SI630 Homework 3 Report

Xinyi Ye, Yang Yu, Ruge Xu

March 2022

1 Part 1: Annotating Data

1.1 Measuring Inter-Annotator Agreement

Problem 6

We performed the following agreement analyses:

1. We computed r and the compute α using the ordinal and nominal level of measurements for the group member's annotations.

   The Pearson's correlation r is:

Annotator	user1	user2	user3
user1	1.000000	0.767748	0.797113
user2	0.767748	1.000000	0.780097
user3	0.797113	0.780097	1.000000

   Table 1: Pearson's correlation of our group

   The Krippendorff's alpha α with nominal metric is: 0.524.

   The Krippendorff's alpha α with ordinal metric is: 0.779.

2. We first compared the difference between r and the α scores. The correlation r's values are close to the α with ordinal metric, and higher than the α with nominal metric. This shows that the Pearson's correlation is based on ordinal values, which means the data has a predetermined or natural order.

3. Then we compared the difference between the ordinal and nominal α scores. The α with ordinal metric is higher than α with nominal metric. This means that the annotated ratings are more like ordinal values, and the α with ordinal metric is better for agreement measurement.

   We would use the α with ordinal metric in practice to measure agreement in this setting.

School of Information
Natural Language Processing: Algorithms and People (SI630)

Problem 7

This time, we computed the agreement of all the other annotators on our group's items.

First, we looked at the Pearson's correlation r for each group. The r of Group 24 is shown below as an example.

Annotator	user_29	user_30	user_28
user_29	1.000000	0.820823	0.598647
user_30	0.820823	1.000000	0.794964
user_28	0.598647	0.794964	1.000000

Table 2: Pearson's correlation of group 24

We found that our group had relatively high r values compared to other groups, except for Groups 9, 13, 15, 19, 21, and 25, which had higher r values than ours.

Then, we looked at the Krippendorff's alpha α value for all the other annotators.

The α with nominal metric is 0.426, the the α with ordinal metric is 0.689. Both are slightly lower than our group's.

The Krippendorff's alpha α values for each group are:

group	nominal	ordinal
group_01	0.254144	0.675410
group_02	0.274390	0.296482
group_03	0.298969	0.730159
group_04	0.210784	0.593197
group_05	0.133028	0.488636
group_07	0.311978	0.331481
group_08	-0.304833	-0.406607
group_09	0.826858	0.914960
group_10	0.064615	0.062382
group_11	0.129129	0.037555
group_13	0.424603	0.819315
group_14	0.441964	0.561404
group_15	0.862745	0.943700
group_16	0.193548	0.550265
group_17	0.073059	0.264493
group_18	0.006024	0.100000
group_19	0.893281	0.886555
group_20	0.244060	0.616681
group_21	0.646154	0.861446
group_22	0.038902	0.457645
group_23	1.000000	1.000000
group_24	0.437186	0.710306
group_25	0.784861	0.965473

Table 3: Krippendorff's alpha for each group

We see that Group 9, 13, 15, 19, 21, and 25 also have higher ordinal α than our group.

After comparing with the guidelines of other groups, we found that our guideline is somehow different from others' and there is room for improvement.

First, some other guidelines take into account aspects that we have not considered, such as the length of the answer,sarcastic or joking responses, external links, etc. Also, some other guidelines have not only positive examples, but also negative examples provided for references, and even provide a comparison between the different rating criteria, which our guideline does not cover. In addition, some other guidelines divide the criteria corresponding to each rating into two parts, satisfied and unsatisfied, while our guideline mixes all the criteria together.

