HW5

Training/Evaluation Data Acquisition Through AMT

Bing-Je Wu

IST 736 – Text Mining

Table of Contents

[Introduction 2](#_Toc39246838)

[Analysis and Models 2](#_Toc39246839)

[About the Data 3](#_Toc39246840)

[Results 6](#_Toc39246841)

[Conclusion 7](#_Toc39246842)

[Reference 7](#_Toc39246843)

# Introduction

What is the review of the restaurant? How many stars does it have? How many reviews do they have? Is the rating on Yelp or Google? Those are the questions that people usually ask. The scenario happens quite often when people go out to eat and have no clue for which restaurant to go. In that case, people tend to reply on the recommendation from others. And, the easy way to get the recommendation from others is to look at the review system from different platform. Each platform has their credibility on different topic. For example, google are quite famous on its attraction review or sightseeing spot review. Many rely on the google map to find a place to go or to hang out. Yelp is specialized on restaurant review. Public love to use Yelp to leave the review for restaurant about their dining experience, food, services and decoration…etc.

Analysis and Models  
The Cohen’s Kappa statistic measure was used to find the degree of agreement between two annotators. The Kappa value give a quantitative measure of the agreement in any situation in which or more independent observers evaluating the same thing. The equation of the Kappa value is . is the probability of random agreement and is the probability of expected agreement.

**Table 1.** *Matrix of two annotators*

|  |  |  |  |
| --- | --- | --- | --- |
|  | | B | |
| Yes | No |
| A | Yes | a | b |
| No | c | d |

The equation of the observed proportionate agreement is . The equation for the expected probability is (.

**Table 2.** *Interpretation of Cohen’s Kappa*

|  |  |  |
| --- | --- | --- |
| **Kappa** | **Agreement** | **Percentage of data that are reliable** |
| **< 0** | **Less than chance agreement** | **0 – 4%** |
| **0.01 – 0.20** | **Slight agreement** | **4 – 15%** |
| **0.21 – 0.40** | **Fair agreement** | **15 – 35%** |
| **0.41 – 0.60** | **Moderate agreement** | **35 – 63%** |
| **0.61 – 0.80** | **Substantial agreement** | **64 – 81%** |
| **0.81 - 0.99** | **Almost perfect agreement** | **82 – 100%** |

## About the Data

Data were sampled from Yelp restaurant reviews. Six restaurant reviews were sampled. There is one one-star review, two two-stars reviews, one four-stars review and two five-stars reviews. The ‘label’ column was created with rules, number of stars below 3 labeled as ‘negative’ and number of stars above 3 labeled as ‘positive’. It is esteemed as the data ground truth. Three restaurant reviews were labeled as ‘n’ and the other three were labeled as ‘p’.

**Table 2.** *Dataset*

|  |  |  |
| --- | --- | --- |
| Attribute | Description | Data Type |
| stars | The Yelp rating system, from 1 star to 5 stars | Numeric |
| text | Yelp restaurant review | String |
| label | User defined sentiment based on stars as ground truth, ‘n’ or ‘p’ | String |

**Figure 1.** *Dataset*A screenshot of a social media post

Description automatically generated

The data were used for the Amazon Mechanical Turk (AMT) experiment. The experiment was designed to have five assignments for each of documents, restaurant reviews. In order to get the result in time, a short time period was set up as 2 days. There was no requirement for Turkers to complete the tasks. Five Turkers were expected to be hired for annotating the sentiment of the restaurant reviews. However, in lack of understanding how the AMT workflow works, there were 18 Turkers fulfilled the request. The project was done within four hours. With the price of $0.2 per Human Intelligence Task (HIT) and 30 assignments in total, the total cost is $7.2 for the entire project. Table 3 shows the information how the six documents and annotation were distributed into 18 Turkers. It is clear to see that one of the Turkers, annotator Q, has systematic bias, favoring the ‘positive’ label. Another Turker who could be the potential systematic bias is annotator J. That Turker has annotated three ‘positive’ labels when there is only one document actually is ‘positive’. However, the result of the experiment cannot hold the credibility because of lack of evidences. If there are more instances that all six documents were responded by the same Turker, the analysis could be stronger and persuasive.

