Homework 5

Kyle Walter

11/7/2021

# Introduction

Today and as each day goes by it seems like the world is being filled with more and more data on absolutely everything under the sun. However; there are areas where the the various pieces of information may be missing. Often times categorization may be missing or incomplete and the need to complete it before the rest of the work can conintue on.

There are methods that can do this. One way is to explore other similar datasets and use them as starting point for classifying text. Another is to used unsupervised Machine Learning and look at the groupings that the data ultimately returns. Lastly, and most time consuming is to use actual manually work to classify the data prior to being consumed for other purposes.

The last can not only be time consuming, but also quite expensive. There are however; projects that have gone on at mass scale for some time utilizing human experts doing some this work. One such example is the Penn Tree Bank used in the Natural Language Processing community for their sentiment and word sense libraries.

There are some version of crowd sourcing that has happened since, captcha for one pops up when a user is logging into some internet based accounts. It usually asks the user to look at pictures for some type of object and has collected information to help Libraries digitize old text and for AI to recognize common road obsetcales.

On a different scale Amazon has introduced a service called Mechanical Turks which helps align human users with tasks which can be setup to facilitate human review. The question is since these people are not experts in a domain, are they able to correctly classify and how long would it take to get the information needed?

# Analysis

For the Analysis a tiny subset of a data set of reviews from Amazon that was posed on Kaggle was used. The Original analysis centered on a sentiment analysis. Except the data is tagged with sentiment, the postualization is that the start rating represents the sentiment of the review, with a review of 1 and 2 stars being negative, 3 stars being neutral, and 4 and 5 stars being positive.

The data set has 3 million records and with Amazon Turks being a paid service, there is some cost associated. In addition, since experts in sentiment are not one of the criteria that Amazon allows for selection, though they do specify sentiment tagging as one of the NLP categorization tasks. In order to measure the postulate, a subset of 30 records was chosen for the purpose of classification. The choice was done in Python using pandas’ sample function, given 3 stars a weight of 1 and the other starts a rate of .45. This lead to an even balance of proposed positive, negative, and neutral classes. In order to allow for replication with the same results a random state was provided of 37. The number takes advantge of a computer’s pesudo random. As long as another researcher use the same seed number with the data they same results will be seen.

## 3 10  
## 4 6  
## 1 5  
## 2 5  
## 5 4  
## Name: Sentiment, dtype: int64

Now that all three classes were balanced, the next step was to apply the data to the Mechanical Turks site for the application of tags. Doing a little research online, on how much to pay for tag, the generally sentiment was between 2 and 5 cents. As the data set was small and choice quite limited, 2 cents was decided on.

The Other question that needed to be answered was around how many Turks to hire. The decision was around 5 Turks, as 5 people per question would hopefully lead to consensus on what the sentiment should be calibrated to.

With this set the data set was loaded into Turks and the cost of 4.50 was paid, the 2 cents by 150 sample (the 30 were each supplied 5 times) and amazon has a mandatory fee for facilitating the platform.

Turks allows for a few requirements to be set. There are some basic ones like language proficiency. English however was not a select-able requirement, though one that would have been used as a non-native speaker may not understand the sentiment. Since these were Amazon reviews, location likely would not have mattered as Amazon would likely show them to anyone in the English Speaking world even if they were buying the same product on the other side of the world. It also would not make sense to request a specialty as people at all levels would be reading the results when selecting a product from the site.

Once submitted a waiting game started. Given the small data set, it seemed to go quite quickly, 48 minutes later Amazon Mechanical Turks site had sent an email informing that process had completed and the tags assigned. The Data set was approved and results downloaded.

Amazon’s system allowed the Turks to tag with the written word samples so the the Data set of stars needed to have them added. Using python the tags were applied to a new column in the sample data set. Unfortunately the the questions were returned out of order.

In order to fix the problem the sample data set was exported with the index. Since the data set was, this set was also used for manual sentiment tagging. Reviewing each of the questions in the set since it was small would act as another measure for the model to see how well the sentiment taggers did.

Now that the data was manually tagged and with index it was brought in with the Amazon Data. The Amazon Data was pivoted so the scores for Positive, Neutral, and Negative could be seen for each question and Positive and Negative of 3 were used to as the score decider, those not meeting the threshold were deemed as neutral.

The data was then merged with the manual tags using the question text and sorted by the index to put it back in the order of the original samples.

One last check was done to see how the Turks and manual tagging produced. Here are the results of the Turks

## Positive 15  
## Negative 9  
## Neutral 6  
## Name: AMTTag, dtype: int64

and the results of the hand tagged sentiments.

