**Artificial Intelligence**

**Training/Evaluation Data Acquisition Through AMT**

**Maya Mileva**

**HW4 IST 736**

A picture containing drawing

Description automatically generated

**Introduction**

Less than a decade after breaking the Nazi encryption machine Enigma and helping the Allied Forces win World War II, mathematician Alan Turing changed history a second time with a simple question: "Can machines think?" At its core, AI is the branch of computer science that aims to answer Turing's question in the affirmative. It is the endeavor to replicate or simulate human intelligence in machines. The expansive goal of artificial intelligence has given rise to many questions and debates. So much so, that no singular definition of the field is universally accepted.

It is useful to think of AI as being like a toddler. It can learn, develop and improve its capabilities over time, but it is not smart enough and does not have the emotional intelligence to know the context and impact of its decisions. If AI is utilized with care, these outcomes can be improved, including those related to diversity and inclusion. However, if AI is used carelessly, this can undermine its benefits and limit success for our people and the organization as a whole.

Though people love technology so deeply, they sure have an anxious relationship with it. They are worried about AI stealing their job, their house, or their significant other. People like to express their opinion, especially in Social Networks. . Social media is a part of people`s day-to-day life that cannot be ignored. It has both positive and negative effects. Most of the people are addicted to many social media platforms like Facebook, Twitter, Instagram, etc. and it is considered odd not to be connected. Social media has grown to become a large platform for entrepreneurs, businesses, organizations and various other professionals who seek identification and recognition at a moderate cost.

The best way to know what people think is analyzing the Social Media. There are so many tools that can be used for this purpose nowadays. But what about using people to do the computer’s job? What about a task, that human can do better than machine? For example, sentiment detection can be hard for a computer if there is sarcasm or hidden message. How do you teach machine to recognize it? Crowdsourcing platforms such as Amazon Mechanical Turk have become popular for a wide variety of human intelligence tasks. One of the tasks that the platform can be used for if sentiment detection. Working with unlabeled data has been always challenging, and now the there is one solution.

**Analysis and Models**

**About the data**

Five positive and five negative tweets about AI are used in this project. Labels were removed and csv file was uploaded in Amazon’s Mechanical Turk website: **(**[**https://requester.mturk.com/create/projects/new**](https://requester.mturk.com/create/projects/new)**)**

The goal of the project was to show how to deal with unlabeled data. The positive tweets contained words like “love”, “like” and the negative: “hate”, nasty”. Two of the reviews contain sarcasm, for the purpose of determining how good the Turks are going to perform.

**Positive Tweets Negative Tweets**

|  |  |
| --- | --- |
| A close up of text on a black background  Description automatically generated | A picture containing food  Description automatically generated |

Ten workers were hired initially, and the payment was $0.02. File with positive and file with negative tweets was uploaded. The only requirement for them was to be located in USA, English speaking people would get the sarcasm easier. The spam control measures included monitoring the average response time and the time for task completion. The average time per assignment was 2 minutes and 10 seconds.

For the second experiment only, master workers were required and the payment there was 0.3. This was another measure to avoid spams. Only one file without labels was uploaded. The point of the experiment was to determine if the master workers can recognize sentiment better. The average time per assignment was 7 minutes and 52 seconds.

In addition, third experiment was performed with more than 10 and not experienced workers, for the purpose of analyzing different results and calculating Kappa values among those workers.

**Data Exploring**

**Experiment 1**

File with five positive and file with five negative reviews was used. All reviews were labeled by the workers for 32 minutes.

|  |
| --- |
| A picture containing large  Description automatically generated |

8 Turkers worked on neg tweets and 12 worked on pos tweets. It is interesting to observe how many hits did each Turker do.

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

Positive and negative sentiment can be confusing. The time workers determine each was different.

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

None of the Turkers had less than 10 seconds response time, which was good. To detect potential Bots, consistent average response time was observed.

**Low**

A screenshot of a cell phone

Description automatically generated

**High**

**A screen shot of a social media post

Description automatically generated**

Workers can favor more positive or negative sentiment. In this experiment. The labels each Turker give were examined.

