IST-736 Text Mining Homework-5

Training/Evaluation Data Acquisition Through AMT

Sharat Sripada (vssripad@syr.edu)

Introduction

This week’s homework is around an interesting exploration about how a platform like Amazon Mechanical Turk (AMT) can be leveraged to label data. Subsequently, we will employ techniques like Cohen’s Kappa to determine the level of agreement or disagreement between the Turkers.

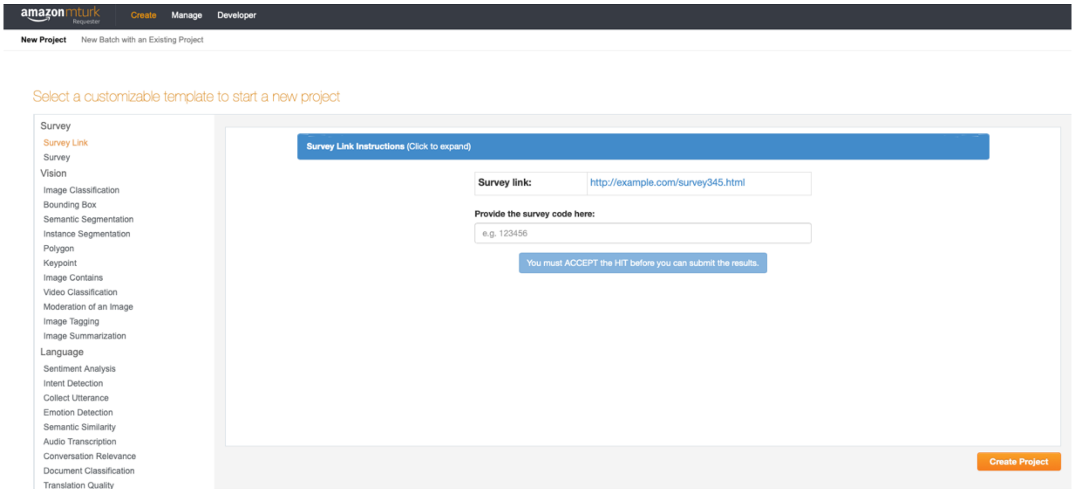
In the report we will briefly introduce the AMT platform, define important terminology, and see how it can be used to label the dataset introduced in Homework-1. That is, a corpus of tweets related to topic artificial intelligence.

About the AMT platform

A common requirement with datasets in machine-learning or artificial intelligence is to have the data labelled sufficiently so models can be trained and eventually make predictions. For the dataset specifically, we are looking to tag sentiment on a bunch of tweets.

Using the AWS credentials, login to URL www.mturk.com

The platform offers role as *Worker or Requester*. To stage data and getting the AMT workforce to tag the data, we would need to use the role of *Requester* hence. Further, on sign-on select *New Project* and choose *Sentiment Analysis under Language* and hit the *Create Project* button as shown below:



Select a customizable template to start a new project

Survey link: http://example.com/survey345.html

Provide the survey code here:

