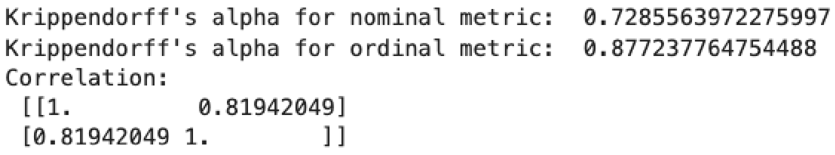
Homework 3 Report -Group25

Problem 6

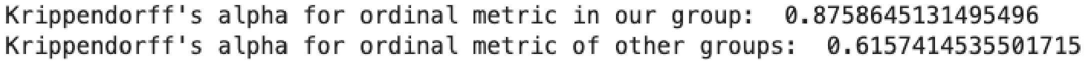
1. The three scores are shown below:



```
Krippendorff's alpha for nominal metric:  0.7285563972275997
Krippendorff's alpha for ordinal metric:  0.877237764754488
Correlation:
 [[1.         0.81942049]
 [0.81942049 1.        ]]
```

2. In the scores of our group: the correlation score is 0.819. The α for nominal and ordinal level measurement are 0.728 and 0.877. We found that the correlation score is between those two **αs.**
3. We found that the nominal score is lower than the ordinal score. In our setting, I think ordinal level score should be used because our labels are ordinal data.

Problem 7



```
Krippendorff's alpha for ordinal metric in our group:  0.8758645131495496
Krippendorff's alpha for ordinal metric of other groups:  0.6157414535501715
```

Here, the reason why alpha in our group 0.876 which is slightly smaller than the 0.877 above is we used the original annotation data in the previous problem while using the summarized train data in the problem7. We found that the alpha of all the other annotators on our group items is obviously lower than our alpha. The major reason for the difference is that our definition of label 5 is different from other groups' definitions. In our definition, formal replies with any reasonable explanation can be considered to be 5. However, different groups can have different opinions. And in Problem8, we analyze the difference in detail.

Problem 8: Analysis of annotator agreement and the disagreements and the different annotation guidelines:

1. Question ID: t3_n68hwz

   Question：What are some unique and harmless pranks to play on your coworkers in an office setting?

   Reply ID: gx5l1aw

   Reply：Replace every m key with the n key and replace every n key with the m key. Some will call you a monster but they will call you a nomster.

Group 25		Group 01	
User 34	User 35	User 13	User 14
4.0	4.0	1.0	2.0

   Why Different: Group 25 considers that this reply has some extra explanation in its second sentence, and the first sentence answers the question, so it is graded 4. But for Group 01, they have some different criterias for different types of questions. In this question, this is a "What" type question, they considered the helpfulness as to what extent the answers align with the question and share the appropriate information with the person who asked the question. User 13 may think that the answer does not contain any keywords related to the question, is playing tricks with wording and circle around the question and is less helpful. User 14 may think the reply is without any other information in the context explained. In general, the difference in the understanding of "additional explanation" was the main reason for the difference in ratings between the two groups, with Group 25 considering any addition after a very short answer to be an "additional explanation", while Group 01 did not.

Need to change the rating? Why? : No. Our final true rating is 4.0.

2. Question ID: t3_n46meh

   Question：LPT Request: How to be more participative in classes?

   Reply ID: gwu3k44

   Reply：If somebody else says something you can always say “I really like what x said about y, I agree with this point a lot and &lt;something something that rehashes what that person said&gt;”

Group 25		Group 07		
User 34	User 35	User 40	User 41	User 42
3.0	2.0	5.0	5.0	5.0

   Why Different: For Group 25, user 34 may think the replay answers the question, but it is a short answer, do not provide any further explanation, user 35 may think the answer is not completely answered. For Group 07, user 40, 41, 42 have same rating 5, they may think this question satisfy their all criterias, relevant to the subject of question, clearly articulates a claim and demonstrates solid use of supporting evidence, also the writing does not affect reading comprehension and does not contain parts that needs further research. In general, Group 25 has a clear requirement for additional explanations, while Group 7 does not, which is the main reason for the difference in rating.

   Need to change the rating? Why? : No. Our final true rating is 2.0, because we think the reply does answer part of the question, just one scenario of the situation, which is not enough to answer the question.

3. Question ID: t3_ni01ho

   Question：People who lie to embellish stories about how you were mistreated by someone, why not tell the truth?

