Jeffrey Levesque

jlevesqu@syr.edu | Text Classifier Evaluation and Human Annotation

IST-736: HW5

PRofessor gates

**Introduction**

Today, an abundance of available data has encouraged many existing statistical methods to apply predictive and prescriptive forecasting. Classification algorithms often require train data to include target labels, many times consuming to construct. However, crowd sourcing tools such as Amazon Mechanical Turk (MTurk), allow individual(s) to upload their dataset, then pay another to label the rows of the data for small price. This allows data scientists to worry less on data acquisition, and more on analysis and reporting.

While MTurk has been around since 2005[[1]](#footnote-1), many research organizations have either not heard of the platform, or not considered utilizing the tool. Within the platform, requesters can hire multiple workers, or restrict the type of workers based satisfied credentials[[2]](#footnote-2). This allows requesters to selectively restrict who can work on the given dataset. Furthermore, the idea of a “small price” is relative on the size of the dataset. Though careful inferencing could potentially reduce the scope and size.

**Analysis**

While MTurk provides many great functionalities, a cross-sectional analysis between will be made between each worker, and against the overall ground truth. Three MTurk workers will be hired, at $0.02 per task, tasked to label sentiments (positive, negative, or neutral) for 3236 rows of a given dataset. This equates roughly $48.15. Furthermore, a pairwise Kappa score will be conducted for each of the following cases:

* Ground Truth vs. MTurk #1
* Ground Truth vs. MTurk #2
* Ground Truth vs MTurk #3
* MTurk #1 vs MTurk #3
* MTurk #2 vs MTurk #3
* MTurk #2 vs MTurk #3

In general, the Kappa score will be a measure of (observed accuracy – expected accuracy) / (1 – expected accuracy). Using a contingency table, the sum of the diagonals divided by the sum of all cells constitute the *observed accuracy*. Furthermore, the *expected accuracy* can be calculated by summing of each column total divided by the total sum, then dividing the entire result by the total sum. In general, the kappa scores are measure of agreement between two observers. In this study the kappa score will be used to compare each of the above cases. Scores > 0.75 indicate excellent results, 0.40-0.75 moderate results, and > 0.40 poor[[3]](#footnote-3).

**Data Preparation**

Data was acquired by utilizing the twitter api, executed through custom python code[[4]](#footnote-4). Specifically, a single twitter username “taralibinksi” (former U.S. Olympic gold figure skater) was collected for a duration limited by two factors:

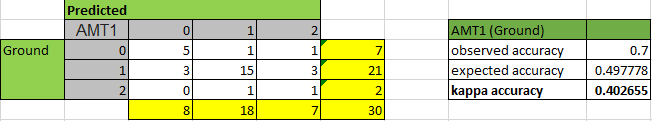
* Rate limit: number of api requests
* Count: maximum number of tweets to collect

The twitter api limits the number of requests to 900, and the count of tweets per request 200 per 15-minute interval[[5]](#footnote-5). Therefore, these upper limit thresholds were used[[6]](#footnote-6). However, during initial data gathering, the twitter api by default truncates returned tweets to 140 characters[[7]](#footnote-7). To extend the functionality for obtaining the full length of each tweet, tweet\_mode='extended' needed to be provided to the api[[8]](#footnote-8). However, this limitation was discovered shortly after the MTurk workers completed corresponding tasks. The corresponding csv file was stored locally[[9]](#footnote-9), as well as the associated MTurk batch results[[10]](#footnote-10), which was converted to a reduced form[[11]](#footnote-11). Moreover, each of the kappa scores was computed within the csv spreadsheet.

While the initial motivation of this assignment was to potentially prepare a dataset for an upcoming study[[12]](#footnote-12), the scope of this assignment was greatly reduced. Specifically, the MTurk workers completed 3236 tasks. However, only the first 30 were utilized for this assignment. This nearly matches the dataset size (50 rows) from the first assignment[[13]](#footnote-13).

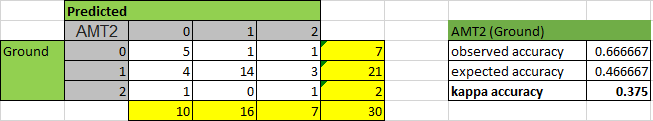
**Results**

The ground truth was compared against the prediction from the first mechanical turk worker. The Kappa score of 0.402 indicates a moderate result:



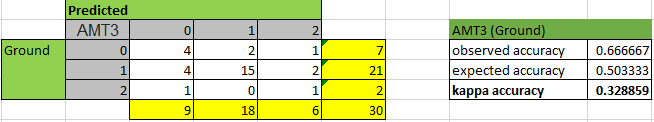
**Figure 1:** first amazon mechanical turk prediction against ground truth.

The ground truth was compared against the prediction from the second mechanical turk worker. The Kappa score of 0.375 indicates a poor agreement:



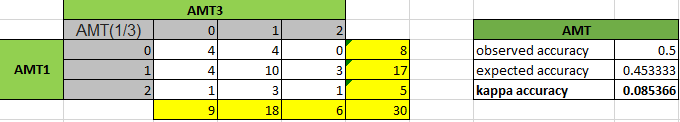
**Figure 2:** second amazon mechanical turk prediction against ground truth.

