SI 630: Homework 3

Data Annotation and Large Language Models

Yiwen Yang (yangiwen), Yu Yan (kuminia), Zheng Yuan (yuazheng)

Kaggle username: yangiwen

March 31, 2023

Problem 6

1. Compute r and the compute α using the ordinal and nominal level of measurements for the group member's annotations and report them (three scores total).

Here is the r for our group result.

annotator annotator	kuminia	yangiwen	yuazheng
kuminia	1.00000	0.48331	0.48832
yangiwen	0.48331	1.00000	0.71069
yuazheng	0.48832	0.71069	1.00000

Here are the two α

```
simpledorff.calculate_krippendorffs_alpha_for_df(group_data,experiment_col=['Premise_ID','Argument_ID'],
                                                 annotator_col='annotator',
                                                 class_col='persuasiveness')
```

```
0.22394745686275064
```

```
def interval_metric(x,y):
    return (x-y)**2

simpledorff.calculate_krippendorffs_alpha_for_df(group_data,experiment_col=['Premise_ID','Argument_ID'],
                                                 annotator_col='annotator',
                                                 class_col='persuasiveness',
                                                 metric_fn = interval_metric)
```

```
0.54558144167421
```

2. In 2-3 sentences, comment on the difference (if any) between your group r and the α scores. Which is higher and what do you think this means?

Overall, our r score is higher, with a correlation coefficient of 0.71 between Yuazheng and Yangiwen, and a correlation coefficient of nearly 0.5 between Kuminia and the two. Therefore, there is a relatively large difference in opinions between Kuminia and the other two, but the range is still acceptable. From the perspective of Krippendorff's α, the highest score is the ordinal level of α, which we obtained as 0.546, while our normial level is only 0.224. In conclusion, the correlation between our annotation scores is better than the inter-rater reliability and agreement among our annotations.

3. In 2-3 sentences, comment on the difference (if any) between your group's ordinal and nominal α scores. Which is higher and what do you think this means? Which one should you use in practice to measure agreement in this setting?

For the normial level, we get 0.224 and for the ordinal level we get 0.546. After investigation, we found that an ordinal level Krippendorff's α value between 0.4 and 0.6 is considered moderate consistency, indicating that our results have some degree of correlation. The α value of the normial level at 0.224 indicates that our data results would be better in an ordinal situation. We think we should use the ordinal level for the project, that is because we get a higher score in this situation, therefore, we also choose regressor for the part 2 work.

Problem 7

annotator_x	group-02_person-0	group-02_person-1	group-02_person-2	group-03_person-0	group-03_person-1	group-08_person-0	group-08_person-1	group-08_person-2	group-09_person-0	group-09_person-1	group-09_person-2	group-17_person-0
annotator_x												
group-02_person-0	1.000000	0.614138	0.797407	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
group-02_person-1	0.614138	1.000000	0.628971	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
group-02_person-2	0.797407	0.628971	1.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
group-03_person-0	NaN	NaN	NaN	1.000000	0.878775	NaN	NaN	NaN	NaN	NaN	NaN	NaN
group-03_person-1	NaN	NaN	NaN	0.878775	1.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN
group-08_person-0	NaN	NaN	NaN	NaN	NaN	1.000000	0.493671	0.629804	NaN	NaN	NaN	NaN
group-08_person-1	NaN	NaN	NaN	NaN	NaN	0.493671	1.000000	0.723596	NaN	NaN	NaN	NaN
group-08_person-2	NaN	NaN	NaN	NaN	NaN	0.629804	0.723596	1.000000	NaN	NaN	NaN	NaN
group-09_person-0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.000000	0.467707	0.580288	NaN
group-09_person-1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.467707	1.000000	0.271405	NaN
group-09_person-2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.580288	0.271405	1.000000	NaN

```
simpledorff.calculate_krippendorffs_alpha_for_df(
    problem_7,
    experiment_col=['Premise_ID','Argument_ID'],
    annotator_col='annotator_x',
    class_col='persuasiveness_x'
)
```

```
0.3177428982759226
```

```
simpledorff.calculate_krippendorffs_alpha_for_df(
    problem_7,
    experiment_col=['Premise_ID','Argument_ID'],
    annotator_col='annotator_x',
    class_col='persuasiveness_x',
    metric_fn = interval_metric
)
```

```
0.5807992465016146
```