A screenshot of a cell phone

Description automatically generated

There are only 10 workers so it`s hard to determine. Probably the response time can help determine if some of them are bot.

A screenshot of a cell phone

Description automatically generated

It is really interesting how somebody can read a tweet for 0.1min… Most of the text was short, but it seems like half of the Turkers are bots.

**Experiment 2**

For the second experiment only one csv file without labels was uploaded. The demand there was all the workers to be masters, and they were paid more. Do they give the right labels? Let`s see.

A screenshot of a cell phone

Description automatically generated

They all labeled the sentiment correctly and they all agreed with each other.

A screenshot of a cell phone

Description automatically generated

Master workers seem to be reliable with small data sets.

A screenshot of a cell phone

Description automatically generated

There were 5 positive and five negative tweets in the original submission, and the results show that they were correctly classified.

**Experiment 3**

For this experiment only one csv file without labels was uploaded. The payment was again $0.2 and there were no other demands. Five workers were hired to work on each task. The average time per assignment was 6 min and 5 sec. The results from this experiment were more interesting.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generatedThe labels given this time, don`t much completely the initial ones.

More than half of the Turkers agreed on the sentiment.

A screenshot of a cell phone

Description automatically generated

The disagreement seems to be in the positive tweets. Worker`s seems to agree on the negative.

A screenshot of a cell phone

Description automatically generated

Cohen's kappa statistic, κ , is a measure of agreement between categorical variables X and Y. For example, kappa can be used to compare the ability of different raters to classify subjects into one of several groups. Kappa also can be used to assess the agreement between alternative methods of categorical assessment when new techniques are under study. According to Cohen's original article, values ≤ 0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41– 0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement.

**A screenshot of a cell phone

Description automatically generated**

Calculating kappa and inter-rater reliability when there are multiple reviewers is challenging . The reviewers have different number of reviews labeled, which doesn`t make the task easier. ReviewID was numbered from 1 to 10, matching the review for easier interpretation.

A screenshot of a cell phone

Description automatically generatedThey all have different number of turks. Top five were used for calculating Kappa. The table below illustrates the problem even better. Worker ‘A2RKUDGK5PQ44X’ labeled 8 tweets( 1,3,5,6,7,8,9,10 – missed 0 and 2). Worker ‘A2L746JBCNW066’ did only 6 and so on. To compare them using kappa, for review 0 and 2 the fist worker won`t have any values.

**A screenshot of a cell phone

Description automatically generated**

The maximum number of reviews is going to be 8. The comparison for the top five workers can be only on the five that they all have.

A screenshot of a cell phone

Description automatically generated

**TURKER 1** ['Positive', 'Positive', 'Positive', 'Negative', 'Negative']

Kappa=0.2258

0.2258

**TURKER 2** ['Neutral', 'Positive', 'Neutral', 'Positive', 'Negative']

Kappa=0.1836

0.2258

**TURKER 3**  ['Positive', 'Positive', 'Positive', 'Negative', 'Negative']

Kappa=0

0.2258

**TURKER 4** ['Negative', 'Negative', 'Negative', 'Positive', 'Negative']

Kappa=-0.3333

0.2258

**TURKER 1** ['Positive', 'Positive', 'Positive', 'Negative', 'Negative']

Kappa=0.4666

0.2258

**TURKER 3**  ['Positive', 'Positive', 'Positive', 'Negative', 'Negative']

**TURKER 2** ['Neutral', 'Positive', 'Neutral', 'Positive', 'Negative']

Kappa=0.2727

**TURKER 4** ['Negative', 'Negative', 'Negative', 'Positive', 'Negative']

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Var # | T1 | T2 | T3 | T4 | % agreement |
| 1 | **P** | **Neu** | **P** | **N** | **0.50** |
| 2 | **P** | **P** | **P** | **N** | **0.75** |
| 3 | **P** | **Neu** | **P** | **N** | **0.50** |
| 4 | **N** | **P** | **N** | **P** | **0.50** |
| 5 | **N** | **N** | **N** | **N** | **1.00** |
| Study Interrater Reliability |  |  |  |  | **0.81** |

