**ASSIGNMENT 7**

**Training/Evaluation Data Acquisition Through AMT**

**Introduction:**

For machines to perform sentiment classification task on vast amounts of text data, the first and foremost task is to label the data into different categories as required and then come out with the sentiment. Labeling the text is the most daunting task after which models application and accuracy can be easily calculated. For machines to correctly annotate text is a challenging task when the text can be interpreted in different ways. For this purpose, we use Kappa Value between the annotators to see how well they agree on labeling of the text data. Kappa is a method to find the level of agreement between the annotators for the same text. This value is in range from 0 to 1 (decimal) or 0 to 100%. The higher the Kappa value, the more the agreement between the annotators helping to reduce misclassification of labels and correctly annotating for higher accuracy. Amazon Mechanical Turk is an online platform where fast and cost-effective annotations of the text data can be achieved.

The goal of this assignment is to use the Social Media data to find the sentiment of public towards AI and calculate the Kappa Values for the same text between my annotations and the annotations provided by Workers from Amazon Mechanical Turk. This task aims at finding out level of agreement between me and other AMT workers as well as agreement between the AMT workers themselves for the same text.

**Data Definition:**

The Data used for this task is the same data which was collected for HW1 using Web Scraping. The data is Tweets collected from Twitter for analyzing public sentiment towards AI on social media. The tweets were cleaned for @, # and other issues earlier itself. For this task, out of all the tweets, 50 tweets were used to be annotated by AMT workers. The data containing 50 tweets was in a .csv file which was uploaded on AMT for workers to annotate. The ground truth according to my annotations was not included in the .csv file. The tweets had mixed annotations among the 3 categories defined positive, negative or neutral for each tweet. There were 35 positive tweets, 5 negative and 10 neutral as per the ground truth.

**Methods:**

The 50 tweets with their ground truth was to be compared with annotators labeling. For this task, I requested annotations from 5 different workers for each tweet. I registered myself on AMT platform as a requester using my Amazon account which is needed for getting the annotations done. Then project was created by selecting Sentiment Analysis section. The .csv file was uploaded with all the qualifying specifications defined for workers who can take up this task along with payment for each worker for each tweet. Before starting their annotations, the workers were provided sample of tweets for each category for them to understand what I think is positive, negative or neutral. 1 example for each of the 3 categories were defined for them to be easily do the task based on my understanding of the labels. After the annotations were completed by the workers, I approved the annotations and downloaded the .csv file to view the results. Calculated Kappa Values between the ground truth and each of the worker pairwise using Python library cohen\_kappa\_score from Sklearn package where for positive, negative and neutral categories were substitutes as 1,2,3 respectively and then pairwise calculation is done. Also, I calculated Kappa values between the 5 workers pairwise to see how the different workers annotated the tweets.

**Results:**

**Experiment Design:**

For completing this task, I hired 5 workers from the AMT. The 5 workers who were required to complete this task were paid $0.01 for each tweet annotated. Each worker has 50 tweets to work on. Initially, I thought that same 5 workers would take up the task of annotating the tweets but after the results were published, I figured out by comparing the WorkerID’s Column that for each tweet there can be different workers doing the annotations. Not all the tweets were annotated by same 5 people as was visible from WorkerID. Before the annotation was published for the workers to start working, I specified certain requirements for a worker to match for him/her to be qualified to take up this task. The location of the workers was specified to be from USA or India. This requirement was specified as the tweets contained AI sentiment which was better interpretable by people from these countries. The technology level of the 2 countries is very high to be able to understand the sentiment better and thus would produce more accurate labels and a higher kappa. The proficiency was defined to be English as all the tweets are in English and a person fluent in English would do better at figuring out the sentiment by considering negation, words with different meanings, context etc. The workers were not required to be masters for the reason of finding if someone is not a genuine worker. This experiment was also carried out to see the performance of workers who are not yet masters and if they can do the tasks with same efficiency or not as Master Workers. There were other options which I didn’t try such as Occupation criteria, Income etc. for all people to understand basically what is being said by referring the example of what is what for all the 3 categories in general.

