This data set was used to train a CrowdFlower AI gender predictor. [You can read all about the project here](https://www.crowdflower.com/using-machine-learning-to-predict-gender/). Contributors were asked to simply view a Twitter profile and judge whether the user was a male, a female, or a brand (non-individual). The dataset contains 20,000 rows, each with a user name, a random tweet, account profile and image, location, and even link and sidebar color.

**Inspiration**

Here are a few questions you might try to answer with this dataset:

* how well do words in tweets and profiles predict user gender?
* what are the words that strongly predict male or female gender?
* how well do stylistic factors (like link color and sidebar color) predict user gender?

**Acknowledgments**

Data was provided by the [Data For Everyone Library](https://www.crowdflower.com/data-for-everyone/) on [Crowdflower](https://www.crowdflower.com).

Our Data for Everyone library is a collection of our favorite open data jobs that have come through our platform. They're available free of charge for the community, forever.

**The Data**

The dataset contains the following fields:

* **\_unit\_id**: a unique id for user
* **\_golden**: whether the user was included in the gold standard for the model; TRUE or FALSE
* **\_unit\_state**: state of the observation; one of *finalized* (for contributor-judged) or *golden* (for gold standard observations)
* **\_trusted\_judgments**: number of trusted judgments (int); always 3 for non-golden, and what may be a unique id for gold standard observations
* **\_last\_judgment\_at**: date and time of last contributor judgment; blank for gold standard observations
* **gender**: one of *male*, *female*, or *brand* (for non-human profiles)
* **gender:confidence**: a float representing confidence in the provided gender
* **profile\_yn**: "no" here seems to mean that the profile was meant to be part of the dataset but was not available when contributors went to judge it
* **profile\_yn:confidence**: confidence in the existence/non-existence of the profile
* **created**: date and time when the profile was created
* **description**: the user's profile description
* **fav\_number**: number of tweets the user has favorited
* **gender\_gold**: if the profile is golden, what is the gender?
* **link\_color**: the link color on the profile, as a hex value
* **name**: the user's name
* **profile\_yn\_gold**: whether the profile y/n value is golden
* **profileimage**: a link to the profile image
* **retweet\_count**: number of times the user has retweeted (or possibly, been retweeted)
* **sidebar\_color**: color of the profile sidebar, as a hex value
* **text**: text of a random one of the user's tweets
* **tweet\_coord**: if the user has location turned on, the coordinates as a string with the format "[*latitude*, *longitude*]"
* **tweet\_count**: number of tweets that the user has posted
* **tweet\_created**: when the random tweet (in the **text** column) was created
* **tweet\_id**: the tweet id of the random tweet
* **tweet\_location**: location of the tweet; seems to not be particularly normalized
* **user\_timezone**: the timezone of the user