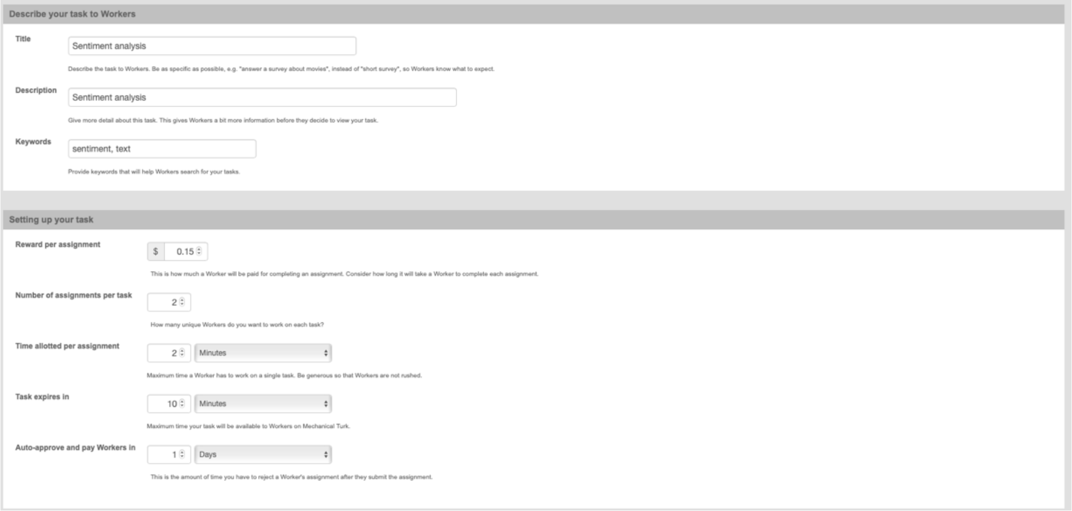
e.g. 123456

You must ACCEPT the HIT before you can submit the results.

Prior setting up the task it is important to understand the terminology. Following are a few:

- Worker: A worker refers to anyone with an MTurk worker account. Workers browse tasks posted on MTurk and choose to accept a task, work on it, and submit it when it is done.
- HIT: Expands as Human Intelligence Task. A HIT is a single unit of work that you want to complete. For example, we will use a collection of 30 tweets, each of those tweets would be a single HIT.
- Assignment: You can request one or more workers to complete each of your HITs. The work submitted by each worker for each HIT is called an Assignment. In our exercise, we will use 2 workers for 30 tweets amounting to a total of 60 assignments.

Knowing this helps us define the task clearly:



Describe your task to Workers

Title: Sentiment analysis

Describe the task to Workers. Be as specific as possible, e.g. "answer a survey about movies", instead of "short survey", so Workers know what to expect.

Description: Sentiment analysis

Give more detail about this task. This gives Workers a bit more information before they decide to view your task.

Keywords: sentiment, text

Provide keywords that will help Workers search for your tasks.

Setting up your task

Reward per assignment: $ 0.15

This is how much a Worker will be paid for completing an assignment. Consider how long it will take a Worker to complete each assignment.

Number of assignments per task: 2

How many unique Workers do you want to work on each task?

Time allotted per assignment: 2 Minutes

Maximum time a Worker has to work on a single task. Be generous so that Workers are not rushed.

Task expires in: 10 Minutes

Maximum time your task will be available to Workers on Mechanical Turk.

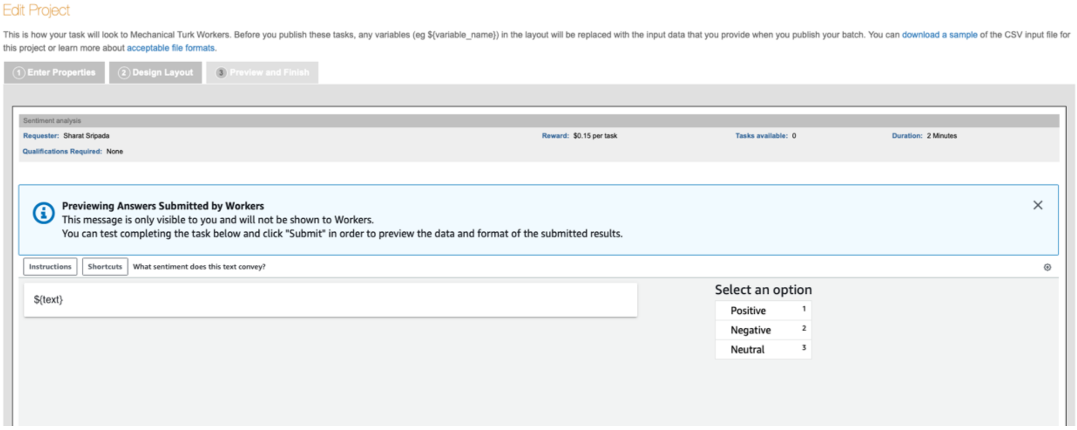
Auto-approve and pay Workers in: 1 Days

This is the amount of time you have to reject a Worker's assignment after they submit the assignment.

