**SYS 6018: Systems Engineering Competition 3**

Team (Competition 3-1): Sai Bollempalli, Elizabeth Homan, Benjamin Greenwald

**What is the problem?**

This data challenge revolves around individuals’ feelings about self-driving cars. Using text from twitter, it is possible to discern emotional reactions and general sentiment toward automobile automation. The training data set has been read and each tweet has been manually assigned a value between 1 and 5, a Likert scale for sentiment (1=most negative to 5=most positive). The challenge given is to create a model that accurately predicts sentiment for future tweets (using this same Likert scale).

**Who might care about this problem and why?**

Consumer brands are often interested in market research and understanding consumer perceptions to gauge a brand’s potential success. Automobile manufacturers incur significant capital expenditure prior to launching new brands and sentiment analysis of twitter data could provide them with a better idea of consumer preferences. With this knowledge, developers could address potential customer concerns by adjusting their media campaigns, or simply reinforce their advertising with themes found across a large number of positive tweets.

Additional groups that might be interested in this problem include lawmakers, who might craft legislation to ensure their constituents’ wellbeing; potential owners of self-driving cars, whose purchasing decisions might be driven by the “hype” surrounding automated vehicles; and investors, who inherently care about the sentiment of consumers.

**Why might this problem be challenging?**

Sentiment analysis is challenging because individual words have different meanings out of context, so it is usually not sufficient to divide blocks of texts into their word components and analyze based on these isolated words. Making things even more complicated, user generated data on Twitter is not always easy to interpret as a human reader, let alone as a computer. Tweets do not follow typical grammatical rules, spelling is often incorrect, slang words are used that may not be found in any dictionary, and special characters such as emojis are not read in the same way by every programming language.

In addition to the above, the sparsity of data matrices – the frequency of a word in a tweet relative to the whole corpus of text – is largely populated by null values. This is a critical problem since, by applying statistical methods on a term document matrix, we are effectively modeling on a data frame where 99% of all values are null. Inverse document analyses and different weighing methods further add to the complexity. Even a linear regression can be hard to apply in the context of text mining. This is because of the vast number of predictors (words) generated, especially in the cases of bigram or a trigram analyses.

**What other problems resemble this problem?**

Any kind of sentiment analysis using text would resemble this problem. This could be understanding customer feelings based on a transcript of customer support call, identifying comments made by users where there are no explicit ratings, or search engines using labeled data sets to categorize emails as spam (to name a few).