   Reply ID: gyz3vbs

   Reply：Pfft, so anyway I’m reading Reddit this user u/Normalizesteroidz comes out of nowhere and demands this answer to an r/Askreddit question and is like super in my face about it. Honestly it was so confrontational and aggressive… I just got the heck outta there before it got ugly. I think they knew I wasn’t playing around and backed off. They’re honestly lucky I didn’t downvote them. People were all cheering me on too.

Group 25		Group 01	
User 34	User 35	User 14	User 13
5.0	5.0	4.0	1.0

Why Different: For Group 25, they think this reply just answers the question by sharing the user's own experience, and have extra explanations, the words length > 300. However, for Group 01, user 14 also thinks the reply understood the question and provided predominant rationale or justifications behind the answer, for question "why", the answer replied predominantly including an explanation for the given scenario and the length of the reply greater than or equal to 300 characters and less than 500 characters. But user 13 gave 1.0, they may think that the reply fails to obtain a thorough explanation for the given scenario. In general, the difference in the length of the responses resulted in different ratings. Group 01 felt that a rating of 5 was required for a 500+ word count, but Group 25 did not.

Need to change the rating? Why? : No. Our final true rating is 5.0.

4. Question ID: t3_n5krrt

Question：What desperate attempt has the company you worked for tried to keep you around?

Reply ID: gx1s26q

Reply：When I was 19 I was working at a car dealership in the service department as a lot attendant (park cars, clean the service drive through area, bring cars out to customers when they come to pick them up, sometimes help move new vehicles, did some filing in spare time), I was supposed to be paid biweekly and they said my first pay would be after a month because of some filing issue. Well I didn't get my paycheck after the first month and I needed to pay bills so I talked to my manager about it and she just wrote me a check, not a paycheck with taxes and deductions taken off. I wasn't happy about it but I needed the money so I took it. 2 weeks go by and I'm supposed to be getting a paycheck again but there's still some sort of administrative issue so my manager told me she would have it to me by the end of the next week. Well surprise, surprise there was another issue, I told my manager I couldn't keep working there if I wasn't being paid properly, she offered me a position as a detailer (better pay and if I can do jobs faster I get more money, which is ideal for me because I am a fast worker) if I waited it out another week. So I said I would. The next week rolled around and I did get my paycheck, so I was happy about that, but when I went to ask my manager about the detailing job she said I would have to wait until another girl finished up her last 2 weeks. I agreed again, because I really wanted the detailing job, and the next week they hired another lot attendant (I assumed to take over my position when I moved to detailing). But he was useless, he barely did any work so I had to run around doing his job too. I voiced my frustration to my manager and she said that I would be in detailing soon, so try not to worry. Well the 2 weeks passed quickly and I asked my manager again about the detailing position, when she decided to tell me they already hired someone to fill the spot, but I was so great as a lot attendant and I should really stay in that position. I quit on the spot.

Group 25		Group 08	
User 34	User 35	User 38	User 39

5.0	5.0	1.0	4.0

Why Different: For Group 25, they think the reply is complete and covers lots of details of experience, so rates 5. However, for Group 8, they separate the questions to objective and subjective. For this subjective question, user 38 may think that the reply raises some disrespectful response, user 39 may think the reply gives a short story or personal experience to support the opinion, so get 4. In general, Group 08 has given more thought to the categorization of questions and has placed higher demands on the quality of responses. "The answer is attracted to read with meaningful content and supportive explanation. Reasoning: Even if the question is an imaginary setting, the response still provides meaningful content (where inspires some thoughts and reflections), and lead people to think about the depth part of the question, or the content itself is educational. ", meet this criteria can get 5.

Need to change the rating? Why? : Yes. Our final true rating is 4.0 now. Here, we are inspired by Group 08 that the disrespectful content should reduce some rating, because we did not notice the disrespectful content in this reply before, but this should be highlighted in the point 4 rating.

5. Question ID: t3_nnydy0

   Question：People who don't watch trailers for upcoming movies/series you know you will be interested in, Why?

   Reply ID: gzx54pc

   Reply：Because I don't want to see the best bits before I'm supposed to.

