**Data Analytics II:**

**Homework #4**

**Madeline Warndorf**

1. The classification task that is relevant to my final project is to do a sentiment analysis of tweets that mention texting/operating a cell phone and driving (within the same tweet) to see if using a cell phone while driving is a bad behavior or a good behavior. The dataset that I will use to carry out this task is created from GW’s SFM Twitter crawler. The duplicates were removed and random sample of 1,000 tweets were selected out of the original dataset for monetary reasons (tw\_txt\_driving\_real.csv). The random selection is shown in Jupyter notebook. Fifteen attention checkers were also added to the 1,000 selected tweets. Therefore, the dataset that will be used in 1,015 tweets that reference texting/using a cell phone while driving. The annotations that I need to do to accomplish this goal is to define what it means for a tweet to have the opinion that using a cell phone while driving is a “bad” vs. “good” opinion. There will also be the annotation of “Neither/Unknown” which will state that the tweet isn’t referring to using a cell phone while driving, it is just referring to only using a cell phone (texting, tweeting, using Snapchat, etc.), only driving, or the tweet does not have enough information to determine the opinion.
2. Summary Instructions: Choose the primary opinion category that is expressed by the tweet.

Long Instructions: Determine if the tweet indicates that the person who posted it/retweeted it approves texting/using a cell phone while driving or disapproves texting/using a cell phone while driving.

* 1. Approves of using phone during driving:
     1. The tweet says “I am driving right now”
     2. The tweet mentions a negative view on a texting while driving law.
     3. The tweet supports others texting and driving.
  2. Disapproves of using phone during driving:
     1. The tweet says “Don’t text and drive”
     2. The tweet says “Say no to distracted driving”
     3. The tweet mentions seeing somebody else texting and driving and gives a negative view point about them doing it.
     4. The tweet states that texting and driving is dangerous.
  3. Neither/Unknown
     1. The tweet/retweet says “My girlfriend is driving me crazy”
     2. The tweet/retweet says “I love texting my mom”
     3. The tweet/retweet is cut off before a mention of using a cell phone while driving.
     4. The tweet/retweet mentions seeing somebody use their cell phone while driving but does not state their opinion on the matter (either through text or emojis).

1. The csv that will be annotated is txt\_driving\_withAC.csv. It includes the 1,000 randomly sampled tweets as well as the attention checkers. The write-up for the HIT is saved in HIT\_code.html.
2. The GOLD HIT samples were pulled from txt\_driving\_real.csv to prevent the attention checkers from being annotated.
   1. What is the distribution of annotations in these samples?

This distribution of the annotations is

Using GOLD\_Hits\_Annotated.csv:

=COUNTIF(C2:C101, “Disapproves”)

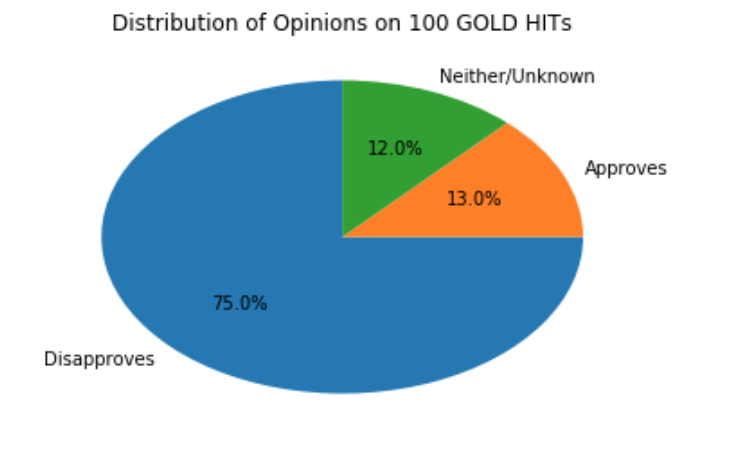
Disapproves: 75,

=COUNTIF(C2:C101, “Approves”)

Approves: 13

=COUNTIF(C2:C101, “Neither/Unknown”)

Neither/Unknown: 12.



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

I have a feeling that there will be more “Disapproves” because this data was scraped at the beginning of April which is Distracted Driving Awareness Month. I went ahead and added 50 other random tweets to my Gold Hits to see if there was any change in the distribution.

When I added the 50 new tweets the distribution didn’t change much:

Using Gold\_Hits150\_Annotated.csv:

=COUNTIF(C2:C151, “Disapproves”)

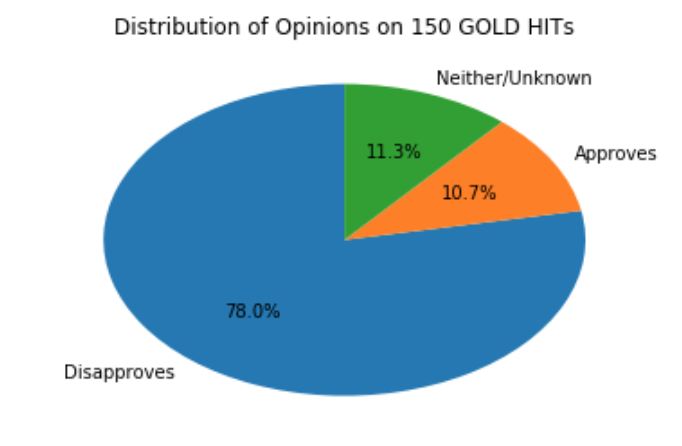
Disapproves: 117

=COUNTIF(C2:C151, “Approves”)

