Name: Morgan Gere

Course: IST 736- Text Mining

Date: 5/13/2022

**Training/Evaluation Data Acquisition Through AMT**

**Introduction**

AMT (amazon mechanical turk) is a “crowdsourcing marketplace” (*What Is Amazon Mechanical Turk?).* It allows a requester to place something (text, image, video) and have a worker complete some set tasks for it. There are many ways this can be used with many features that can be selected. The goal is to create microtasks that the workers can preform on their own from anywhere with an internet connection (*What Is Amazon Mechanical Turk?).*

As an investigation into the workings and to gain experience using AMT, tweets collected from homework 2 about Artificial Intelligence (AI) will be put up for workers to annotate into 5 levels of sentiment. This will be compared to an annotation “Gold Standard” created by manual annotation of the same data. 5 non-master annotators (workers) and 5 master annotators (master workers) will be used.

Cohen’s kappa score (k-score) will be used to find pair-wise score and an average will be found. The scores will be used to compare workers with master workers as well as workers and master workers vs the created gold standard. Kendall’s Tau score (t-score) will be used to compare the master workers with one another. The t-score will be used to find an average score to see how strongly the classifications are correlated. Finally, the t-score will be used with each master worker and the gold standard.

**Method**

Data

From the mongo database created in homework 2 the first 50 tweets were collected and placed into a pandas data frame. The \_id column from the mongo database was dropped and the data frame was exported into and csv file.

Graphical user interface, text, application, email

Description automatically generated

After data was manually annotated for the gold\_standard sentiments and 10 coders were acquired from AMT and their annotated sentiment classifications were manually added to the CSV file. This file was then read back into python using pandas and placed into a data frame.

Text

Description automatically generated

Acquisition from AMT

Two separate batches were placed on the AMT’s website. One was for 5 worker and one for 5 master workers. Both were the same excluding this. The reward for each line annotated was selected to be $0.05 plus the cost of using AMT ($2.50) The task was selected to allow for 2 hours of time and only be available for 7 days. The sentiment labels to be used were given with a brief explanation. “Very Negative” - The Sentiment towards AI is against it, evil, world ending. “Somewhat Negative”- The Sentiment towards AI is not good. example this AI isn’t working as intended. “Neutral”- The Tweet is either a fact statement or not talking about AI. “Somewhat Positive”- The Sentiment towards AI is not bad; example AI is convenient. “Very Positive”- The Sentiment towards AI is for it, Amazing, it’s the future. These labels were also given with the instructions to examine with Artificial intelligence in mind. The header was selected to say, “What sentiment does this tweet convey towards AI?”

The time was selected to make sure the person doing the annotations was attempting to do it in one sitting. If they were to stop and come back later, it is possible for there to be large difference in their mood which could change how they rate sentiment. It is hard to account for spammers or people that are not taking the annotation seriously without getting very specific on the variables. As time was a perceived issue not many options were selected. This was to make sure that enough time remained to accomplish the project. This is ultimately why two sets were chosen to be able to have a good comparison of workers vs master workers.

Data Preprocessing

After the data frame was created the labels needed to be transformed into numeric representations before any scores could be generated for comparison. Each column of the data frame was converted into a list and run through a function that replaced the label with a numeric representation. An example is “Very Negative” was replaced with 0.

Text

Description automatically generated

Pairwise Kappa and Average Kappa Score

Using cohen\_kappa\_score from sklearn each worker was run with the other four workers to get pair-wise kappa scores. The mean of these scores was obtained using numpy. This same process was done using only master workers.

Text

Description automatically generated Graphical user interface, text, application

Description automatically generated

AMT pair-wise kappa with “gold standard”

Using cohen\_kappa\_score from sklearn each worker was run with the gold\_standard. This process was repeated for master workers.

Text

Description automatically generated Graphical user interface, text, application

Description automatically generated

Combined Kappa

The cohen\_kappa\_score from sklearn was used with each worker and master worker combination. Numpy’s mean function was used on these new scores and all the pair-wise kappa scores obtained previously.

