SI 630 Part 2 Report

Homework 3

Chongdan Pan, Limeng Liu, Shirley Wang

Kaggle Leaderboard Team Name: Group2

Problem 6. (5 points) Perform the agreement analyses

1. For the Krippendorff ***α***, the ordinal value is about ***α***, while the nominal value is near 0.33. Since we have three members for annotation, the result is

	pandapcd	xuelw	liulim
pandapcd	1	0.658533	0.696062
xuelw	0.658533	1	0.648942
liulim	0.696062	0.648942	1

2. Our correlation is higher than the ordinal Krippendorff ***α***, that's probably due to the fact that Krippendorff ***α*** is calculated by the result of three annotators, which may have more diversity. However, the difference between pearson correlation and ***α*** is not very large, we still have a relatively consistent correlation between any two annotators, implying that our group guideline works well.

3. For Krippendorff ***α***, the nominal one is very low because it considers the distance between 4, 5 and 1, 5 as equal, while it's not the case in our annotation process. The distance between 4, 5 is definitely smaller than 1, 5, hence we should put the ordinal Krippendorff ***α*** into practice.

Problem 7. (5 points) Comment on the difference (if any) between your group's and other students' agreement and what you think is the causeOnce all the annotations are released, compute the agreement of all.

Other annotators on our group's item, have the Krippendorff ***α*** as 0.73, and ours is 0.62. Their correlation is around 0.7, however, since some other groups have only 2 members, we simulate the third member's score by calculating the group rating mean, which may increase the correlation, and hence the comparison between correlation is not reasonable.

The difference between the Krippendorff ***α*** is probably because in our guideline of subjective questions, we focused more on whether the answer have a strong reason or not (for example, our guideline need to first identify the type of the question and for each type the answer should have relatively complete or slightly strong evidences), however, in other groups, they have other criterias like grammar, styles, writing quality, answer length, etc. which we think may not be relative to the helpfulness of the answer. That will cause them to give higher ratings on these questions.

Problem 8. (5 points) 10 replies and what the ratings were and describe why you think the two were different, adjudicate the rating and decide what is the final "true" rating, examining all the evidence.

1. What is your all time favorite Disney animated file?
ITS THE CIRCLE OF LIFEEEE
Us: 4, 1, 4, other: 5 (group 1)
Internal: List item without reason, however the answer is a song not a film, the controversy is that whether we can consider the film song as the representation of a film
Group 1's rationale: fully and directly addressed the question, reply to the question of which and what with detailed information
Differences: they did not follow their guideline
True Rating: 2 (answer need relevant knowledge to understand, but did not provide enough information, we will need to add this to the guideline)

2. What is the best comeback to an insult you have ever heard or said?
Your mom goes to college.
Us: 4, 2, 4, other: 1 (group 16)
Internal: whether respectful, question itself may not be "good"
Group 16's rationale: Insults the original poster for asking the question
Differences: difference in identifying whether the answer is respectful, for example, cannot have dirty words, in our guideline, we did not directly point out how to treat the dirty phrases
True Rating: **2**

3. Why would anyone want to have kids?
Many people want a traditional sort of family. They want to raise children with a partner. Even if you don't want kids, I think the motivation is easy to understand.
Us: 4, 5, 4, other: 2 (group 8)
Internal: subjective to identify the solid of the reasoning
Group 8's rationale: Dull response / Slightly off the topic
Differences: we think the responder answer the question, but the reason is not strong, but they are judging whether the answering is interesting
True Rating: 4

4. Young cat got very sick quickly and died earlier this week—found his tiny nose prints on bedroom window and got very upset. What unexpected things should you prepare yourself for when a pet passes suddenly?
Deciding what to do with all her bedding, cat cave and other bits and pieces. I had to put my elderly cat to sleep a few weeks back-she had been reasonably OK but then had a rapid decline so it was quite a sudden end. She had 3 beds (because she had different favourite sleeping spots depending on the time of day and where the sun was), blankets, litter trays, liners, enormous bags of litter, a water fountain etc. It took me a while to feel up to laundering all her beds, blankets and cushions. I've donated the items in good condition to the local shelter, together with all the unused cat food I had stocked. I had her litter trays set up in the downstairs bathroom, and going into the room now and seeing a clean and empty floor where her litter trays and continence pads used to be still throws me.
Us: 5, 5, 4, other 2.6 (group 11)
Group 11's rationale: May be difficult to identify the direct response to the question
Differences: they did not understand the question since it is a long answer
True Rating: 5

