**Data Analytics II:**

**Homework #4**

1. Identify a classification task that is relevant to your final project. Write a paragraph answering the following questions:
   1. What is the goal of this task?

The goal is to classify the text data that I crawled from twitter as pro-vaccine or anti-vaccine or neither

* 1. What dataset will you use to carry out this task?

I used the data that I got back from social feed manager with the query “anti-vaccine OR vaccine”

* 1. What annotations do you need to accomplish the goal identified in part a?

I need to write instructions about how to complete this task- like what the task is about and what the options are

I also need to clarify how to deal with the situations that are confusing

1. Write annotation instructions for Turkers given the task, goals, and dataset identified in #1. Include at least three examples of correctly annotated data for each annotation field identified in part 1c.

The description for it is:

|  |  |
| --- | --- |
|  | please indicate whether the displayed tweet is anti-vaccine or pro-vaccine or neither. |
|  |  |

And above each text data that requires to be classified, I stated that:

Please choose what sentiment you think this tweet conveys. Please only look at the tweet text body (not hashtags) to decide whether the attitude this tweet expressed is anti-vaccine or pro-vaccine or neither. If the tweet expresses a positive attitude towards vaccines, or people or incidents that are pro-vaccine, or it urges people to be vaccinated or it is relating the outbreaks to people who are not vaccinated, it should be counted as pro-vaccine, if the tweet expresses a negative attitude towards vaccines, or people or incidents that are pro-vaccine or related to vaccines, or it is discouraging people to get vaccinated or if it’s expressing that accidents or consequences happen because of vaccines, or if its expression is against mandatory vaccine order or law or vaccine manufacturers, it should be counted as anti-vaccine. Tweet that is anti anti-vaccine should be considered as pro-vaccine. The attitude a retweet expresses is the attitude that original tweet expresses. For example, if someone retweet a tweet that is pro-vaccine, then the attitude this retweet expresses should be pro-vaccine as well. If the tweet is just stating a fact, or not related to vaccine, or the main idea this tweet wants to convey has nothing to do with vaccine, then it should be counted as neither.

3 examples of Neither:

RT @GaviSeth: Today is the anniversary of the death of Maurice Hilleman, the American biologist who developed over 40 vaccines. We have him‚Ä¶

RT @ForeignPolicy: A recent study found that thousands of Russian accounts used to spread disinformation had seized on anti-vaccine messagi‚Ä¶

RT @ViraBurnayeva: "Water that contains more than 0.2 mg (200 mcg) #Aluminum per LITER should not be used for drinking or to prepare infant‚Ä¶

3 examples of Pro-vaccine:

RT @marcialangton: @Anthony\_Mundine You need to shut up. Babies are contracting measles because idiots like you believe these anti-vaccine‚Ä¶

3. Get Vaccinated : Vaccines are very good for our babies, ourselves, this generation and even the generations yet unborn to prevent some Cancers and reduce mortality rate from Vaccine preventable diseases in the world! #GetVaxxToday against HPV.! <https://t.co/qPevf6fEmC>

Measles: no vaccine, come another time. The interest in measles vaccination is increasing and the state is urging the Czechs to be vaccinated. https://t.co/o7xsM6FgJ0 via @iDNEScz @GIRPBrussels

3 examples of Anti-vaccine:

RT @Jimcorrsays: Did you know that since 1986 Vaccine Manufacturers cannot be sued because the amount of lawsuits against them at the time‚Ä¶

@AquaVelvaBoy @tammyhealy21 @kidoctr @ChrisJohnsonMD @HenningTveit @WendyOrent @doctorsensation Risks are low? Then why has there been $4+ billion paid in vaccine injury compensation with only about 1% of the cases being reported? So who's really being di

@NYCMayor Hopefully you will be run out of town on a rail by mothers of vaccine injured children https://t.co/Mf6eJekAXk

1. Create a CSV file containing the data to be annotated. Write a HIT that randomly displays samples and, using the instructions created in part 2, requests annotations for these samples.

Shuffled the samples in HW4.ipynb, the shuffled samples are in shuffled\_tweets.csv

1. Using the dataset identified in #1b, annotate 100 samples (these are your GOLD HITs).
   1. What is the distribution of annotations in these samples?

The distribution in my 100 samples is that:

Neither: 44/100

Pro-vaccine: 37/100

Anti-vaccine: 20/100

* 1. Consider the distribution identified in part a. Is sampling necessary? If so, annotate more samples as needed.

Yes. Sampling is necessary. Now my GOLD HITS contains balanced results of anti-vaccine, pro-vaccine and neither

1. Write a script to randomly insert the 100 samples identified in part 4 into the HIT that you created in part 3. You may create a new CSV to do this if possible. Given a set of responses from MTurk, demonstrate the ability to determine whether a given annotation refers to a GOLD HIT and, if so, whether it was annotated correctly.

Inserted the 100 GOLD HITS samples in the end of All\_HITS.csv and then shuffled all the samples.

The shuffled new HITS after inserting balanced GOLD HITS are in shuffled\_all\_HITS.csv file

The responses back is in Batch\_results.csv, and my annotation is in the Input.Annotation column(if there’s annotation in the Input.Annotation column, then it is a GOLD HIT). The response’s annotation is in the Answer.sentiment.label column, and I checked if the responses annotated correctly in the notation=gold\_HITS\_annotation column

1. Write a script to measure the accuracy of a given annotator on the GOLD HITs. Also measure the average time spent per HIT

Average time spent per HIT is 8.6842seconds

for annotation=gold\_HITS\_annotation column: =IF(ISBLANK(AE2)=FALSE,(AF2=AE2),"")

The accuracy is 0.95

The formula For accuracy column:

=COUNTIF(AI:AI,TRUE)/COUNTA(AE:AE)

1. Once you have convinced yourself that you can adequately identify cheaters, and that the instructions for your HIT are clear, copy your HIT to the MTurk worker website (not the Sandbox).