**Table 3.** *Turkers participation and annotation with the ground truth, label*

A screenshot of a cell phone

Description automatically generated

**Figure 2.** *Average vote score for each of document (‘p’:1, ‘n’:0)*

A screenshot of a cell phone

Description automatically generated

Figure 2 shows the bias between the Annotators and the ground truth. The sentiment labeled from Turkers were calculated for voting. The vote score is the average of the labels score (‘n’:0, ‘p’:1) for each document. If the vote score equal to 1, it means the sentiment label is the same as the ground truth. As the vote score getting lower, it shows that the disagreement between Turkers and the ground truth. By taking closer look at Figure 2, it pointed out that two documents, document 4 and document 5, have the same sentiment. Both documents are positive reviews based on the ground truth. Its review contains positive words or phrase without too much ambiguous sentences or tone, such as ‘Famous’, ‘better’, ‘placed was packed’, ‘premium’, ‘illuminating review’, ‘hadn’t heard code violation’…etc.

**Table 4.** *Annotator P vs Annotator Q*

|  |  |
| --- | --- |
| A screenshot of a cell phone  Description automatically generated | ( |

**Table 5.** *Ground Truth vs Turkers*

|  |  |
| --- | --- |
| A screenshot of a cell phone  Description automatically generated | ( |

# Results

The Amazon Mechanical Turk provides a solution for people who have data but need to be labeled. Usually, when talking about labeling a data set, it requires a group of people who are experts in a specific area. It is costly and might be hard to find those talents. However, the AMT platform offers the solution for this problem. A requestor can create a project on AMT and set a bar for hiring workers who meet the requirements. However, how is the credibility turns out between labeled by Turkers and experts (here means ground truth) or between Turkers?

Since there are two annotators having overlaps on 4 documents, document 2, document 3, document 5 and document 6, the Cohen’s Kappa measure can be used to compare the degree of agreement on two Annotators. Table 4 shows the matrix between Annotator P and Annotator Q. The Kappa value was calculated as the table shown above. Both annotators, Annotator P and Annotator Q have poor agreement and the data are not reliable as the ground truth. However, this result cannot represent the conclusion because of the sample size. If there are more labeled data to compare, the Kappa value could have improvement in this case.

For the degree of agreement between the ground truth and Turkers, a Kappa value was calculated as Table 5 shown above. The Kappa measurement is 0.467. According to the Table 2, Interpretation of Cohen’s Kappa,this value is in ‘moderate agreement’ category. It seems like the annotations from Turkers are not too far from the ground truth but also not too close to the ground truth as well. It may have some reasons. One is that the annotators might be systematic bias, such as annotator Q from Table 3. The other thing needs to be mentioned is that the samples are not enough for advanced analysis between Turkers. A complete sample, having several Turkers finish all six documents annotation, should be obtained more for identifying the reliability. Based on the result of the experiment, the statement of ‘AMT is a viable approach for obtaining training labels’ cannot not be confirmed in this case. Although the result is not satisfying, it still has the possibility that AMT can be another approach to get the training labels, with restrict requirements on hired workers. By combining the factors, such as cost, time, and resources, into the consideration, using Amazon Mechanical Turk as an alternative to obtain the training data is still a viable approach.

Conclusion  
Everyone has their own judgement based on education, experiences, living background, personality…etc. When people read at a statement, it is nature that they have different interpretations in different way at different time. For example. The ambiguous statement , “their food is so greasy. I cannot wait to go there again”, can be labeled as ‘positive’ or ‘negative’ in this case. If a person who likes to use sarcasm may sense the negative attitude from the sentences. However, a person who does not use sarcasm may feel like the reviewer really want to go to that restaurant again and he or she may truly like greasy food.

Retaining a training label from others could be easier than obtaining labels from experts. However, each individual has their own thought and idea. To get the general training label, closing to the ground truth, people’s thought need to be combined and voted. In that case, using the statement mentioned above, some may think it is a positive review. But it will be a negative review when majority of people think it is a negative review because of the tone of sarcasm.

As that being said, if one has the limited budget and time pressure to finish a project, AMT platform could be a solution for obtaining training labels. And, in order to get the authenticity of the labels, the requirements of workers must be considered thoroughly.

# Reference

Cohen's kappa, from Wikipedia. [Web Link](https://en.wikipedia.org/wiki/Cohen%27s_kappa) ; 2 Kappa Online Tool [Web Link](http://vassarstats.net/kappa.html)  
3 Kappa Value Calculation, from YouTube. [Web Link](https://www.youtube.com/watch?v=DfNo32nL_fo&t=105s) ; 4 Cohen's Kappa in Excel, from YouTube. [Web Link](https://www.youtube.com/watch?v=AfgFyzGGlto)