Group 25		Group 13	
User 34	User 35	User 08	User 09
2.0	3.0	5.0	5.0

Why Different: For Group 25, user 34 may think this reply just answers a part of the question, user 35 may think this answer just answers the question but did not offer extra explanation. However, for Group 13, user 08 and 09 both rated 5.0, this may think this reply satisfies all the guidelines including length, on topic, good quality, and on offensive joking…content. In general, Group 13 tries to give rating based on different aspects but points are not deducted strictly based on whether the answer provides additional explanation, which is different from Group 25.

Need to change the rating? Why? : No. Our final true rating is 3.0, because the reply does answer the whole question although this is a short answer.

6. Question ID:t 3_n3e4qy

   Question：How to deal with jealousy?

   Reply ID: gwpafi3

Reply : Your girlfriend is never going to meet Billie Eilish, date her, or dump you for her. This is reality. You are not the only person your girlfriend will ever find attractive. This is also reality. Right now your fragility and lack of perspective on reality is honestly very concerning. Are you talking to a counselor or therapist who can help you with these things?

Group 25		Group 16	
User 34	User 35	User 48	User 49
4.0	4.0	2.0	1.0

Why Different: Group 25 thinks that the reply contains the complete answer and gives some extra explanation and also suggestions so gave 4.0. For Group 16, user 48 may think the reply contains some inside jokes or references that may not be obvious to an outside reader, user 49 may think the reply is not at all related to the question so gave 1.0. In general, the different understanding of this reply by different annotators may cause this question's rating difference.

Need to change the rating? Why? : No. Our final true rating is 4.0.

7. Question ID: t3_ng6ios

   Question : People who actually enjoy their job, where do you work and what do you do?

   Reply ID: gypdir8

   Reply : I work at a furniture store. It's a lot of work moving around furniture all the time but it's fun! The only real headache that comes with the job is the customers who don't know how to park, that come with people in the back seat full knowing they're picking up sofas of desks or a bunch of stuff, or people that get mad at us because they brought a vehicle too small for a clearance item the purchased.

Group 25		Group 11		
User 34	User 35	User 56	User 57	User 58
5.0	5.0	2.0	3.0	3.0

Why Different: For Group 25, they think the reply answers the whole question and also give some extra explanation, also the length of reply is long so rated 5.0. For Group 11, users 57 and 58 may think the reply does not provide some additional contextual information (examples, resources) so gave 3.0, user 56 rated 2.0, this may be because the reply does not say why their job is fun so providing an incomplete answer. In general, the different definitions of additional explanations are the reason for the different ratings, for group 25, theory, explanation, example, instruction, reference etc can all be considered as extra details.

Need to change the rating? Why? : No. Our final true rating is 5.0.

8. Question ID: t3_n7zcec

Question：What is a scientific phenomenon that seems fake but is actually real?

Reply ID: gxfjiol

Reply：That there is an above zero probability of you fazing through something solid

Group 25		Group 07		
User 34	User 35	User 40	User 41	User 42
2.0	3.0	5.0	5.0	4.0

Why Different: For Group 25, user 35 may think the replay answers the question, but it is a short answer, do not provide any further explanation, user 34 may think the answer is not completely answered. For Group 07, user 40, 41 have same rating 5, they may think this question satisfy their all criterias, relevant to the subject of question, clearly articulates a claim and demonstrates solid use of supporting evidence, also the writing does not affect reading comprehension and does not contain parts that needs further research, user 42 may think the reply need some extra parts. In general, Group 25 has a clear requirement for additional explanations, while Group 7 does not, which is the main reason for the difference in rating.

Need to change the rating? Why? : No. Our final true rating is 2.0. Because this does answer some of the questions but not very effectively.

9. Question ID: t3_n6szbe

Question：Look for books to know and understand women?

Reply ID: gx9h02w

Reply：Well done you for addressing this problem so young; that says a lot about your character! It's simple really; women are individuals, people, just like you; so read widely, and try to read as many good female authors as you can. \[not all female writers are feminist, some are writing books as bad as the worst male porn-obsessed authors.\] I suggest Girl, Woman, Other by Bernadine Evaristo; all about women of all ages and types in London, with their common thread being they are all POC. It's a very funny, interesting and sometimes shocking story. I think you'd find it fascinating! If you want to increase your understanding of how society all over the world treats women, read Rebecca Solnit's Men Explain Things To Me. It's quite a horrifying read about all the ways women are treated badly, by the law, by society, by the media. It's in 7 shortish essays so you don't have to read it in one go! Good luck.

