Natural Language Processing Final Project Progress Report

Solomon Foster & Lazar Zamurovic

**Research Question/Statement**: We have changed the data, though not the general theme of our proposal, given the difficulty of acquiring a good corpus of threats with outcomes attached. Instead, we have decided to conduct sentiment analysis on our own phone text conversations. We are in the process of importing/cleaning data from various conversations from our mobile devices and will pursue a number of questions related to the dynamics of these conversations. Possible questions include: how positive are each of our conversation partners? How does our positivity vary by recipient? Is there a difference in language by us depending on the gender of the recipient? and more. We would like to be able to quantify the positivity/negativity of the texts we receive and send.

**Data**: We are using our own phone text conversations and some free software to download them. There **are** concerns related to our use of these data: for one thing, we are not certain about the ethics of conducting analysis on personal conversations. We think, however, that when we present data to the class we will take sufficient steps to conceal the identities of our conversation partners so as to eliminate most concerns.

For sentiment analysis, we are using a tagged corpus that provides a level of positivity for 10,000 common English words. We downloaded these data from Storylab, a project at UVM.

**Existing tools:** We will probably use standard open-source python libraries like nltk and pandas.

**Design/Coding**: We have not begun coding, though I expect we will use techniques similar to those used in our second assignment, including the creation of some ngram models. It is going to be potentially difficult to assign sentiment to slang words, as well as emojis. We will likely use <UNK> and <OON> tags and figure out ways of dealing with unseen ngrams, including smoothing. We also plan to look for a pre-built sentiment analysis corpus for emojis and/or common slang words. This will help us deal with these situations when they are encountered.

**Evaluation:** We will address evaluation using some generic ideas about our texting patterns and our relationships with specific people. We know which texts were positive and negative. If our models return a certain text to have a positive sentiment and we can clearly see that the text is negative, we can see that we have a problem. We will be able to use some existing ngram model and classification metrics to evaluate at least parts of our system.