Next, we design the layout or dashboards the worker will see. Here, I customized the xml to comprise only Positive, Negative and Neutral sentiment (and removed the N/A). Also, we provide instructions to help the workers:

```
<full-instructions header="Sentiment Analysis Instructions">
    <p><strong>Positive</strong> sentiment include: joy, excitement, delight</p>
    <p><strong>Negative</strong> sentiment include: anger, sarcasm, anxiety</p>
    <p><strong>Neutral</strong>: neither positive or negative, such as stating a fact</p>
```

Finally, the preview and Finish screen would like this:



Edit Project

This is how your task will look to Mechanical Turk Workers. Before you publish these tasks, any variables (eg ${variable_name}) in the layout will be replaced with the input data that you provide when you publish your batch. You can download a sample of the CSV input file for this project or learn more about acceptable file formats.

Sentiment analysis

Requester: Sharat Sripada

Reward: $0.15 per task

Tasks available: 0

Duration: 2 Minutes

Qualifications Required: None

Previewing Answers Submitted by Workers
This message is only visible to you and will not be shown to Workers.
You can test completing the task below and click "Submit" in order to preview the data and format of the submitted results.

What sentiment does this text convey?

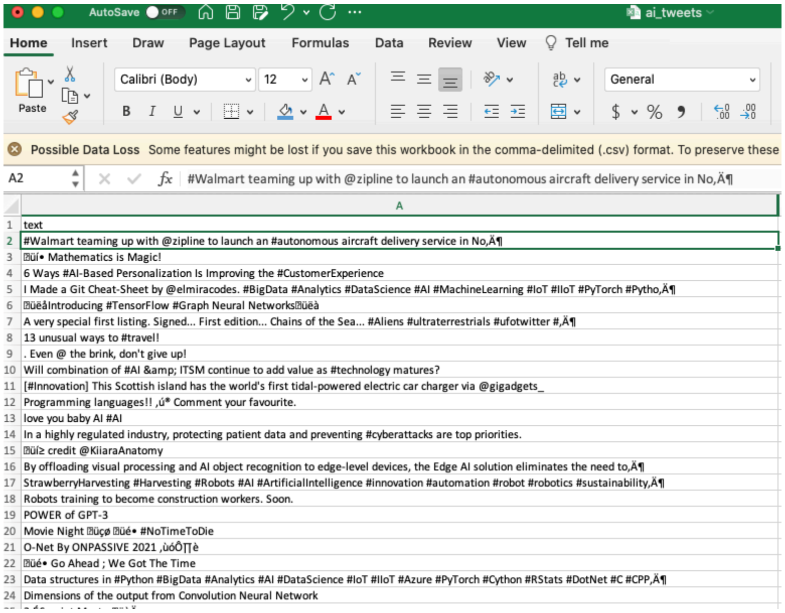
${text}

Select an option

Positive	1
Negative	2
Neutral	3

Data source, EDA, and labelled data on AMT

Once the dashboard and task has been setup on AMT, the final step is to Publish the data. For this purpose, we would create a csv file with the following content:



Possible Data Loss Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these

A2 #Walmart teaming up with @zipline to launch an #autonomous aircraft delivery service in No,Ä¶

