Neha Mathew
Netid: nehaam2

CS 410: Text Information Systems

Project Progress Report

1. **Tasks Completed**
    - Curated a small trial data set by manually collecting reddit comments about a course to try out the model.
    - Found and implemented good data preprocessing techniques for sentiment analysis. Since most of the text is user generated, it is necessary to clean, normalize language, and remove noise in the information to be classified. Lowercasing, punctuation removal and number removal were done. Also tried stemming. Stop word removal is already done by VADER.
    - Used the VADER model to classify the preprocessed text.
    - Used the VADER model to classify the raw text (only blank space cleaning and lowercasing was done) and found that this gave better results.
    - Started groundwork for building a simple website to host results of analysis.

2. **Pending Tasks**

    The tasks that are pending are querying data to get posts using Reddit API, Getting additional data about courses using Course Explorer API and building and hosting a simple website. A stretch goal is making a chart/graph to add to the website.

3. **Challenges**

    A question was raised about how sarcasm would be handled by the model and I haven't been able to find a good way to detect and deal with it.

    Currently I'm using google colab (to carry out sentiment analysis) but I don't know whether/how I can use it to host the website or if I should switch to working locally.

    Learning how to build a website and connect it to the backend is taking longer than expected!

Neha Mathew
Netid: nehaam2

Answers to previous questions from reviewers for better understanding

One thing I thought of is what will happen to the analysis if the class had reviews that were good, but something about the class changed (ie. different professors, class gets restructured), so the reviews changed at some point.

I was thinking of adding the date/year that the post was made to the final output along with the sentiment. The users would still have to use this to cross reference and find what aspects (if any) of the class has changed.

Are you building on top of the prior year project or implementing it from scratch again with extensions?
I will be implementing it from scratch with extensions.