5. Trying to make a point to someone. Do you wear the same clothes at school and when you are out with your friends or do you change them?
For a long time i didn't changing my clothes all day long, from morning toilet to evening shower. I changed this at high school, when i started to sweat way more.
Us: 3, 2, 3, other: 5 (group 7)
Group 7's rationale: subject agreement: immediately relevant to the subject of the question, clear organization, good writing quality
Differences: their guideline did not categorize different scenario and have ambiguous settings - need the annotator to identify 4 and 5 by reading the answer - did not provide clear rating criterion
True Rating: 3

6. What are some of your pet peeves?
I can't stand people who eat and chew loudly
Us: 3, 1, 3, other: 5 (group 23)
Group 23's rationale: relevancy (15), context (10), supporting details (0), understood questions (6), writing quality (8), appropriate length (5) = total (44) – absolutely helpful
Differences: the guideline is confusing and the helpfulness should not have much wrights in writing quality and length
True Rating: 3

7. When was the last time you tried to activate any dormant superpowers you may have?
A spider bit me. I tried to activate the superpowers as soon as I woke up the next morning upside down.
Us: 4, 4, 5, other: 2 (group 24)
Group 24's rationale: the answer is sarcastic or a joke

Differences: the question is a unrealistic question, but probably the other group think it is a joke related to the Spiderman
True Rating: 4

8. Which late night Tv show host do you love the most and which one do you hate the most?
Love Graham Norton. Hate Ellen DeGeneres.
Us: 3, 3, 3, other: 5 (group 18)
Group 18's rationale: The answer can be either very concise and short, that directly answers the questioner what to obtain with no ambiguity
Differences: we care more about the reason to the answer, but the other group think the answer is enough without a reason
True Rating: 3

9. When you imagine your Dream House, what is it like? Modern? Rustic? In the countryside, or in a City?
A large castle just east of town, on top of a man-made hill that I will have to build before building my castle. Ideally, large enough to block out the morning sun from the city for an hour or two. There would even be a moat around the castle. Not a lava moat or a water moat, but a traditional ditch moat filled with poison ivy and rattlesnakes. Yeah, that's right Chinese Restaurant, try to leave a fucking flyer on my door now. You'll have to get past my moat, the drawbridge is up.
Us: 2, 2, 2, other: 4.67 (group 24)
Group 24's rationale:: include examples and information that is useful or advantages knowledge eo the subject, easy to understand, clear answer the question, interesting and humorous
Differences: we focus on the bias and discrimination of the answer and have identified as 2
True Rating: 2

10. What advice can you give to young adults that are wasting their 20s?
Make any kind of investment, preferably one you can't touch if control will be an issue. Just leave yourself anything to work with when you either get your shit together or have an emergency.
Us: 3, 3, 2, other: 5 (group 7)
Group 7's rationale:: subject agreement - Immediately relevant to the subject of the question, organization - Clearly articulates a claim and demonstrates solid use of supporting evidence, writing quality - Writing does not affect reading comprehension and does not contain parts that needs further research
Differences: we think suggestions need more explanation and reasons, the answer does not have "supporting evidence", they violate their guidelines
True Rating: 3