Text

Description automatically generated

Master’s Kendall Tau Scores

A function was created to print out only the score obtained from scipy’s stats (Kendall’s Tau Score). The function was used on each possible combination of master workers. An average score was obtained using numpy’s mean function. Each master worker was placed in the function along with the gold\_standard. One worker, found to have a high kappa score, was also used with the gold standard to find it’s t-score.

Graphical user interface, text, application, chat or text message

Description automatically generated

Graphical user interface, text, application

Description automatically generated

**Results**

Acquisition from AMT

In total 10 workers were hired. Each worker annotated the entire 50 tweets. The total cost was $2.50 dollars for each worker and $5.00 was paid to AMT directly. A total of $30.00 was spent on both sets of annotators. It took around 24 hours for each set to be returned.

Average Kappa Scores

The average kappa score with workers was found to be around 0.0003 while the master workers average kappa was found to be around 0.092.

AMT pair-wise kappa with “gold standard”

The pair-wise kappa scores of the workers and the gold\_standard are:

A screenshot of a computer

Description automatically generated with medium confidence

The pair-wise kappa scores of the master workers and the gold\_standard are:

A screenshot of a computer

Description automatically generated with low confidence

Combined Kappa

The combind workers and master workers kappa was found to be around 0.03.

Master’s Kendall Tau Scores

The overall average Tau score was around 0.16 and the Tau scores compared to the gold standard are:

Text

Description automatically generated

**Conclusion**

Kappa Scores

When looking at both the workers and master workers average kappa scores they are both very low. This indicates that the data is not being processed properly or at least not the same by each worker/master worker. A good kappa score will be around a 0.7 with a 0.9 being preferred. Looking at the annotator pair-wise kappa scores with the gold\_standard within the workers one score is elevated at 0.33 this is still rather low but since the classifiers are ordinal its possible, they are just placing something that is “Somewhat Negative” in “Neutral” or similar classification differences. Meaning these annotations could still have been produced properly. This worker had his Tau score produced. The remaining workers would not be used. While looking at the master workers gold\_standard pair-wise kappa scores they are higher ranging from around 0.15 to 0.34. Any of these could be the same as the one found with the highest score in the workers kappas. To examine if this is possible the Kendall’s Tau score was used.

Master’s Kendall Tau Scores

For Kendall’s Tau score the overall value on the master workers is a weak correlation. Yet when looking at the individual scores compared to the gold\_standard, there are two (0.25 and 0.27) which had a moderate correlation and one (0.43) that exhibits strong correlation. These three annotators would be selected from the masters. The worker with the high kappa score was also run with the gold\_standard to see how it’s Tau score would fair. It was found to be 0.37 which is a strong correlation and would be used as well.

Future Acquisition

For future uses of AMT the understanding of how quickly the results can be returned will change the way of going about designing the experiment. Since the time was perceived limited, less features were selected, more would have been ideal. Using master and non-master workers allowed for some insight into spamming and the level of work being completed that would not have been gained firsthand otherwise. More time needed to be invested into the actual process of setting up the AMT’s page. Yet it was also found that this can be avoided by using master workers and making sure that more than one annotation is used along with a small number of test questions and either Cohen’s Kappa, Kendall’s Tau, or Pearson Correlation to evaluate if the worker is completing the task properly.

If AMT is utilized properly, it can produce good results. It requires some amount of manual annotation before hand to use as a comparison. This is required because not all the workers will participate in the project as designed due to many possibilities. These include lack of understanding, trying to spam for pay, lack of knowledge and many more. While there are features that can be selected to help reduce these effects, they are not guaranteed.

**Citation**

*What Is Amazon Mechanical Turk? - AWS Documentation*. https://docs.aws.amazon.com/AWSMechTurk/latest/AWSMechanicalTurkRequester/WhatIs.html.