Overall, we found that our ordinal and nominal level Krippendorff's α were slightly lower than those of other groups, but we believe that the range is acceptable.

Additionally, considering the correlation coefficients, we found that our group's level of correlation is fairly moderate, with most groups' coefficients falling between 0.4 and 0.75, which is consistent with our results.

Through comparing the differences in guidelines among the various groups, we discovered that the potential reason for our lower results may be due to our guidelines not including three specific parts: emotional appeal, overall writing style, and neutral language of argument. In future guideline adjustments, we will focus on addressing these three points.

Problem 8

This is the ID and scores for the texts have the biggest absolute difference in mean rating between your group and some other group. The persuasiveness is our group's scores and the persuasiveness_x is others. The diff is the absolute difference.

Premise_ID	Argument_ID	persuasiveness	changeability	premise_agreement	persuasiveness_x	changeability_x	premise_agreement_x	diff
m6j48e	gr8b7e6	4.000000	3.666667	3.333333	1.0	5.0	5.0	3.000000
e62jxn	f9on1t1	3.666667	2.666667	5.000000	1.0	2.5	2.0	2.666667
lgeqip	gmqz16l	2.333333	3.666667	4.000000	5.0	3.0	3.0	2.666667
heirw1	fvrh8zt	2.666667	2.333333	2.666667	5.0	1.0	2.0	2.333333
cmrrgp	ew4hclg	1.333333	4.000000	3.000000	3.5	3.0	3.5	2.166667
o9pyz7	h3d4kwe	1.333333	1.000000	4.333333	3.5	2.5	3.5	2.166667
h8mamc	furo6om	3.666667	3.333333	1.666667	1.5	2.0	1.0	2.166667
jyvfqs	gd8o2kz	2.000000	2.333333	4.333333	4.0	1.5	4.5	2.000000
gihcx1	fqem647	1.000000	2.333333	3.000000	3.0	2.5	3.0	2.000000
gez1fy	fpque4g	2.000000	3.333333	3.333333	4.0	2.5	4.0	2.000000

Here is the detail for the arguments of the ten texts:

1. To be honest based on a lot of your replies to others you simply don't want to pay or have paid for something that doesn't benefit you as much as it may benefit someone else.
You signed up for that loan knowing exactly the same as everyone else who did, you had the option to party every weekend knowing that the government may cancel student loan debt but you chose not to in the hopes you made the right decision. Sometimes you make a decision that could have played out better, you're still benefitting from the

cancellation.

Think of it in terms of the stock market as its very relevant right now. We're both giving a book about 4 different stocks, we both have the same information and amount of money we have to invest with. You read this book 5 times over, analysing data and patterns and put as much as you can afford into a sure fire winner. I skim the book, see something I like the idea of, put a few $ in and hope for the best, the stock i chose booms and I gain like 4000%, you chose a steady increase stock gaining 15% a year.

Are you being punished because you chose what to do with your money when we both had the same information? No you're just not benefitting as much as me based on your own thought process and decision.

Picture this situation with no student debt you're saving 2k a month and you're friend is partying instead of saving, you're choosing not to have that lifestyle and save for your future, your friend is choosing to have a good time now with the hopes they won't fall on hard times. Really the most sensible option is you should both find a nice balance between the 2.

Our group score: 4.0

Group 17 score: 1.0

Our group gave a high score because we believed that the argument covered the content described in Premise Justification within the context and the presented points were relatively strong (although one small point was not fully covered in Premise Justification). Therefore, the final average score was 4.