Approves: 16

=COUNTIF(C2:C151, “Neither/Unknown”)

Neither/Unknown: 17



1. The script is the same script used in question 2. The csv file used was Gold\_Hits150.csv. I did this in my Jupyter notebook source code.
2. I was the only one who annotated my GOLD HITs in MTurk sandbox. The script to find the accuracy of a given annotator (I took to mean a worker) is shown in the source code. Also finding the measure of the average time spent per HIT is shown in the source code.
3. I have included 15 attention checkers that are labeled and randomly distributed throughout the 1,000 tweets. I also believe that each HIT should take the worker at least 5 seconds minimum to read the tweet and decide what to classify it as. I also believe that it should take no more than 20 seconds to make the decision as well. That being said, if the worker shows that on average, they take longer to read and decide per HIT I will not hold it against them, but I will double-check their work if it is often over 20 seconds.
4. The timing that I took to go through 150 tweets was approximately 9 seconds per HIT. 9 seconds times 1,015 HITS is approximately 2.5 hours needed. Since MTurk can only do full hour allotments, that means 3 hours should be given to this assignment. If a reasonable hourly rate is $6 per hour, then I should pay $0.02 per HIT.

Additional qualification that I added for my HIT include workers that have an approval rate greater than 95%.

1. My Attention Checkers raised flags on several workers in the first Batch\_Request. Two workers failed 3 different questions in the checks and one worker failed 4. All of the workers that failed an attention check were double-checked and were not paid for the failed AC HITs.
   1. How many workers did your algorithm identify as cheaters?

I found a lot of cheaters from my cheater flag algorithm. The first Batch caused52 cheaters to be identified (only 22 were actually cheating) and 525HITs rejected (they are shown in First\_MTurk\_Batch\_Results.csv). I then checked the new responses and found no cheaters.

* 1. HITs were rejected if:
     1. It was a failed Attention Checker.
     2. If the tweet is clearly mislabeled.
        1. Example of Approves that were mislabeled:
           1. “I am texting and driving”
        2. Example of Disapproves that were mislabeled:
           1. “stop texting and driving”
           2. #justdrive, #UtextUpay, etc.
        3. If the tweet does not talk about texting and driving or using a cell phone while driving and it is not labeled as Neither/Unknown.
     3. Though it is clearly stated in the instructions, tweets that mention bills, laws, and texting while driving statistics were not rejected if they were not labeled “Disapproves”.

1. The other general descriptive statistics are in the source code.

The frequency of annotations for each of the classes that I generated are:

Using Cleaned\_Final\_MTurk\_Results.csv

=COUNTIF(E2:E3046, “Disapproves”)

Disapproves: 1,871

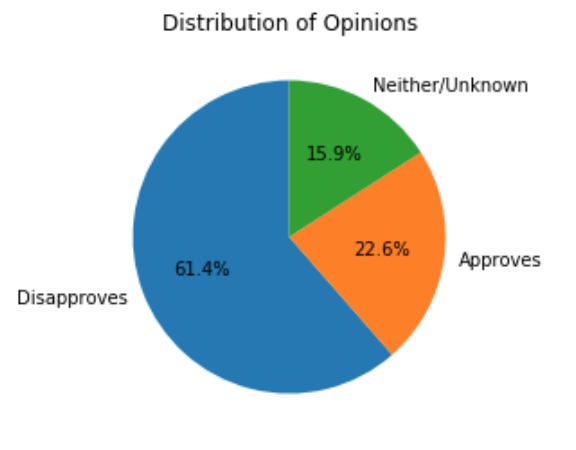
=COUNTIF(E2:E3046, “Approves”)

Approves: 689

=COUNTIF(E2:E3046, “Neither/Unknown”)

Neither/Unknown: 485

The distribution looks similar to the distribution of the GOLD HITs from Question 4. It just has more approves and neither/unknowns. This could be because each HIT was annotated 3 times which could result in different answers HIT.



1. The calculations are shown in my source code. The inter-rate reliability reflected in my Fleiss’ kappa score was 0.286. I didn’t think the task was too hard, however it did rely on how people interpreted what the tweet was saying (sarcasm vs. taking it at face value). I did receive one email saying that my instructions were confusing to them even though they classified all their HITs correctly. I believe that this task would have been easier if I broke down the instructions more. The method that I chose to select which annotations to keep was simple majority as well as executive decision since there were possibilities where the answers could be tied.
2. I did the 10 cross-fold validation in the source code. The metric that I decided to use is micro averaged (because my dataset is imbalanced and it is multilabel) precision because I am interested in knowing the measure of specificity that the model is able to produce. I really want to make sure that tweets that do have the opinion of supporting texting while driving are caught and classified correctly.
3. I believe that my crowdsourced data was semi-successful in accomplishing the goal that I defined in #1. I believe that a lot more tweets were probably labeled as “Neither/Unknown” when they shouldn’t have been. To help prevent this in the future, I would modify my approach by making sure my instructions were understood. I’m not sure how this task could have been broken down any smaller than it was. I also provided examples of each annotation that could have been seen in the tweets. I would also put a “location” qualifier on my HITs to set it to USA so that language is not a problem as well. I would also remove the Twitter links at the end of the tweets.