**Experiment Outcome:**

The labels along with 1 example of each category was defined for the ease of workers to work on. After creating new project sentiment analysis, publish batch of 50 tweets as a .csv file, it took around 2 hours for the completion of task. The task initially was being completed very fast with around 60% of it done in 20 minutes but later the speed was reduced. For the last 5 to 10 tweets, it took 1 hour to be annotated. I paid total of $5 for the whole annotation task. For each tweet each worker got $0.01 which came to $2.5 and Amazon took fees of $2.5 for using their platform. The fees amazon takes varies based on the size of data and the amount being paid to workers. Also, for selecting variety of requirements for a worker to qualify, the fees are different. Some of the annotations where I found the workers couldn’t figure out the sentiment were “How AI Is Moving Right Onto Devices”, this tweet every worker classified as Neutral but for me this is Positive as they didn’t figure out well that AI is moving onto devices means something positive which would save time, give better user experience rather all of them saw it as a simple sentence being said as a fact. I used this sentence at the end “This futuristic bus is 100% driverless and fully electric AutonomousVehicles” which seems positive but none of them seem to pick it up and all said it is neutral except 1 who annotated this as negative. For this case, the simple interpretation for Worker 1 saying negative was completely opposite to what was expected label. There were no spammers found as I put some factual general tweets to test this situation also. This sentence was one used for this test “Xenobots"" are programmable ""living robots"" created from the skin and hea cells of frogs” which conveys the fact that robots are made from skin of frogs just a simple fact and they can figure out it out properly. The average Kappa for the workers among themselves came out to be around 0.32 where the average Kappa for workers compared to my ground truth came out to be around 0.304. There seem to be higher levels of agreement between the workers of AMT than agreement between the ground truth and then taken as an average of Kappa Values. The AMT workers share similar marginal distributions classifying tweets almost in each category to be roughly the same between them and not the ground truth along with the misclassifications.

The Kappa Values between the ground truth and workers pairwise can be seen in the below table.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Me & W1 | Me & W2 | Me & W3 | Me & W4 | Me & W5 |
| Kappa Values | 0.22 | 0.29 | 0.27 | 0.49 | 0.25 |

The table for Kappa Values between the Workers pairwise can be seen in the table below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | W1 | W2 | W3 | W4 | W5 |
| W1 | 1 | 0.10 | 0.47 | 0.28 | 0.29 |
| W2 | 0.10 | 1 | 0.42 | 0.32 | 0.42 |
| W3 | 0.47 | 0.42 | 1 | 0.44 | 0.33 |
| W4 | 0.28 | 0.32 | 0.44 | 1 | 0.48 |
| W5 | 0.29 | 0.42 | 0.33 | 0.48 | 1 |

Worker 4 and the ground truth seem to be more in line with each other. The level of agreement for the ground truth and Worker 4 is high as compared to other workers. Almost, half the time there is a agreement between the worker 4 and the ground truth. The level of agreement between Worker 5 and Worker 4 is observed high for pairwise Kappa for AMT workers. Worker 1 and 2 have the least level of agreement between them based on the Kappa Values which can be used to see the differences between them on how they encode the labels.

The number of classifications for each category for ground truth and 5 workers is also calculated to see what type of worker the person is. Neutral or Polarized Worker based on number of classifications in each category for that worker.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Positive | Negative | Neutral |
| Ground Truth | 35 | 5 | 10 |
| W1 | 19 | 4 | 27 |
| W2 | 27 | 6 | 17 |
| W3 | 23 | 4 | 23 |
| W4 | 24 | 5 | 21 |
| W5 | 22 | 4 | 24 |

The workers tend to give more neutral reviews on an average than positive. The tweets in this data are 70% positive and 20% neutral whereas for AMT workers this is almost 50% neutral meaning Neutral Sided Workers.

Marginal distribution or chance agreement is less for Actual Neutral being Predicted as Positive for Ground truth vs Workers than among the workers i.e. False Neutrals more for pairwise Workers vs Workers and Ground truth based on the confusion matrix. Predicted Neural is high for Ground Truth vs Workers than between the 5 workers. More wrong neutrals are predicted which are actually positive making the positives wrongly classified as is the case for Ground Truth. Marginal distribution varies a lot among the workers and workers vs Ground Truth.

The AMT platform is good option to get large amounts of text data annotated in a very fast and cost-effective manner. Other text mining like surveys etc. tasks can also be performed using AMT. The Kappa values for non-master annotators is very less which might be high when master annotators come into play. Marginal distribution has a noticeable difference and does help to see the differences that the annotators have between them and their agreement/disagreement on the texts.

**Conclusion:**

The tedious task of getting huge amount of text data annotated in a limited time and cost with good reliability can be achieved through proper worker and task requirements specification on AMT. But if large number of qualifications for the annotators are desired, it can be costly as Amazon charges almost equal amount of fees being paid to workers, so we need to specify accurate number of worker qualifications based on task at hand and the desired outcome. The results take very less time to be generated. Marginal distribution is noticeable in this task. Positive tweets being mistakenly classified as Neutral is the case for the most parts. The thin line between the positive and neutral is not properly interpreted which can be improved by getting more test cases annotated by masters with other qualifications such as age group defined, education level or occupation who change the Kappa Values to be better. Raw agreement can be high and is for most of the times, but the marginal distribution is equally important which shows noticeable variation between the workers and the ground truth taken pairwise where they don’t agree. Bias among annotators can be seen based on if they give more neutral/positive/negative labels for a large data to say if they are Polarized or Neutral Coders. With adequate qualifications and enough of text data along with train examples, would produce better results for a huge data to be annotated. It also helps to see subtle differences in the way different people annotate the same data which can also lead to sentimental analysis among the workers participating in the task. This task helped to see how the texts can be labelled differently which can result in different sentiments figured out from the same data if we observe it minutely.