Group 25		Group 08	
User 34	User 35	User 38	User 39
5.0	5.0	1.0	5.0

Why Different: For Group 25, they think the reply is complete and covers lots of details of extra explanation, so rates 5. However, for Group 8, they separate the questions to objective and subjective. For this objective question, user 38 may think that the reply has too many personal opinions so rates 1.0, user 39 think this has full explanation with explanations so rates 5.0. In general, Group 08 has given more thought to the categorization of questions and has placed higher demands on the quality of responses, too much personal opinion and offensive content can result in getting a very low rating.

Need to change the rating? Why? : Yes. Our final true rating is 4.0 now. We are inspired by Group 08 that different types of questions (subjective or objective) can have different criterias, we need to consider the personal opinion and offensive content when rating.

10. Question ID: t3_n9zw6j

Question：what states have the most spanish speakers?

Reply ID: gxqo8k6

Reply：Per capita I'd guess border states - California, Arizona, Texas, and New Mexico. Probably California in absolute numbers, just because so many people live there *and* it's a border state.

Group 25		Group 01	
User 34	User 35	User 14	User 13
4.0	4.0	3.0	1.0

Why Different: Group 25 considers that this reply has some extra explanation, and the first sentence answers the question, so it gets rating 4. However, for Group 01, User 14 may think that the reply has some information explained, user 13 may think this answer is not helpful and fails to directly answer the question. In general, the difference in the length of the responses resulted in different ratings. Group 01 felt that a rating of 4 was required for a 300+ word count, but Group 25 did not.

Need to change the rating? Why? : No. Our final true rating is 4.0.

Problem 9: Specific improvements that could be made to our own guidelines:

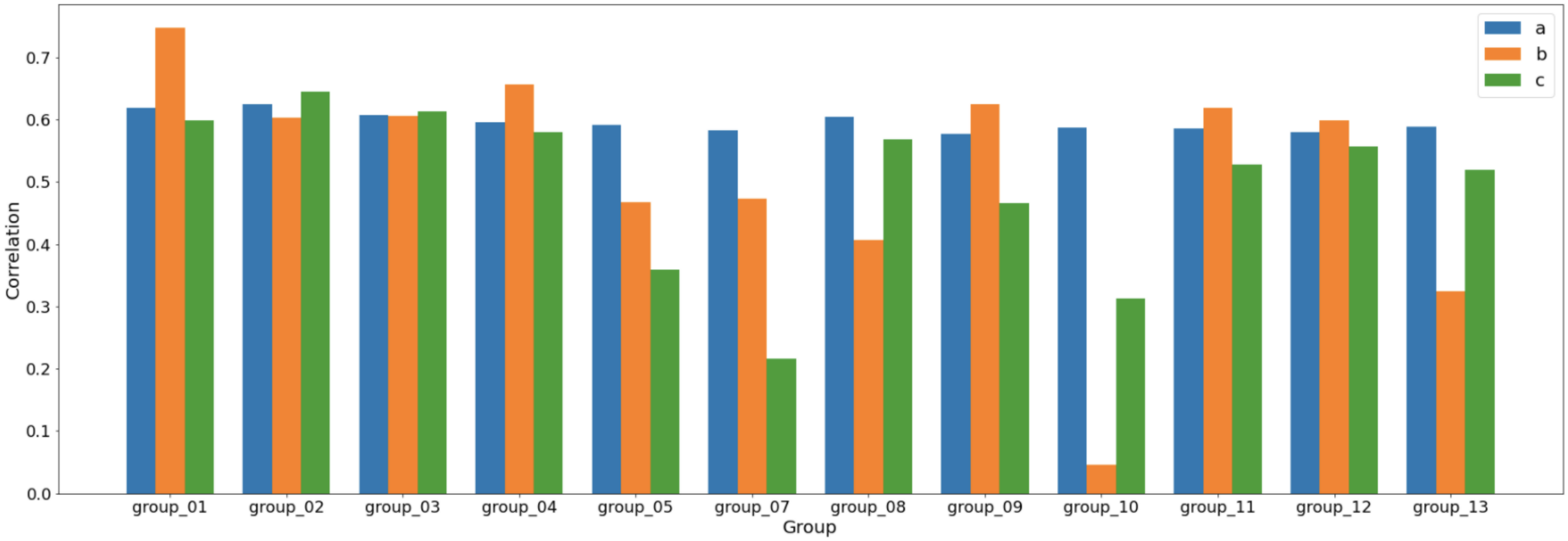
- For different rating scores, we need to specify the range of text lengths for the responses, for example, 500+ for (5), 500-300 for (4), 300- for (3)(2)(1);
- We should consider more carefully the different types of questions that can cause different scoring guidelines. For different types of questions (subjective or objective, different types of questioning, such as what, why, how...). Improve more detailed scoring rules by categorizing them, so that scoring is more reasonable.
- We should consider additional factors that may affect the rating, such as on or off-topic, well formedness of responses, offense, harassing, or illegal responses , sarcastic, joking, or misinformative responses, negated or guessed responses.