From that percent of agreement, the kappa values probably should be higher with no negatives.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | T1 | T1 label | Tweet # | PoN |  |
| 0 | A2RKUDGK5PQ44X | Positive | 1 | P | 0 |
| 1 | A2RKUDGK5PQ44X | Positive | 3 | P | 0 |
| 2 | A2RKUDGK5PQ44X | Positive | 5 | P | 0 |
| 3 | A2RKUDGK5PQ44X | Negative | 6 | P | 1 |
| 4 | A2RKUDGK5PQ44X | Negative | 7 | P | 1 |
| 5 | A2RKUDGK5PQ44X | Negative | 8 | N | 0 |
| 6 | A2RKUDGK5PQ44X | Negative | 9 | N | 0 |
| 7 | A2RKUDGK5PQ44X | Negative | 10 | N | 0 |

The first Turker for example disagree with 2 of the labels (75% agreement).

Another approach for calculating Kappa was tried. T\_ID show all the Turks.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ReviewID** | **T\_ID** | **sentiment** |  | **True sent** |  |
| **0** | 1 | T\_0 | N | **P** | **1** |
| **1** | 1 | T\_1 | P | **P** | **0** |
| **2** | 1 | T\_2 | P | **P** | **0** |
| **3** | 1 | T\_3 | P | **P** | **0** |
| **4** | 1 | T\_4 | P | **P** | **0** |
| **5** | 2 | T\_5 | N | **P** | **1** |
| **6** | 2 | T\_3 | P | **P** | **0** |
| **7** | 2 | T\_6 | P | **P** | **0** |
| **8** | 2 | T\_7 | P | **P** | **0** |
| **9** | 2 | T\_8 | P | **P** | **0** |
| **10** | 3 | T\_3 | P | **P** | **0** |
| **11** | 3 | T\_9 | N | **P** | **1** |
| **12** | 3 | T\_2 | P | **P** | **0** |
| **13** | 3 | T\_5 | P | **P** | **0** |
| **14** | 3 | T\_6 | P | **P** | **0** |
| **15** | 4 | T\_10 | P | **P** | **0** |
| **16** | 4 | T\_9 | N | **P** | **1** |
| **17** | 4 | T\_11 | N | **P** | **1** |
| **18** | 4 | T\_6 | P | **P** | **0** |
| **19** | 4 | T\_5 | N | **P** | **1** |
| **20** | 5 | T\_3 | P | **P** | **0** |
| **21** | 5 | T\_2 | P | **P** | **0** |
| **22** | 5 | T\_12 | P | **P** | **0** |
| **23** | 5 | T\_13 | P | **P** | **0** |
| **24** | 5 | T\_5 | P | **P** | **0** |
| **25** | 6 | T\_2 | N | **N** | **0** |
| **26** | 6 | T\_3 | N | **N** | **0** |
| **27** | 6 | T\_14 | N | **N** | **0** |
| **28** | 6 | T\_1 | N | **N** | **0** |
| **29** | 6 | T\_9 | N | **N** | **0** |
| **30** | 7 | T\_15 | N | **N** | **0** |
| **31** | 7 | T\_9 | P | **N** | **1** |
| **32** | 7 | T\_2 | N | **N** | **0** |
| **33** | 7 | T\_3 | N | **N** | **0** |
| **34** | 7 | T\_5 | N | **N** | **0** |
| **35** | 8 | T\_6 | N | **N** | **0** |
| **36** | 8 | T\_2 | N | **N** | **0** |
| **37** | 8 | T\_7 | N | **N** | **0** |
| **38** | 8 | T\_16 | N | **N** | **0** |
| **39** | 8 | T\_9 | N | **N** | **0** |
| **40** | 9 | T\_5 | N | **N** | **0** |
| **41** | 9 | T\_17 | N | **N** | **0** |
| **42** | 9 | T\_9 | N | **N** | **0** |
| **43** | 9 | T\_6 | N | **N** | **0** |
| **44** | 9 | T\_2 | N | **N** | **0** |
| **45** | 10 | T\_5 | N | **N** | **0** |
| **46** | 10 | T\_9 | N | **N** | **0** |
| **47** | 10 | T\_11 | N | **N** | **0** |
| **48** | 10 | T\_2 | N | **N** | **0** |
| **49** | 10 | T\_16 | N | **N** | **0** |

It seems like the most disagreement was in Review 4***: “ Siri is my best friend, she never lies to me, buy groceries, pay my bills, tells jokes, sometimes even buy staff I don`t need. Artificial Intelligence make life better!’***”. Maybe some of the workers don`t know who is Siri, or they didn`t like the fact she buys staff without permission.