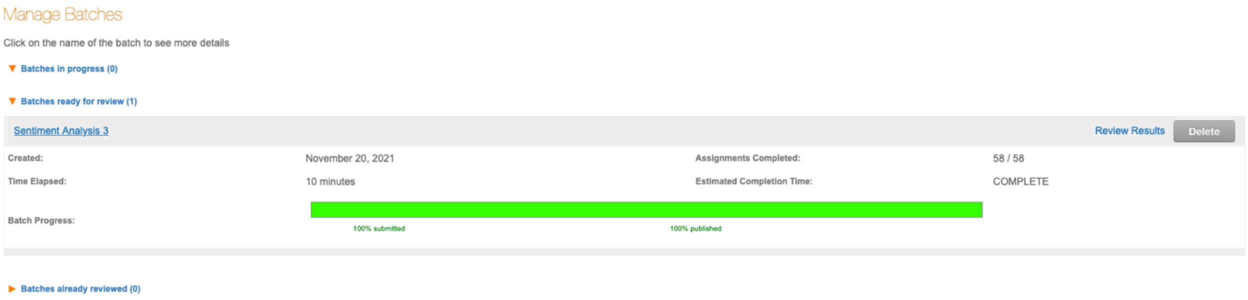
	A
1	text
2	#Walmart teaming up with @zipline to launch an #autonomous aircraft delivery service in No,Ä¶
3	üí• Mathematics is Magic!
4	6 Ways #AI-Based Personalization Is Improving the #CustomerExperience
5	I Made a Git Cheat-Sheet by @elmiracodes. #BigData #Analytics #DataScience #AI #MachineLearning #IoT #IIoT #PyTorch #Pytho,Ä¶
6	üëåIntroducing #TensorFlow #Graph Neural Networksüëà
7	A very special first listing. Signed... First edition... Chains of the Sea... #Aliens #ultraterrestrials #ufotwitter #,Ä¶
8	13 unusual ways to #travel!
9	. Even @ the brink, don't give up!
10	Will combination of #AI &amp; ITSM continue to add value as #technology matures?
11	[#Innovation] This Scottish island has the world's first tidal-powered electric car charger via @gigadgets_
12	Programming languages!! ,ú* Comment your favourite.
13	love you baby AI #AI
14	In a highly regulated industry, protecting patient data and preventing #cyberattacks are top priorities.
15	üí≥ credit @KiiaraAnatomy
16	By offloading visual processing and AI object recognition to edge-level devices, the Edge AI solution eliminates the need to,Ä¶
17	StrawberryHarvesting #Harvesting #Robots #AI #ArtificialIntelligence #innovation #automation #robot #robotics #sustainability,Ä¶
18	Robots training to become construction workers. Soon.
19	POWER of GPT-3
20	Movie Night üçø üé• #NoTimeToDie
21	O-Net By ONPASSIVE 2021 ,ùóÔ∏è
22	üé• Go Ahead ; We Got The Time
23	Data structures in #Python #BigData #Analytics #AI #DataScience #IoT #IIoT #Azure #PyTorch #Cython #RStats #DotNet #C #CPP,Ä¶
24	Dimensions of the output from Convolution Neural Network

Where the column *text,* comprises tweets obtained using the *tweepy* library. However, prior writing tweets into the data-frame and csv file we filter, and clean tweets based on the following:

- Remove RT tags from each tweet using re.sub()
- If a http(s) URL appears in a tweet, we will pass since we do not expect Turkers to spend time researching on the URLs

Obtaining results from AMT platform

Having uploaded the data to AMT, we can now publish it and allow workers to work on tagging the sentiment on tweets. The task can be monitored in the Manage section and completed in a span of a few minutes. Below is a snapshot of the completed assignment and task:



Manage Batches

Click on the name of the batch to see more details

▼ Batches in progress (0)

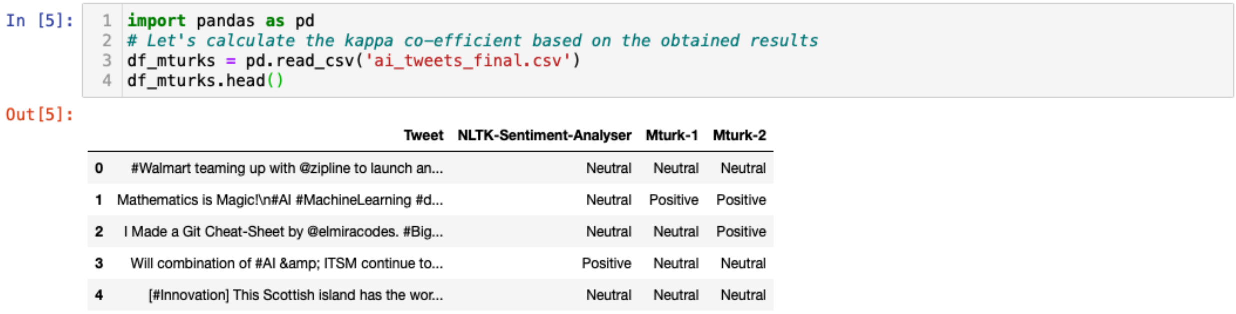
▼ Batches ready for review (1)