After comparing our guidelines with Group 17, we found that their group also considered emotional appeal as a bonus and would increase the score when they perceived stronger emotional appeal. Additionally, their changeability for this issue was 5, which means that the score they gave was subject to change.

After our discussion, we ultimately gave scores of 3, 4, and 4, with an average score of

3.67. However, even with this, there were still significant score differences between our group and Group 17.

2. > One way you can change my mind is this: Name a law set by a nation that, if broken, would not eventually lead to violence being used against you if you refused to comply with all subsequent punishments resulting from that crime.
/r/MasterGrok mentions symbolic laws and suicide to which you responded
> I would argue that a law that isn't enforced isn't really a law at all
?????????

Our group score: 3.67
Group 3 score: 1.0

Compared to the third group, our group did not consider overall writing style as a factor in scoring, and we did not use it as a reference for grading. The garbled text in the previous answer may have caused the third group to give a lower score. Additionally, we found that the third group considered the length of the argument as a criterion for grading. Perhaps the length of the argument was not long enough for them, but after discussion, we did not think that length was a very important criterion since length does not necessarily represent quality. The third group also considered emotional appeal, which we think is one of the elements that we need to consider in the future. The previous answer did not evoke a strong emotional appeal. After discussion, we finally gave scores of 3, 2, and 1, with a tie being scored as 2, which was closer to the third group's scores.

3. Wages are already tax deductible. As is that payroll tax. Is your goal to tax businesses more?

Our group score: 2.33

Group 21 score: 5.0

We believe that this is a very subjective judgment. The author used a sarcastic approach to respond to the premise, and although the answer was intense, it did not fully cover all the content of premise justification. After comparing, we found that there was not much difference between our guidelines and those of group 21. Therefore, we think that the difference in the scores we gave may be due to our different life experiences. Finally, we did not adjust the scores given, and the average score remained at 2.33.

4. Has every single person you've ever helped turned on you? Have you ever received help from another person when you needed it?
Nobody can guarantee that others will always be there for you no matter how much you help or care about them, but nobody can make it through life alone. Helping others makes it **more likely** that others will be there to help you when you need it.
Plus, it feels good to help other people, and that's not wrong.

Our group score: 2.67

Group 17 score: 5.0

I think that the emotional appeal of this argument may have led group 17 to give a high score. However, if we do not consider emotional appeal (which is not included in our guidelines), this answer is just a very subjective one. Different people have different feelings about helping others, and it is not necessarily true that people need others to survive. These are points that our group cannot agree with. In the end, we gave scores of 3, 3, and 3, with an average score of 3, because we considered emotional appeal.

5. I've been saying this since 2017. Sanders/Warren 2020

Our group score: 1.33

Group 17 score: 3.5

In fact, the members of our group are not very familiar with the political issues in the United States, so many related issues need to be searched; Our changeability also reaches 4 for this argument. In addition, the answer is very brief, which seems to have nothing to do with the premise justification. Finally we gave a score of 2,2,1, with an average score of 1.67.

6. I don't see why this has to be a dichotomy. All the fears about cures and tests can be true, yet those tests and cures can also be beneficial to many people. Just because you fear what preventative measures can do to people who currently have that disability, doesn't mean you automatically support banning the preventative procedure or support boycotting it?

Our group score: 1.33

Group 18 score: 3.5

The 18th group includes "Neutral language of argument" as one of the criteria for judging, and we believe that this could lead to a high score for this argument as it demonstrates a very neutral response. Additionally, they also include "No sarcasm (or 'trying to one up') in argument" as a scoring point, which was mentioned in the gray area. Often, sarcasm can lead to misunderstandings, especially for non-native English speakers, although it can enhance the effectiveness of language expression. This is an area for improvement.

The final scores have been adjusted to 2, 3, and 1, with an average score of 2.