1.2 Examining Disagreements

Problem 8

The table below shows the 10 instances that have the biggest absolute difference in mean rating between our group and the other group.

id	rating difference	group	rating
t3_nir04f	3.333333	group_10	4.666667
		group_12	1.333333
t3_nlsfqo	3.333333	group_01	5.000000
		group_12	1.666667
t3_nl6dx3	2.833333	group_08	1.500000
		group_12	4.333333
t3_nl4icj	2.666667	group_07	4.666667
		group_12	2.000000
t3_ng8b9p	2.666667	group_12	2.333333
		group_21	5.000000
t3_njrkw6	2.500000	group_08	2.500000
		group_12	5.000000
t3_nai64u	2.500000	group_08	2.500000
		group_12	5.000000
t3_n5lm2s	2.333333	group_09	1.000000
		group_12	3.333333
t3_nd4zey	2.166667	group_08	2.500000
		group_12	4.666667
t3_nmnfge	2.000000	group_12	3.000000
		group_18	5.000000

Table 4: 10 instances that have the biggest absolute difference in mean rating

The texts in these 10 responses are:

1. id: t3_nir04f

   question: What is the lore of the videogame / show you are currently playing / watching?

   reply: Some anti social guy joins the literature club of his high school and tries to get closer to its members who each have their own issues

2. id: t3_nlsfqo

   question: What methods do you use to prevent or stop addictive smartphone behavior?

   reply: I recognize that smartphones/tablets and other small computers or bots are part of the human development to attach to artificial intelligence appendages like external body attachments on robots that will eventually become implants. There was a time for example you needed to physically read encyclopedia for world information and maps for directions, now you phone can read the encyclopedia and map data for, too you. In the future I would not be surprised if the external attachments (phones etc.) get smaller but more advanced so eventually what were external implants become internal implants... Now you have all information in your mind. You can pull a map that sends info to the optic nerve or answer advanced physics questions in seconds with brain implants that help process information faster. This helps keep humans advancing with robots and or A.I. Seems scary but look how much influence what was once just a phone to make 2 way calls has become and you can sorta see humans being groomed for a technological Singularity. Knowing or sensing these things helps me to control and manage my attachment level to technologies. In deeper thought I feel some Extraterrestrials have met similar fates with technological Singularity and by doing so they become more dominate in the universe. Somethings to consider.

3. id: t3_nl6dx3

   question: What's the stupidest, most impulsive thing you've ever done that ended up working out perfectly?

   reply: Went on a blind date with some random woman I'd never even spoke on the phone with. I married her.

4. id: t3_nl4icj

   question: What is a cringy scene from a show/movie that no one talks about?

   reply: Stranger things at the end of the last season when that one girl is singing the song during the intense scene.

5. id: t3_ng8b9p

   question: You have 10 seconds to talk to aliens. What u'll say?

   reply: Me: Hey! Wanna help me find out if our species are sexually compatible ? Alien: Well, I'm all in for that, but normally, our species can only have sex once per year. Me: Why ? Alien: We reproduce by the female litterally **eating** the male partner's gonads during sex. They take about a year to grow back. So, you naughty boy, ready to go to bed and try out? Me: Erm... You know what, we just met and I wouldn't want to cause a diplomatic incident.

6. id: t3_njrkw6

   question: Easy middle grade books in German?

   reply: I'd say to absolutely go with something with a theme that at least sparks some interest on your part. If that means reading translations - go for it! Here are some children's/teen geared books that I really liked growing up. Michael Ende - great writer, loved the Wunschpunsch, no idea how well it holds up for an older audience - he wrote some books that maybe are more suited for older teens than that and the Jim Knopf ones like Momo and the Neverending Story. Otfried PreuBler's books - I loved Krabat I read Scandinavian children's lit authors like Lindgren, Lagerloef, and Jannsen in German - the translations are top-notch. Some of Janosch's books are pure genius. Der Froschkoenig is the bomb. (they are illustrated children's books, mind, but have a particular folly to them) I never got around to reading moers' 13 1/2 leben des Kaept'n Blaubaer - but I see it mentioned often on the forum as a hidden gem.

7. id: t3_nai64u

   question: What was your "terrible roommate" horror story?

   reply: My roommate a new college freshman had a fetish for older men. She had a boyfriend named Doug and he was like 56 he used to stare at me and our other young female roommate. He told us to call him daddy and left anal beads out on the coffee table. So gross. When we told our roommate that her bf was creepy and we didn't want him coming over anymore she picked up a dinner plate and threw it at the wall behind me it didn't hit me but the glass pieces did scratch my arm. She ended up screwing us over by leaving no notice but believe me it was worth it.