The ground truth was compared against the prediction from the third mechanical turk worker. The Kappa score of 0.328 indicates a poor agreement:



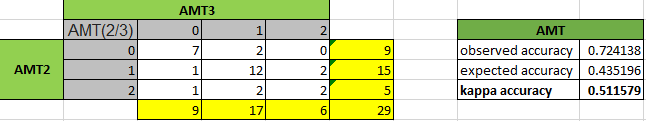
**Figure 3:** third amazon mechanical turk prediction against ground truth.

The prediction from the first mechanical turk worker was compared against the third mechanical turk worker. The Kappa score of 0.085 indicates no agreement:



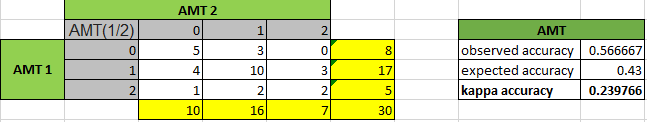
**Figure 4:** first amazon mechanical turk prediction against third amazon mechanical turk.

The prediction from the second mechanical turk worker was compared against the third mechanical turk worker. The Kappa score of 0.511 indicates good agreement:



**Figure 4:** second amazon mechanical turk prediction against third amazon mechanical turk.

The prediction from the second mechanical turk worker was compared against the first mechanical turk worker. The Kappa score of 0.239 indicates no agreement:



**Figure 5:** first amazon mechanical turk prediction against second amazon mechanical turk.

The above results indicate that second and third amazon MTurk workers have the best inter-agreement levels, while the first with second, and first with second have worse agreement levels respectively. Furthermore, the third MTurk worker had the worse agreement level with the ground truth at 0.085, while the other two MTurk workers roughly equal at values under 0.40. Based on the given results, the MTurk workers did not sufficiently suffice acceptable Kappa values. However, limitations several limitations described earlier may have contributed to poor kappa values. Specifically, the original tweet dataset obtained from the twitter api truncated numerous tweets. This provided a level of confusion to MTurk workers. However, the ground truth had the benefit of reviewing tweets directly from the twitter website. Furthermore, using 30 row entries may not be enough to derive context, such as Kappa values for pairwise inter-agreement.

**Conclusions**

Amazon Mechanical Turk (MTurk) allow research studies to utilize unlabeled data, by crowdsourcing the labeling responsibilities. During this process, any number of workers can be selected for hire. Prioritization can be established by incentivizing the rate of payment per task. Furthermore, overall scores can be improved by incorporating multiple workers and averaging the total scores. Additional benefits of MTurk include selectively choosing workers by any number of criteria’s, including geography, age, education, or professional status.

Though results in this study have not shown great results, earlier analysis have stated that poor data submission to MTurk has made numerous tasks ambiguous. To fully maximize the benefits of crowdsourcing, it is imperative to submit a dataset ready to be reviewed. Failure to perform this task can be cost inhibitive, with poor data results. Moreover, rather than submitting the full original dataset, a careful inference strategy could be beneficial, at least initially.

Overall, the MTurk concept is greatly beneficial. The use of the crowdsource platform is not new and currently being used by large and small organizations alike. While labeling data can be better controlled if performed by an internal research group, consideration should be based on optimal time management. Furthermore, the larger the data, the better the case for utilizing platforms like amazon mechanical turk. Research groups should ideally focus on deriving research findings, and the best way to tell the corresponding story.

1. <https://en.wikipedia.org/wiki/Amazon_Mechanical_Turk> [↑](#footnote-ref-1)
2. <https://github.com/jeff1evesque/ist-736-hw/blob/master/hw5/resources/mechanicalturk.pdf> [↑](#footnote-ref-2)
3. <https://github.com/jeff1evesque/ist-736-hw/blob/master/hw5/resources/Interrater_agreement.Kappa_statistic.pdf> [↑](#footnote-ref-3)
4. <https://github.com/jeff1evesque/ist-736-hw/blob/master/hw5/assignment.py> [↑](#footnote-ref-4)
5. <https://developer.twitter.com/en/docs/tweets/timelines/api-reference/get-statuses-user_timeline.html> [↑](#footnote-ref-5)
6. <https://github.com/jeff1evesque/ist-736-hw/blob/8d4dfa3e4a6f9369bc42479125193e54caf9e683/hw5/assignment.py#L50-L51> [↑](#footnote-ref-6)
7. <https://stackoverflow.com/a/47393913> [↑](#footnote-ref-7)
8. <https://github.com/jeff1evesque/ist-736-hw/issues/97> [↑](#footnote-ref-8)
9. <https://github.com/jeff1evesque/ist-736-hw/blob/master/hw5/data/twitter/taralipinski.csv> [↑](#footnote-ref-9)
10. <https://github.com/jeff1evesque/ist-736-hw/blob/master/hw5/data/mturk/batch_results.csv> [↑](#footnote-ref-10)
11. <https://github.com/jeff1evesque/ist-736-hw/blob/master/hw5/data/mturk/batch_results_reformatted.csv> [↑](#footnote-ref-11)
12. <https://github.com/jeff1evesque/ist-736> [↑](#footnote-ref-12)
13. <https://github.com/jeff1evesque/ist-736-hw/blob/master/data/sample-sentiment.csv> [↑](#footnote-ref-13)