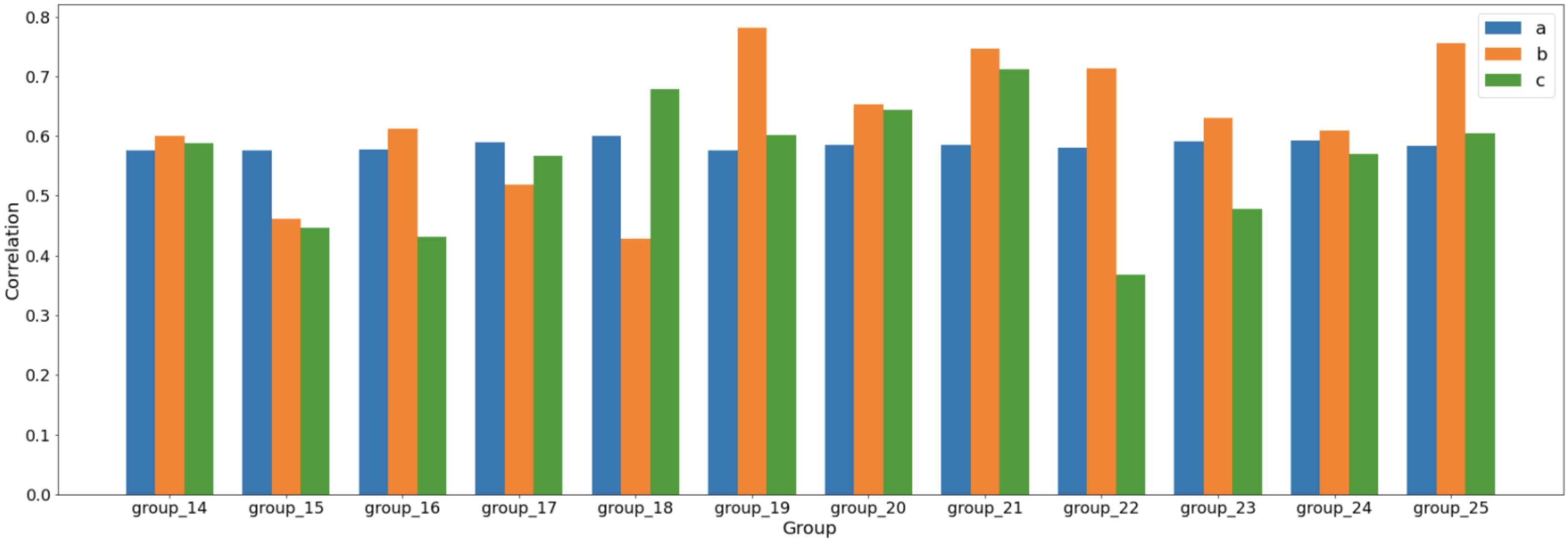
Problem 11 & 12

Please check the submitted notebook for any detail of the trained model.
The kaggle name is **Group25**

Problem 13 Evaluation

Since we have too many groups, I separated the graph into two parts. Here, we found that group_10 has very low correlation in their devset b, which means the model cannot predict the labels of Group10 well. The reason for that is because they did not have good agreements in their group's annotation. We checked their annotation agreement, the ordinal α is just 0.2029, which is pretty low. We also checked their group guideline and did not find some huge gaps between ours. Some group members of group 10 might not understand the guideline well. Thus, the major reason for the low score of their annotations is the disagreement within the group.


Correlation
a
b
c
group_01
group_02
group_03
group_04
group_05
group_07
group_08
group_09
group_10
group_11
group_12
group_13
Group


Correlation
a
b
c
group_14
group_15
group_16
group_17
group_18
group_19
group_20
group_21
group_22
group_23
group_24
group_25
Group