Sentiment Analysis 3

Created:	November 20, 2021	Assignments Completed:	58 / 58
Time Elapsed:	10 minutes	Estimated Completion Time:	COMPLETE
Batch Progress:	100% submitted 100% published		

▶ Batches already reviewed (0)

Once the results are available in AMT, the data can be downloaded for further analysis (as csv file). The csv is appended with an additional column, comprising sentiment tags obtained from the NLTK sentiment intensity analyzer. Below is a snapshot of the resulting data-frame:



```
In [5]:
import pandas as pd
# Let's calculate the kappa co-efficient based on the obtained results
df_mturks = pd.read_csv('ai_tweets_final.csv')
df_mturks.head()
```

Out[5]:

	Tweet	NLTK-Sentiment-Analyser	Mturk-1	Mturk-2
0	#Walmart teaming up with @zipline to launch an...	Neutral	Neutral	Neutral
1	Mathematics is Magic!\n#AI #MachineLearning #d...	Neutral	Positive	Positive
2	I Made a Git Cheat-Sheet by @elmiracodes. #Big...	Neutral	Neutral	Positive
3	Will combination of #AI &amp; ITSM continue to...	Positive	Neutral	Neutral
4	[#Innovation] This Scottish island has the wor...	Neutral	Neutral	Neutral

Results and accuracy

In this section, we will use the Cohen's Kappa score to correlate and compare the sentiment tagging between the Turkers and NLTK Sentiment Analyzer – in the following combinations:

- Between Mturk-1 and Mturk-2
- Between NLTK-Sentiment Analyzer and Mturk-1
- Between NLTK-Sentiment Analyzer and Mturk-2

But first, what is a Kappa score. Kappa score finds origins in the field of psychology, and it used for measuring the agreement between two human evaluators or raters (e.g., psychologists) when rating subjects (e.g., patients). It was later appropriated by machine-learning to measure classification performance. The metric is also commonly known as Cohen's Kappa coefficient after the American statistician and psychologist Cohen Kappa.

Kappa can be commonly defined as:

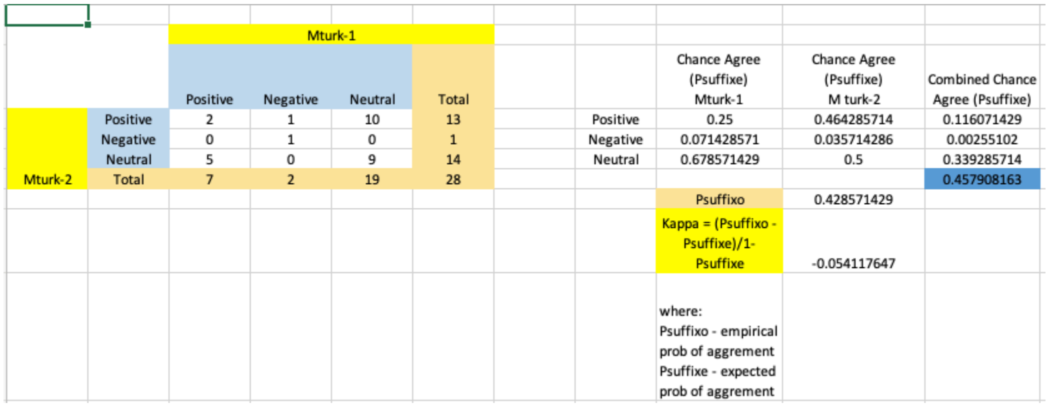
KappaScore = (Agree - ChanceAgree) / (1 - ChanceAgree)

where:

- Agree - proportion of agreement between two parties
- ChanceAgree - expected proportion of chance agreements
- KappaScore - (-1, 1) where 1 is perfect agreement

