier that used only the text of the profile descriptions, coming in around a 62% accuracy score. The combination of both text fields resulted in the most accurate classifications, with an accuracy score of about 67%. These scores do not tell all, as all 3 classifiers seemed to favor classifying users as female over the other two options, and all 3 also seemed to have difficulty classifying as a brand. This could be because the training data was not split up in the best way, or might be because there was simply more female data and less brand data within the dataset.

Ultimately, it seems that my classifier has a shot at beating the original model based on my accuracy scores. However, it is unfortunately not clear since this particular Kaggle dataset was posted for exploration rather than competition. My next steps for this data set might include altering the way I train my model, or attempting to use a different type of classification, such as some type of Neural Network.