Linguistic Self-Expression on Dating Websites

Elizabeth Dellea and Kelly Sadwin
Faculty Sponsor: John Barr
Extended Abstract

March 10, 2016

Introduction and Background

The inspiration for this research was to discover whether generalizations about gender and language hold true in online communications. Past linguistics studies have demonstrated that women use more filler words and qualifiers (I think, maybe, just, only, etc) than men. Recently, women have been encouraged to eliminate such words from their speech to sound as confident as men.

Our question, on reading those pieces of prescriptive advice, was whether these speech patterns are as prevalent as the advice would suggest. Many of the studies cited are over fifteen years old, and speech patterns are constantly changing. Perhaps the advice is based on out of date data or, worse, a slap-dash solution to a systemic issue.

Testing these hypotheses would require text samples from online sources that are clearly tagged with gender. We considered more popular forms of social media, such as Twitter, but the gender of the user is not consistently identified. More personalized websites such as Facebook would restrict access to profiles outside of the researchers' social spheres.

The logical choice was online dating websites, where users must self-identify their gender, in addition to other demographic and psychographic information. Frequently, users also fill out a short profile describing themselves. These profiles serve as the text samples in the corpus of this research project.

Methods

The dating website selected for data acquisition was OkCupid.com. As a free site, it afforded the opportunity to make two dummy profiles (one bisexual male and one bisexual female) and examine the profiles of potential matches for those profiles.

The two profiles were listed with as few restrictions as possible on matching protocols, stripping away location and gender requirements to gain a wide variety of user data. A custom-designed web-scraping bot in Python downloaded the text of matched profiles into a SQLite database, tagging with age, gender, sexual orientation, location, and gender that the user is currently looking for in a match. Each profile in the database was given an identifying number. Account usernames were tracked only for the duration of the scraping to prevent duplicates.

Natural Language Processing (NLP) techniques provided the basis for initial inquiry. Individual linguistic features were isolated for analysis, including total word count, unique word count, and percentage of words that are adjectives and adverbs. Profiles were also examined for sentiment, which generated a score for polarity (positive or negative) and subjectivity (objective or subjective).

Additionally, profiles are categorized by tags and generated as word clouds. Words that appear more frequently across profiles are displayed in larger text than less frequently used words. This method of data visualization may provide insights regarding more specific analysis procedures to undertake.

Results and Reflections

Preliminary analysis of the resultant data suggests that the way men and women use language in this context is not as drastically different as other studies concluded. Default word clouds with only stopwords removed (e.g. the, and, of) are strikingly similar across all genders and sexual orientations. There are no immediate trends found among the isolated linguistic features recorded.

Profiles also were clustered according to machine learning techniques, but so far these experiments have not yielded conclusive results. Clustering is a process by which a computer finds similar traits among samples in a group of data and groups them together. With fine-tuning, a successful clustering could potentially separate users of different demographics to demonstrate language differences between these populations.

In our research, we used many code libraries with sparse documentation or in unexplored contexts. Through this process, we became self-reliant in fixing bugs and finding solutions. Despite not finding concrete results at this stage, research and analysis will continue, with new strategies and techniques, through to the end of the semester.