7. The areas of the brain associated with the feet and genitals are adjacent and otherwise highly related. There are plenty much more comical kinks that don't have nearly as much natural relation to sexual areas of the brain without being mostly socially learned. Some people even experience orgasms by having their feet touched. For many it's not even the physical appearance of the foot, but merely the humiliation of putting so much attention into what's commonly considered a gross part of the body. Others simply do see charms in their appearance, and it's not really something that can be explained. It would be like explaining to a horrific alien why humans are attractive. If you don't see it as attractive, you won't be able to see the appeal.
If you're looking to have your view changed in a way that suddenly makes you personally find feet attractive, that probably isn't going to happen.

Our group score: 3.67
Group 17 score: 1.5

The Emotional appeal still plays an important role in Group 17's evaluation criteria. In addition, we believe that gender may also have some influence on the evaluation, although we do not know the specific composition of Group 17. As all members of our group are male, this may lead to bias in our evaluation of topics related to gender, resulting in differences in scores.

After discussion, we decided not to adjust the scores and the average remained at 1.5.

8. When audiences say there are too many women in a film when they are 50%, even as extras, we are a long way off just having enough women to represent reality. Geena Davis wrote an [op ed on this](https://www.hollywoodreporter.com/news/geena-davis-two-easy-steps-664573) a while ago that might change your mind.

Our group score: 2.0

Group 21 score: 4.0

As we have previously compared our guidelines with Group 21 and found them to be very similar, we believe that one of the main reasons for the score discrepancy may be related to gender differences. After discussing, we have decided to adjust the scores to 2, 3, and 2, resulting in an average score of 2.33.

9. Exactly how many people would be enough?

Our group score: 1.0
Group 21 score: 3.0

Our group believes that this argument is full of sarcasm but ultimately lacks a strong argument. It fails to provide enough evidence to support or refute the viewpoint presented in the premise justification. Therefore, our group unanimously gives this argument a score of 1. We will maintain our stance and give scores of 1, 1, and 1, resulting in an average score of 1.

10. this is a dumb argument. a person can major in pre-med and go on to do a career other than surgery. let a man get the accommodations he needs. he'll figure out whether surgery is or isn't the right profession for him when the time comes. no need to fuck up his undergradiate degree over this.
fwiw, i'm a psychotherapist. i talk with psychiatrists a lot. there are a shit load of anxious, narcissistic, personality disordered people in both fields. they do shitty things in their professional lives. that's how it goes. tons of scumbag lawyers out there. people are flawed. seriously flawed.

Our group score: 2.0

Group 21 score: 4.0

This argument is quite emotionally charged, and the last sentence, "people are flawed. seriously flawed," is particularly aggressive. While it does refute the initial point, it did not receive consensus among our group members. Therefore, we decided to adjust the scores to 3, 4, and 2, with an average score of 3, taking into account the emotional tone of the argument.

Problem 9

• **Emotional appeal:** An argument that is emotionally appealing can be highly persuasive. We will award higher scores for the use of more emotional adjectives and expressions that evoke a strong emotional response in the reader and create a sense of empathy.

• **Overall writing style:** A better writing style can improve readability and make an argument appear more compelling, so we will give higher scores for arguments with a strong writing style.

• **No sarcasm expression：** We believe that sarcastic sentences can cause some misunderstandings, which we have mentioned in the "gray area" section of our guidelines. Therefore, we will lower the scores for expressions with sarcasm appropriately.

• **Citing specific evidence and examples to support one's argument (Bonus):** If an argument cites well-known or authoritative viewpoints, we will give it a higher score.

• **Structure (Bonus):** We believe that a well-structured argument can increase its strength, for example, having a clear thesis statement and sufficient evidence to support it, and finally presenting a well-rounded conclusion. Such arguments will receive higher scores.

