GayTwitter: An Investigation of Biases toward Queer Users in AI and Natural Language Processing Models
Ricardo Saucedo
Faculty Mentor: Kevin Scannell, Ph.D.

Natural Language Processing (NLP for short) has gained attraction for its universal applications and importance in decision making for AI technology. This investigation focuses on queer (also known as LGBTQ+) virtual communities, specifically communities found on Twitter. A virtual colloquially known community, GayTwitter™, serves as a safe space for Queer individuals to communicate in their own dialect. While this freedom is pleasant, research has revealed that Google’s NLP API is biased toward queer individuals, enough that its model can decipher “I’m a homosexual” as holding a negative sentiment. To further explore these biases, an implementation of Word2Vec, a word embedding technology, will translate tokenized words from a corpus of Tweets into values in a vector space by employing mathematical analysis. This NLP model enables us to determine words most similar to selected keys; these key words such as “queer” or other queer related terms were examined in this investigation, in hopes of demonstrating a similar negative bias towards queer users as previous research suggested. In this investigation, similarly negative and alarming results were achieved, alluding to the reality that there is a bias in modern-day NLP models, and may someday influence AI technology in making biased or harmful decisions.