## Negative 15  
## Positive 9  
## Neutral 6  
## Name: R Response, dtype: int64

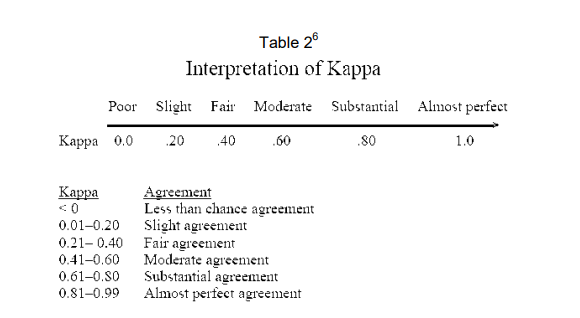
As already seen there is dis balance between all three sets. The Kappa model should help determine how usable the results are.

# The Models

## Kappa Model

The Kappa model is a way to check if two humans complete a task how well they fair against each other. It calculate the variance and takes into acount based on the number of choices how much random chance exists in the observation. The resulting output is a number between -1 and 1. With one being total agreement between the two comparison, 0 being any matches are purely random chance, and anything less than zero meaning matches are less than that of random chance.

The order of results have to be aligned for the math to work out correct, hence in the analysis section the requirement for the index to order the results from Amazon back in the same order as the samples.

Between 0 and 1 a series of ranges have been established to help understand how good results are between measurements. 

Comparing first the results of the star rating vs the sentiment determined by the Turks. The resulting Kappa score was 0.3. While this was fair agreement, it doesn’t bode well for the purpose comparing the results of the Turks vs Professionally tagged sentiment.

The same exercise was run against researcher tagged sentiment vs the Turks tagged sentiment. And a Resulting 0.394 was resulted. Pushing the result closer to the moderate agreement but not enough to drive it into the next category.

## Kmeans

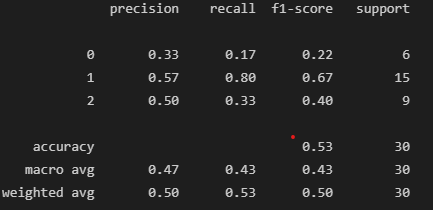
Along with using human classification, Data Scientists have another ability to look for groupings using unsupervised machine learning. kmeans clustering is one such technique, where the algorithm attempts to construct groups based on the distance between variables in the data. Similar to other machine learning techniques, except it is not trying to figure out how to reach a given tag, instead just a number of clusters that need to be further analyzed for review.

For the Kmeans, TFIDF was completed on the text of the samples. This converts the words to count frequency and then looks for words that appear less frequently across the Corpora of reviews and gives them a higher weight. As part of the transformation, the vecotrization does a few other techniques such is single word Tokenization, lower casing of all tokens, and the removal of stopwords, which in theory the TFIDF should minimize if not removed.

At this point placing the vectorized text into the kmeans algorithm, and requesting 3 cluster and setting the random state seed again at 37, since the starting point of the centroids when doing the clusters is at random allows the results to be replicated any other researcher.

The algorithm creates a list of length 30 representing each each test, and an integer value from 0 to 2 representing the three cluster. Since the goal is to find positive, negative, an neutral sentiment. It was viewed against both the reviewer and the turks and cluster map made. For these results it was deemed that 1 represented positive sentiment, 2 represented negative sentiment, and 0 representing neutral sentiment.

The mapping was merged with the samples and applied to Turks results and the like any model review a check of the model’s performance was completed.



Classifcation Report Results

The model shows that positive and negative sentiment are preforming somewhat well. As at random guess the model would preform at 33%. However that is however the level that neutral sentiment is classified at. This may be driven by the fact that the users in given a neutral sentiment of a product may use words that give both positive and negative sentiment which can give the model false information or lead the Turks in the rapid rush to get acquire their funds to flag the information and move on.

A kappa analysis was quickly run to see how the two preformed with each other and resulted in .18 kappa score meaning the the resulting value was low. The size of the data set is likely too small for the model and more examples would provide a more definitive example.

# Conclusion

Amazon Turks provides an interesting way to quickly source tags when information is missing. The sample data set for this exercise was small and might be pointing to a lower return of results than if there were more samples, but the Kappa models comparing the Turks the researcher shows moderate agreement and with more samples might slide even higher.

A machine learning model showed results that are better than a person randomly guessing at a tag with agreement meaning the Turks and model were able to clearly pickup on patterns in the text, but both would likely need more samples for the results to be used in a major model with level certainty.

That said the Amazon’s Turks does show promise for tasks in the future and with less specialized topics could prove quite helpful to advance research tasks.