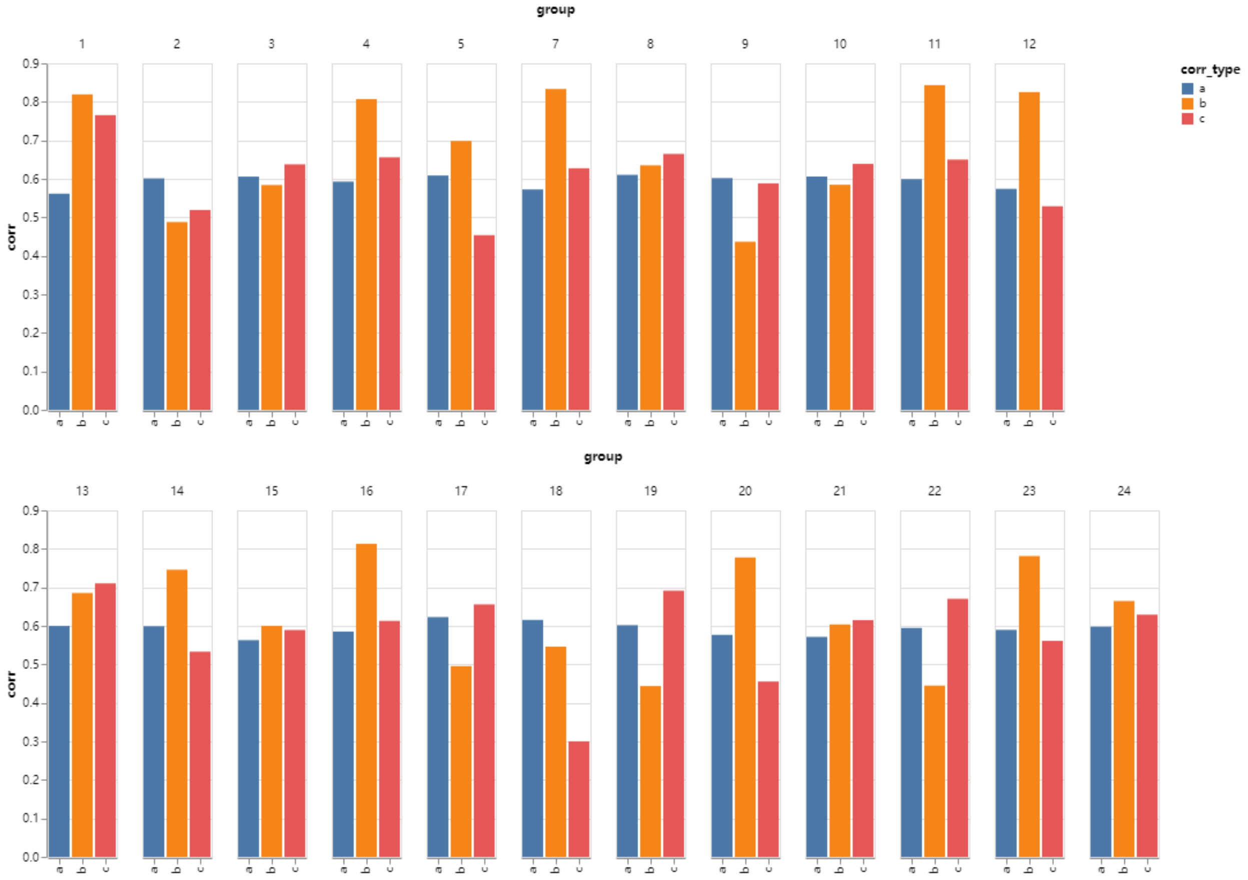
Problem 9. (5 points) At least 5 improvement

1. “Being able to guess the question by looking at the answer is a way to identify a response of a 5” from group 24 can be added to our guideline
2. In List item section - subjective question - add new rule to [2] answer relative topic, but need relevant knowledge to understand the answer
3. Add a new rule to [2], the answer should not have dirty phrases
    - For example:
        - What is the best comeback to an insult you have ever heard or said?
            - [1] not answer
            - [2] (Verbally) dirty phrases
            - [3] (Verbally) Stop it
            - [4] response with humorous
            - [5] response with context and reasons
4. Add guidelines for follow-up questions, if the answer is to a follow-up question, it will be marked as 3
5. Writing Style and Quality: if the answer in a poor writing quality (for example: “we they are”, spelling error which may cause other not able to understand the reply, etc.), the rating should deduct 1 point

Problem 11. (15 points) Report your score on the development dataset

The MSE for our group in the development dataset is 0.409.

Problem 13. (10 points) Make a bar plot of the correlation scores of each group for the three evaluation subsets



In a paragraph, describe what you see and discuss the relative impact of each team's annotations:

- From subset (a) we can find that how the model perform in general without this group's annotation, therefore, a higher correlation in the subset (a) indicates that the group's annotation causes the model to perform worse, such as group 3, group 17, group 18 (correlation higher than 0.6)
- For comparing the subset (b) and subset (c), half of the groups have higher correlation in subset b (model is more able to predict this group's annotation) and half of the groups have higher correlation in subset c (model's prediction is closer to everyone else's prediction).
- For subset (c), if the correlation score is low, which indicates that the groups who assigned the same questions have significantly different guidelines, for example, group 18 has a low subset (c) correlation and it turns out that group 18's guideline is very general and abstractive without detailed information in the differences between the close two labels, for example, it may cause confusion between 4 and 5 and 3 and 4.

- An interesting finding is that for group 3, the correlation for subset (b) is negative (-0.02) when we first trained the model, and when we restart and trained again, the subset (b) correlation is close to other groups. This finding means that for every time the training may result in different correlations and may need further analysis on the reason.