November 8, 2021
Elliott Chen
COMP-4448

Project Proposal: Stress Classification, Linguistics, and Social Media

My final project will be to train and evaluate prediction models dealing with sentiment in online discussions. Specifically, the goal is to determine whether the text's author is stressed or not given the base texts and a variety of additional variables generated through the Linguistic Inquiry and Word Count (LIWC) analysis software provided by the dataset. The source text itself comes from a variety of threads distributed across a number of help, support and advisory boards on the popular website and online content aggregator Reddit. The dataset is sourced from Kaggle, at https://www.kaggle.com/ruchi798/stress-analysis-in-social-media/metadata.

There is a large number of explanatory variables, the vast majority of which are created as mentioned above by LIWC. These range from dealing with basic quantitative analysis (eg words per sentence, the number of pronouns, etc) to language breakdown and grammatical analysis to detection of themes in the text such as anger, health, leisure, and religion. The values in the latter set of columns are of the greatest interest to this project and are given as percentages—as in, the percentage of words in the text that relate to the given theme. More detailed explanation of LIWC is available at http://liwc.wpengine.com/interpreting-liwc-output/. Other explanatory variables in the dataset include Reddit metadeta such as the number of upvotes, which subreddit the text was found in, post_id, etc. Most if not all of the metadata, and much of the LIWC data, will not be necessary for this project. The outcome variable is a binary label detailing whether the text author was stressed or not.

Since there is such a promising number of features beyond the raw text, we plan to utilize logistic regression and decision tree based algorithms to build a classification model. However, as we also have the original text, we can use naive Bayesian analysis separately on that as well.