Problem 10

We use the Trainer class from huggingface, and set the parameters list below.

```
from transformers import AutoTokenizer, AutoModelForSequenceClassification, DataCollatorWithPadding
from torch.utils.data import DataLoader

BASE_MODEL = "microsoft/MiniLM-L12-H384-uncased"
LEARNING_RATE = 1e-2
MAX_LENGTH = 256
BATCH_SIZE = 16
EPOCHS = 20

tokenizer = AutoTokenizer.from_pretrained(BASE_MODEL)
model = AutoModelForSequenceClassification.from_pretrained(BASE_MODEL, num_labels=1, ignore_mismatched_sizes=True)
```

```
from transformers import TrainingArguments, Trainer

training_args = TrainingArguments(
    output_dir="SI630",  # output directory
    learning_rate=1e-5,
    num_train_epochs=10,  # total number of training epochs
    per_device_train_batch_size=BATCH_SIZE,  # batch size per device during training
    per_device_eval_batch_size=8,  # batch size for evaluation
    weight_decay=0.01,  # strength of weight decay
    do_eval=True,
    report_to="none",
    evaluation_strategy="steps",
    eval_steps=100,
)
```

Problem 11

Here is the result we get.

[1510/1510 02:43, Epoch 10/10]

Step	Training Loss	Validation Loss	Mse	Rmse	Mae	Accuracy
100	No log	2.574574	2.574574	1.604548	1.373500	0.317215
200	No log	0.917579	0.917579	0.957903	0.768244	0.452611
300	No log	0.591500	0.591500	0.769090	0.601562	0.475822
400	No log	0.473215	0.473215	0.687906	0.523620	0.541586
500	1.712800	0.530342	0.530342	0.728246	0.555293	0.539652
600	1.712800	0.465432	0.465432	0.682226	0.523129	0.558994
700	1.712800	0.454745	0.454745	0.674348	0.520894	0.541586
800	1.712800	0.471316	0.471316	0.686525	0.531385	0.539652
900	1.712800	0.489198	0.489198	0.699427	0.533843	0.557060
1000	0.416900	0.527170	0.527170	0.726065	0.556198	0.541586
1100	0.416900	0.472400	0.472400	0.687314	0.533005	0.541586
1200	0.416900	0.486160	0.486160	0.697252	0.538096	0.547389
1300	0.416900	0.503015	0.503015	0.709236	0.545652	0.541586
1400	0.416900	0.499521	0.499521	0.706768	0.544505	0.535783
1500	0.357800	0.488904	0.488904	0.699217	0.539102	0.535783

Problem 12

Here is the score we got on Kaggle; the result is 0.7084.

Problem 13

From the bar plot of model's performance over three scores per group, we can see that some groups excel in their first score (taking group average) over other scores, such as group {2,4,6,15}. From their guidelines, we observe that some extra explicit rules may help a lot with annotation. From group 2's guideline, they use rubric that deduct points on aggressive or offensive words; From group 4's guideline, they explicitly specify how to deduct or plus points based on more detailed patterns rather than high-level justifications. Group 5 has the worst performance when taking in-group average as ground truth, and group 14 does not have enough data to analyze. Overall, most group achieve better score in the first and second set than the third set. As only considering

in-group average score leads to a big variance in performance, considering all other groups always give a more stable performance. From the last score, we can see that ratings from a particular group are generally more predictable compared to all other items that the group did not annotate.

Problem 14

From the lmplot, we can see that the linear model does not fit the points well. This implies that changeability in premises does not potentially affect the prediction errors. For all five-changeability score, they share very similar distribution as seen from the box plot. Most points lie between 0.5 and 1, which means that the difference between model rating and annotator rating is within one category unit. Although it is intuitive that changeability on premise may affect annotator's idea on the argument itself, from the data we collected it seems not to be the case. Annotators are meant to judge the argument given the premise, but not depending on how they will change their idea on the premise itself. So, it is a good sign that we can ignore changeability when doing the training part.