8. id: t3_n5lm2s

   question: What is one secret you hid from your friend that if revealed, it will activate the "end" in friend?

   reply: For 20 years his mother and I were having incredible sex. We decided to end it when I married my babysitter.. we realized we were going to get caught. Was not exactly 20. A little less.

9. id: t3_nd4zey

   question: People who work in McDonald's, after working there , would you still eat McDonald's?

   reply: Definitely. I worked there for a year and the hygiene standards at my location were really high. Every food item was on a timer, everything was meticulously clean.

10. id: t3_nmnfge

    question: What piece of technology deserved a second chance?

    reply: The PSP. That thing brings me back so many memories and was definitely ahead of its time. Too bad the PS Vita didn't catch on.

For the replies with id t3_nir04f and t3_ng8b9p, our group thought that the respondents did not give a direct answer, but only a description of the answer, and therefore gave a low rating. The other group thought that they had given enough information and therefore gave a high rating.

For the reply with id t3_nlsfqo, our group thought that the respondent seemed to have given an answer with a long explanation, but the central meaning was almost irrelevant to what the questioner wanted to know. The other group thought that the respondent had replied in the correct form.

The replies with id t3_nl6dx3, t3_njrkw6, t3_nai64u and t3_nd4zey show the great divergence between our group and the group 8. Our group thought that the respondents correctly replied to the questioners' questions, while the other group thought that these were totally dull responses or slightly off the topic, and thus gave a low rating.

For the replies with id t3_nl4icj and t3_nmnfge, our group though that the respondents replied to the questions correctly, but lacked more in-depth explanations, so they were given a moderate rating. The other group thought that the replies were immediately relevant to the subject of the question, thus giving a high rating.

For the reply with id t3_n5lm2s, our group considered the reply to be consistent with the overall content of the questions, despite the reference to sex-related jokes, and therefore gave them a moderate rating. Other groups felt that the reply should be given a low rating for including sexual jokes.

After discussion, we concluded that the differences in ratings existed mainly between the groups; therefore, we used the average ratings of the two groups as the final true ratings.

Problem 9

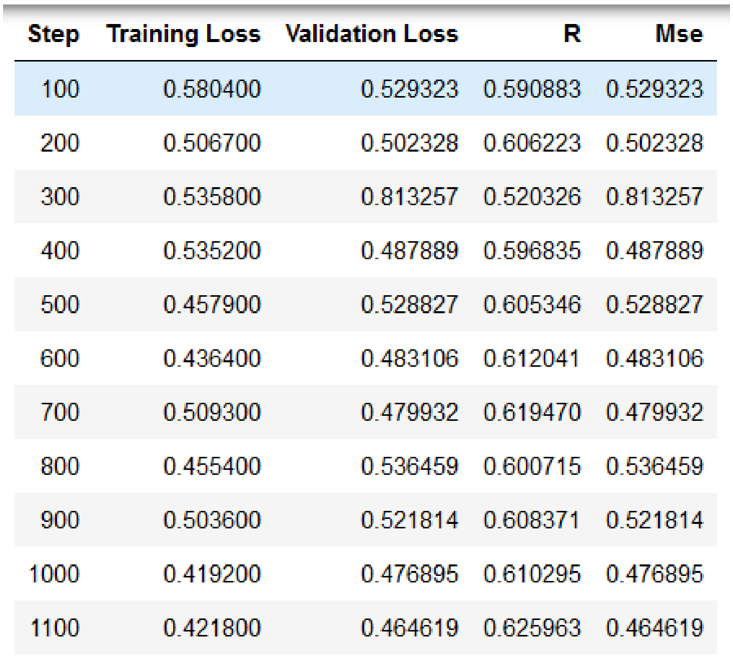
We believe the following improvements could be made to our guidelines to improve agreement among annotators:

- Consider more aspects that were not though of before, such as sarcastic or punny responses, references to external links, etc. Clarify the rating of such kind of replies.
- Specify whether the form of the answer needs to be strict. For example, for wh-questions, does the respondent have to give a clear answer rather than a general description.
- Clarify what rating should be given for responses that involve sensitive topics, such as sex and drugs.
- Clarify the definition of whether or not the reply is relevant. Because everyone is likely to have a different opinion about whether or not the topic is relevant to the question.
- Divide the criteria for each rating into two parts, i.e., positive criteria that once a reply satisfies, then it meets the specific ratings, and negative criteria that once a reply unsatisfies, then it does not meet the specific ratings.
- List not only positive examples, but also negative ones for reference. Add comparisons between replies of different ratings.