Done. Watching how my money is being snatched now

1. Given the timing data provided by your fellow students in part 7, decide on the payment rate per HIT (for a reasonable response rate, you should pay at least $6 per hour). Also decide on worker qualifications for your HIT (suggestion: >95% HIT approval rate).

I decided on 0.01 per task since per task is just one tweet and it should take less than 10 seconds on average. I used >95% HIT approval rate. Wanted to choose those for whom English is native language but it didn’t have the option

1. Launch your HIT on MTurk and collect the results. Each HIT should be completed at least three times, but do not spend more than $100 for this assignment (you may spend more for your final project if you wish). Pay good workers (those that have not been identified as cheaters) promptly.
   1. How many workers did your algorithm identify as cheaters? Do NOT automatically reject HITs from workers that your algorithm identified as cheaters. Instead, examine these cases individually and, if necessary, re-evaluate. NOTE: you may want to install the TurkOpticon plugin for Google Chrome in order to keep track of your reputation in the event that you decide not to pay workers.

I found 3 workers who were identified as cheaters. Their WorkerId are:

A3W0C04EXOY12C, A1GZZGA7ZFUPMB, ARIBVG7WV64QM.

(BTW, the results from MTurk requester website is in Real\_Batch.csv)

There were two or three others who also made a lot of mistakes on my GOLD HITS but they made far fewer mistakes than those three. Those three gave me a kind of feeling that they were just randomly choosing from the options and hoping that I would let it go and pay them

1. Generate descriptive statistics regarding your annotated data. What is the frequency of annotations for each of the classes that you generated? How does it compare to the distribution generated for the GOLD HITs in #4?

The total frequency of my GOLD HITS is 0.1680

The formula for GOLD HITS frequency is: =COUNTA(K:K)/(ROW(K1805)-ROW(K2))

The total frequency of anti-vaccine in the annotations is 0.3684

The formula for the total frequency of anti-vaccine annotation for the GOLD HITS is =COUNTIF(S2:S1805,TRUE)/COUNTA(K:K)

The total frequency of pro-vaccine in the annotation is 0.4868

The formula for the total frequency of pro-vaccine annotation for the GOLD HITS is

=COUNTIF(T2:T1805,TRUE)/COUNTA(K:K)

The total frequency of neither in the annotation is 0.1414

The formula for the total frequency of neither annotation for the GOLD HITS is

=COUNTIF(U2:U1805,TRUE)/COUNTA(K:K)

The distribution generated for the GOLD HITs in #4 (Before down-sampling) is

Neither: 44/100 or 0.44

Pro-vaccine: 37/100 or 0.37

Anti-vaccine: 20/100 or 0.2

After downsampling: Neither=Pro-vaccine=Anti-vaccine=0.33

We can see there’s a huge discrepancy between the two distributions

1. Calculate the inter-rater reliability of your annotations using Fleiss’ Kappa. How easy is your task? Based on these results, choose a method (e.g., simple majority; complete consensus, etc.) for deciding what annotations to retain.

The matrix is in Real\_Batch.csv sheet

(Just realized that when I saved my formula/results in the sheet it only saved the results, not the formula).

Basically I extracted and reordered the annotation based on the their original index before shuffling. Then I counted the number of “anti-vaccine”, “pro-vaccine” or “neither” annotation for every tweet (each tweet is annotated by 3 annotators)

Then I calculated every P\_i for each row using 1/(3\*(3-1))\*((number of anti-vaccine)^2+(number of pro-vaccine)^2+(number of neither)^2-3)

Formula: =(Z2^2+AA2^2+AB2^2-3)/(3\*2)

For each column I calculated the total (aka the total number of anti-vaccine, pro-vaccine and neither across all the annotations)

To calculate the total of each category: =SUM(Z2:Z602)

To calculate pj I used total of each category/the total number of tasks

Formula: =Z603/1803 or = SUM(Z2:Z602)/1803

To calculate P-bar we need to calculate the sum of P\_i, so I summed them all and got 252.66667 Formula =SUM(AC2:AC602)

P\_bar=1/601\*252.6667=0.4204 Formula:= AC603/601

P\_bar\_e=the square sum of all pjs= 0.33166944^2+0.49916805^2+0.16805323^2=0.3874

Formula: =Z604^2+AA604^2+AB604^2

K=(0.4204-0.3874)/(1-0.3874)=0.05386

Formula: =(AD605-AD606)/(1-AD606)

According to this table, there’s only slight agreement

I choose simple majority to decide the annotation. If none of the annotation agree, then I simply assign “neither” as annotation since it seems to cause confusion

1. Use the output from your MTurk HIT as training data for your classification task. Using 10-fold cross validation on a metric of your choice, pick a classification technique. Justify your choice of metric.

Code in Jupyter Notebook the last block. You have to read in the Real\_Batch.csv before running the last block though

I choose accuracy, because any kind of false labeling is bad.

The average accuracy given back from the 10-fold cross validation is 0.51

1. Write a short paragraph indicating whether your crowdsourced data was successful in accomplishing the goal defined in #1. Indicate how you might modify your approach to increase your success rate.

I don’t really think my crowdsourced data was successful in accomplishing the goal define in #1. First of all there are a lot of RTs in the tweets, which should be removed from the tweets to be annotated to increase success rate. Also, I notice that annotators can feel different attitude regarding the same tweet, maybe because it’s text data, and the sentiment it expresses can be different to different people? However, if fund allows, I think one way to improve the results is to have more annotators annotate the tweet and in which case simple majority can be more successful when there’s disagreement regarding what attitude a specific tweet conveys.