Sparse Matrix was created for easier calculation of Kappa.

A screenshot of a social media post

Description automatically generated

A close up of a logo

Description automatically generatedThe final Kappa values are:

|  |
| --- |
| KAPPA |
| 0.3264604810996563,  0.38493723849372385,  0.6989247311827957,  0.2751479289940828,  -0.011612903225806548,  0.4773333333333334,  0.5618479880774963,  0.6632302405498282,  -0.010309278350515427,  0.115979381443299,  0.3287671232876713,  -0.006849315068493178,  1.0,  -0.00512820512820511,  -0.010309278350515427,  -0.01379310344827589 |

**Results**

The goal of the project was to reveal a way to get label data when one is not provided. Amazon’s Mechanical Turk platform was used. Amazon Mechanical Turk is a crowdsourcing website for businesses to hire remotely located "crowdworkers" to perform discrete on-demand tasks that computers are currently unable to do. It is operated under Amazon Web Services and is owned by Amazon.

The workers were paid $0.02 or $0.03, depending on their qualification. The average time for the experiments varied from 20min to 45 min. The labeled data had to be confirmed before obtaining. The problem here was identifying spammers.

For Experiment 1 some of the negative tweets were labeled as positive. Labeling positive tweets seems to be hard for the workers, because they determine some of then as neutral or negative. The average time frame was from 11 sec to 4.18 min. Some of the workers were identified as bot.

For Experiment 2, all the workers were chosen to be Maters. They classified all the sentiment correctly. For the purpose of calculating Kappa values among those workers third experiment was performed.

For Experiment 3 pair-wise Kappa values among the workers are calculated. It was really hard comparing the labels due to the fact that same reviewed labeled different number of tweets. It is really easy to compare only two reviewers y1 and y2. In this case there are ten, all with different number of Turks. ReviewID was numbered from 1 to 10, matching the review for easier interpretation. The results showed slight for fair agreement. It was really interesting to observe the work of the Master Turks, where they labeled everything correctly, and then the work of people, without language of location preference. The AMT worker annotation reliability depends on their class, language and preferences, unless they are Masters ( they have algorithms that label really fast and true).

**Conclusion**

While technology continues to improve, there are still many things that human beings can do much more effectively than computers, such as moderating content, performing data deduplication, or research. Traditionally, tasks like this have been accomplished by hiring a large temporary workforce, which is time consuming, expensive and difficult to scale, or have gone undone. Crowdsourcing is a good way to break down a manual, time-consuming project into smaller, more manageable tasks to be completed by distributed workers over the Internet (also known as ‘microtasks’).

Amazon’s Mechanical Turk can be a great means of recruiting a diverse sample quickly and in a cost-efficient manner; however, the inherent differences observed between an Amazon’s Mechanical Turk  sample and a sample collected using traditional methods might present significant challenges in generalizing the results of the study. These differences include faith-based, technological, educational, age-related, socioeconomic, and employment-related differences. Additionally, the same ethical guidelines that you would uphold with participants collected from any other population must be maintained with Amazon’s Mechanical Turk  workers despite the limits this program places on personal interaction. Third party programs can still collect information about the projects you are running, and workers are under no obligation to keep their work secret.

Sentiment detection can be tricky and using Turks can be helpful, but a all the drawbacks have to be considered. Anyone who has tried to use Mechanical Turk for any large-scale task knows what the biggest problem is: Spammers! The job that they Turkers is important because they are getting computers one step closer to understanding people`s emotions.