2 Part 2: Recognizing Helpful Answers

Problem 11

For this part, We trained the model on the whole train dataset.



Step	Training Loss	Validation Loss	R	Mse
100	0.580400	0.529323	0.590883	0.529323
200	0.506700	0.502328	0.606223	0.502328
300	0.535800	0.813257	0.520326	0.813257
400	0.535200	0.487889	0.596835	0.487889
500	0.457900	0.528827	0.605346	0.528827
600	0.436400	0.483106	0.612041	0.483106
700	0.509300	0.479932	0.619470	0.479932
800	0.455400	0.536459	0.600715	0.536459
900	0.503600	0.521814	0.608371	0.521814
1000	0.419200	0.476895	0.610295	0.476895
1100	0.421800	0.464619	0.625963	0.464619

Figure 1: An Overview of Training

During this part of training, we achieved a performance of 0.63 for correlation score and 0.46 for mean squared error score on development set. Afterwards, we submit our prediction to the Kaggle competition.

2.1 Annotation Evaluation via Classification

Problem 13

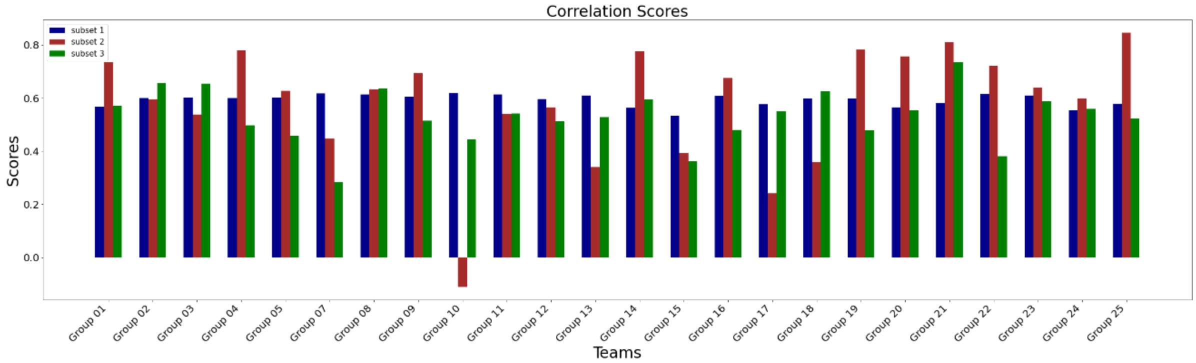


Figure 2: Annotation Evaluation

For each group, we trained a model using data without the group's annotation. Afterwards, we test the performance on 3 subsets of development sets. The resulted barplot is shown in Figure 2.

In the group 10, we can see that group 10's annotations seem to cause the model to perform worse. The correlation between predictions and their annotations is even negative. And the score on the first subset, which shows the general performance is the highest, showing that without the group's annotations, the model can be better. Reading the group's guideline, we can see that the examples within it is weird, where sarcasm or short answers can be scored 5 points. In other groups' annotations, detailed answers without sarcasm are more likely to be highly scored.

We can see that for groups 23 and 24, the scores on three subsets are pretty close, which may imply that their annotations are like the majority and will not effect the average model much. Reading their guidelines, we can see that they are written in good manners with clear tables. These guidelines seem to be pretty standard.

For group 25, we see an interesting thing. The score on the second subset is much higher than the other two. This shows that training without this group's annotations, the resulted model can still predict their annotations pretty well and the score is even better. That is somehow hard to understand. Reading their guideline, we found it is relatively simple. We guess that their annotation style may be simple too, which can be easily predicted. For instance, they may just give those longer answers better points.

Finally, thanks to all the instructors for the guidance!