Final Project Status Report 2

Xi Han, Luzhou Li, Chao Lu

1. Has your problem statement changed at all since the last status update?

The problem we needs to solve remains the same but we have changed our emphasis. Rather than focus on giving each individual word a sentiment score using a brand new algorithm, we choose to first build two models which classify text into two category: positive and negative. The first model is Naive Bayes, whose features are the most important words in the training data, since we've already got a bag of words along with their scores, we just simply find out the first fifty highest score, for example, including positive and negative. Another model is quite easy, even without the training procedure, but we need to optimize the time and space complexity and compare its accuracy to Naive Bayes. Additionally, we can pretend that we don't have any knowledge about the spectrum of sum score of positive, negative and neutral, instead using training dataset to find out the answer.

Another goal is to build a whole new word-score mapping, which requires us to use TF-IDF as one of our feature, and also to find out in the future a way to weigh different factor in order to calculate the final result. The experiment to see the accuracy is to calculate how much the original word-score mapping is different from ours after normalization.

2. What have you accomplished since the last report? Discuss code written and experiments performed.

There are so much progress being made since last report, we even try to implement the paper published on the website of this project, but we eventually abandon that idea due to its complexity. We also discuss through the whole idea behind these two goals and implement partial functionality to see whether there's problem occur when the amount of data reach a pretty high level.

3. Did you finish everything you hoped to by this milestone? If not, why? Are there adjustments that need to be made due to being behind schedule?

We think we are on the right track, are more than confident to finish everything before the deadline, hopefully we can even add more feature in the future.

4. What else do you need to do to complete the project?

we need to take account of the size of the data, and adjust our way to use data structure.