Calculation of Kappa score using manual method

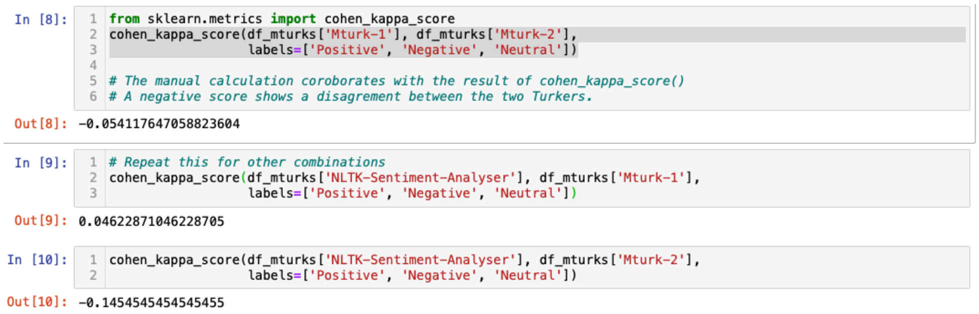
To understand the process thoroughly, see manual calculation of Kappa score between the two Turkers that is, Mturk-1 and Mturk-2.



		Mturk-1								
		Positive	Negative	Neutral	Total			Chance Agree (Psuffixe) Mturk-1	Chance Agree (Psuffixe) M turk-2	Combined Chance Agree (Psuffixe)
Mturk-2	Positive	2	1	10	13		Positive	0.25	0.464285714	0.116071429
	Negative	0	1	0	1		Negative	0.071428571	0.035714286	0.00255102
	Neutral	5	0	9	14		Neutral	0.678571429	0.5	0.339285714
	Total	7	2	19	28					0.457908163
								Psuffixo	0.428571429	
								Kappa = (Psuffixo - Psuffixe)/1- Psuffixe	-0.054117647	
								where: Psuffixo - empirical prob of aggrement Psuffixe - expected prob of aggrement		

Calculation of Kappa score using sklearn

Using sklearn the Kappa score can also we calculated using Python as seen below:



```
In [8]:  from sklearn.metrics import cohen_kappa_score
         cohen_kappa_score(df_mturks['Mturk-1'], df_mturks['Mturk-2'],
                           labels=['Positive', 'Negative', 'Neutral'])

         # The manual calculation coroborates with the result of cohen_kappa_score()
         # A negative score shows a disagreement between the two Turkers.
Out[8]: -0.054117647058823604

In [9]:  # Repeat this for other combinations
         cohen_kappa_score(df_mturks['NLTK-Sentiment-Analyser'], df_mturks['Mturk-1'],
                           labels=['Positive', 'Negative', 'Neutral'])
Out[9]: 0.04622871046228705

In [10]: cohen_kappa_score(df_mturks['NLTK-Sentiment-Analyser'], df_mturks['Mturk-2'],
                           labels=['Positive', 'Negative', 'Neutral'])
Out[10]: -0.1454545454545455
```

NOTE:
The Kappa score between the manual method and the sklearn match. It is -0.05

Summary of results:

	Between Turkers	Mturk-1 vs NLTK sentiment analysis	Mturk-2 vs NLTK sentiment analysis
Kappa score	-0.05	0.04	-0.14

Conclusion

First, what a great platform - AMT! For first time users, it is simple, intuitive and offers just enough customizations to be in control of data. As for the Kappa score, we see the following:

- Between Turkers - A negative score of 0.05 shows disagreement
- Between Mturk-1 and NLTK sentiment analysis - A score of 0.04 shows mild agreement. Note that, since NLTK sentiment analysis was used in homework-1 to label sentiment, it was used as one of the evaluators
- Between Mturk-2 and NLTK sentiment analysis - A negative score of 0.14 again shows disagreement.

Essentially showing poor agreement between any two columns in the data-frame.

Being curious if some of the Turkers were using a common NLTK sentiment tagger, comparisons between each Turker and the standard NLTK library were run. Again, very marginal agreement for one of the Turkers thus showing they possibly used discretion when sentiment tagging tweets.

Couple of possible improvements:

- In the mturk task there is possibility to request for more specific skills. I presume this would give us better or consistent results. This may be specific to the topic of artificial intelligence.
- Some of the tweets were truncated but I kept them assuming Turkers would tag it Neutral since I had a specific instruction, but they ended up tagging Positive/Negative for few instances. Perhaps, providing full tweets would aid Turkers better tag them.