Our task is to identify the gender of Twitter users based on their language use. Previous machine learning models have successfully predicted binary gender (i.e., male vs. female) using language-related features from Twitter and other social media outlets. Our task is interesting because we would like to consider how non-discrete conceptualizations of gender may be reflected in language use. That is, we would like to explore whether language use can predict an individual’s gender identity along a spectrum of gender expression.

Historically, researchers have considered the relationship between language and gender from a perspective where gender is strictly binary. However, we know that, in reality, not every gender identity can be described as male or female. The study of language and gender should reflect this fact, and move towards being less restrictive. Currently, there is little to no work examining the relationship between language and gender from this perspective. We hope that this project will help move the field in that direction.

To address this task, we will obtain data from Twitter’s API. This data acquisition process will require a few steps. First, we must identify a large set of English-speaking Twitter users. As Twitter does not provide a gender identification field for users, we will mine this information by searching in users’ account information, specifically their biographical description, for gender identification of some kind (e.g., *male*, *mom*, *queer*, *trans*). Users that provide this information will comprise our database. For quality control, we will manually examine 1000 users from our database to ensure accuracy of our gender sampling. We will query those users and use their tweets to design features as input to our algorithm.

We will use a number of features that have been used previously in similar models predicting gender from language use. Critically, these will be language-specific features from Twitter, and will not include other Twitter behavior, such as number of followers or number of retweets. Our candidate set of features includes:

* unigrams (binary indicator of word used or not)
* bigrams (binary indicator of two-word sequence used or not)
* trigrams (binary indicator of three-word sequence used or not)
* topics (clusters of associated N-grams)
* proportion of content words used (e.g., nouns, non-auxiliary verbs, adjectives)
* proportion of words from different parts of speech used (e.g., proportion of adjectives to all words)
* proportion of words with repeated character emphasis (e.g., omggggg)

Once we obtain our raw dataset, a number of pre-processing steps will be necessary. We will filter out hyperlinks, special characters (e.g., emojis), and Twitter-specific markers (e.g, hashtags, user mentions with @). Following other work, we will also remove N-gram features that were used by less than 1% of the sample of users in our dataset, in an attempt to restrict prediction to common language.

We will attempt to use a number of machine learning techniques to build our predictive model. As a starting point, we will consider the performance of nearest neighbor classification using 10-fold cross validation. Eventually, we would like to implement algorithms used in previous studies of language use and gender, including Support Vector Machines. We will evaluate the performance of our final model on two held out test sets. One set will come from our dataset of individuals who have identified their gender in their Twitter bio. The other set will comprise a set of randomly sampled English-speaking users from Twitter. Individuals who identify their gender explicitly may use language differently than those who do not. By using two test sets, we can assess to what degree our dataset is biased due to this fact.