Alexander Chatron-Michaud

260611509

COMP 599 Final Project Proposal

In this project, I propose to use the relatively newly available OkCupid dataset to perform classification tasks on users based on their biographical text. This dataset has full profiles from over 70000 users, where each user has written multiple paragraphs answering OkCupid’s biographical question prompts such as “What is the first thing people notice about you” and “What are some things you couldn’t do without?”. This is a data-rich and exciting problem to work on not only because of the huge amount of text available per user, but also because there are also a large number of categorical metrics for each user to perform classification tasks on, such as gender, sexuality, level of education, pet preferences, religion, age and more. Some potentially interesting results may arise if we found that we could determine a user’s level of education, for example, from the text they use to describe their personality alone.

I hypothesize in this project that certain demographic details for users can be determined solely from their writing about their personality and preferences. If this were true, there would be several potential interesting applications of this. One such example could be its use in predicting potential categorical details for users on other dating platforms such as the very popular application Tinder, with similar biographies but lacking fields for education, fitness levels, or whatever else we may be able to predict well using user biographies.

This dataset was originally controversial since it revealed usernames and a hidden “competency score” for each user based on their performance on logical puzzles in OkCupid’s quizzes. Since then, the dataset has been anonymized and no longer has this category, and has now been approved by the company. In my own research I was able to find some work done on this dataset already, though almost all of them were approaches that used the categorical information on users for prediction (rather than predicting those categories), and all of the projects that could be found did not use the text data at all. I think this is an interesting approach that could have interesting results comparable to some works which attempt to classify other forms of writing (i.e. email). It can also potentially be compared to some works who attempt to use blog text to classify the gender of the writer, except that in this case we may predict more than just the gender of the author.

An ideal approach for this task may begin with using variants of the bag of words approach, or perhaps even word embeddings. These features could then be trained using various machine learning approaches, and tested/improved using a development set, and then scored on a test set. Prediction categories with high F-1 score (or other metrics that may be used) may then be deemed viable for prediction from text alone.