Homework 5 – updated by Gates on 4/30

**Training/Evaluation Data Acquisition Through AMT**

Use the documents that you gathered in HW1,

**NOTE1: You do not have to use the docs or data from HW1. You can find, use, or create any data you wish for this assignment.**

and use AMT platform to request five sentiment labels per document. Note that you can set up your tasks in different ways, such as hiring five workers, each annotating all comments, or hiring 10 workers, each annotating half of the comments, etc.

**NOTE2: Class – the goal here is to realize that sometimes you need labeled data, but do not have it. This happens in real life very often. I suggest that you find, use, or create 6 small doc – three (3) positive and three (3) negative – such as movie or restaurant reviews. Make 4 of them really easy. For example, an easy one that is positive would have words like good, like, love, great, excellent, etc. or for negative, words like bad, poor, dislike, etc. Then, make two of the six ambiguous - such as using sarcasm, etc. These are just ideas. You may use whatever you wish. The goal is ot create something that makes sense to you, so that you understand the results.**

**Next, you will practice using something new. Amazon Mechanical Turk (AMT). Learn about it. Google it. Then use it on your 6 (or so) docs. Did the “Turkers” labels correctly? What did they get wrong? How long did it take? How much did it cost? Etc.**

After obtaining the labels, calculate pair-wise Kappa values among the AMT workers, and then calculate the average Kappa value as the overall agreement among the AMT workers. Also calculate the Kappa agreement between each AMT worker's annotations and your manual annotations. Based on these calculations, discuss the AMT workers' annotation reliability.

**NOTE 3: Use YouTube to learn how to calculate Kappa. The math is easy and so this is a great opportunity to practice taking the lead and learning something new without a Guide.**

Describe in your report:

**NOTE 4: Please use the Assignment Format. Remember – the topic is whatever the dataset is about. The Analysis is the AMT and labeling part. The Results are what happened. The conclusions might include whether a restaurant owner (if you data is about restaurants for example0 can use this method to better understand the sentiment of 1 million reviews, etc.**

**Also include the following in the appropriate location in the format assignment….**

* 1. The experiment design: How many turkers do you aim to hire? What is the workload and payment for each turker? What is your requirement for the turkers (language proficiency, geographical location, past work performance, etc.)? Please also explain why you think this is the best choice for your experiment design to obtain the best-quality data in the most efficient way, e.g., your spam-control strategy.
  2. The experiment outcome: How long did it take to obtain all labels? How much did you pay in total? Did any unexpected events occur during the process? Did you find any spammers? If yes, how did you find out and remove spam data? What is the average Kappa agreement among the workers? What are their levels of agreement with your ground truth? Do all AMT workers share similar marginal distributions?
  3. Conclusion: Do you think AMT is a viable approach for obtaining training labels? What lessons did you learn in this experiment?

~~Attach a spreadsheet with the following seven columns: original comment, your label (ground truth, “positive,” “negative,” or “neutral”), and AMT response #1, #2, #3, #4, #5.~~

**~~Submission format:~~** ~~Same as HW1 (up to 4 pages, at least 12-point font, 1'' margin on all sides)~~