NLP HW 2: Annotation

Milton Lin and **Cole Molloy** and **Lois Wong**
Computer Science Department
Johns Hopkins University

1 First Round : annotation, challenges

1.1 Results

We have three annotation files. `Milton.csv`, `Lois.csv`, `Cole.csv`

Table 1: Fleiss' Kappa Scores for Annotations by Milton, Lois, and Cole

Category	Fleiss' Kappa
Tone	0.327
Expertise	0.022
Encouraging	0.300
Respectful	0.150

Table 2: Fleiss' Kappa Scores for Annotations by Milton and Lois

Category	Fleiss' Kappa
Tone	0.267
Expertise	-0.095
Encouraging	0.253
Respectful	0.228

Table 3: Fleiss' Kappa Scores for Annotations by Milton and Cole

Category	Fleiss' Kappa
Tone	0.278
Expertise	-0.011
Encouraging	0.322
Respectful	0.021

1.2 Other Metrics for Agreement

Below are tables for percentage agreement.

Table 4: Fleiss' Kappa Scores for Annotations by Lois and Cole

Category	Fleiss' Kappa
Tone	0.430
Expertise	0.083
Encouraging	0.322
Respectful	0.149

Table 5: Percentage Agreement for Everyone

Category	Percentage Agreement
Tone	73.33%
Expertise	35.83%
Encouraging	52.5%
Respectful	45.83%

Table 6: Percentage Agreement for Milton and Lois

Category	Percentage Agreement
Tone	80.83%
Expertise	51.67%
Encouraging	65.0%
Respectful	66.67%

Table 7: Percentage Agreement for Milton and Cole

Category	Percentage Agreement
Tone	80.83%
Expertise	55.83%
Encouraging	70.0%
Respectful	58.33%

Table 8: Percentage Agreement for Lois and Cole

Category	Percentage Agreement
Tone	85.0%
Expertise	64.17%
Encouraging	70.0%
Respectful	66.67%

1.3 Analysis

For our analysis, we focus on Fleiss' kappa for comparing annotator agreement because it takes into account chance when computing annotator agreement, while simple percentage agreement does not take chance agreement into account.

1.3.1 Parsing Complexity

Overall, our values for Fleiss' Kappa were relatively low. Our best categories, 'Tone' and 'Encouraging', had values of 0.327 and 0.300 respectively. These indicate that there is a greater than chance likelihood of annotator agreement, but is still fairly weak. We interpret this as an indication that the overall task is difficult for humans to perform consistently. We categorize this difficulty as "parsing complexity", where there is some latent feature of the task (perhaps the examples are unclear, or the annotations are hard for humans to apply consistently) that causes a great variability in the annotations from example to example.

1.3.2 Binary Options Limitation

The binary nature of some annotation options (e.g., encouraging vs. discouraging) fails to capture the nuanced reality of the comments, which leads to oversimplification. For example, there is ambiguity in whether 'encouraging' defined with respect to the speaker or audience. In some instances, the comments were stories/complaints about other people, so those comments could arguably be discouraging to the people/things they're complaining about but also reinforce the speaker's points which might be encouraging

1.3.3 Ambiguous Question Phrasing

Our disagreement in categorizing the "expertise" of examples highlight a potential challenge in the interpretation of the question. Amongst annotator pairs, the highest agreement we got received a Fleiss' Kappa of 0.083, indicating a performance similar to random chance. Such a divergence probably stem from ambiguous guidelines and subjective interpretations. For example, it is unclear in the question "Based on the comment, what do you think is the expertise of the speaker?" whether "speaker" refers to the person delivering the talk or the writer of the comment.

1.3.4 Ambiguity in Options

There is additional ambiguity in the meaning/definition of certain categories (e.g., respectful can mean either kind or formal). Different interpretations of such terms can affect consistency.

1.3.5 Mixed Messages

While annotating, we each noticed that certain examples contained sub-sections that exemplified conflicting tags. For instance, a response that begun positive and respectful could devolve into something more neutral and disrespectful. There is no guarantee that a response will uniformly exemplify any of the annotated labels, and as such, it becomes the annotator's responsibility to make judgements. This judgements often can be difficult to make, and we believe contributed to our overall disagreement.

1.4 Ethical Implications

One of the major ethical issues in this annotation task is the subjectivity of concepts such as respectfulness and tone. What is respectful can depend heavily on social cues that vary from culture to culture, so a more homogeneous group of annotators may get greater agreement than a diverse group of annotators. This could inadvertently cause researchers to favor less diverse annotators in order to get the greatest agreement. Additionally, these samples could potentially contain identifying information that could be used to identify them, violating their privacy. This especially becomes an issue depending on how consent was gained from the commentators, as not all people will be comfortable with their message being used for purposes outside of their original comment.

2 Second Round : Improving the Schema

2.1 Proposed Changes

2.1.1 Solution 1: Switching to Comparisons

One of the main issues we saw in annotating was simply parsing complexity, where something about the task made it hard to agree on. We thought that one change that could help with this was switching from an absolute judgement task, where the annotator has to decide on things such as "Strong Positive" vs. "Positive", to a relative judgement task,

where users are shown two samples and instead rank them against each other. This was partially motivated by methods from reinforcement learning from human feedback (Ouyang et al., 2022) where annotators rank from a selection of outputs rather than attempting to make a judgement call on a single example. One benefit of this approach is that users no longer need to make absolute judgements on where the line between each category is, but instead only needs to differentiate between two options. This also helps in samples with few examples as we move from 40 possibilities to $\binom{40}{2}$ options. This can be a benefit, where you are getting a richer feature space to describe the qualities of each response, but also can increase the number of annotations required. In order to maintain some level of consistency for comparing each schema, we also limited this annotation task to 40 samples of pairs, so we annotate the same number of questions as in part one.

2.1.2 Solution 2: Simplifying the Questions

While annotating the tone of the responses, we noticed that it was often difficult to differentiate the lines between a positive response and a strongly positive one. Additionally, we believed that differentiating between a person's education level was incredibly difficult, and we were often just randomly choosing an option (as demonstrated by our Fleiss' Kappa). In order to help this, in our new schema we tried to keep questions as simple as possible, limiting the possible responses to 3 for each question.

2.1.3 Solution 3: Improved Definitions and Priming

We defined 'respect' more explicitly to reduce ambiguity and included some questions to keep in mind:

Respect is defined as showing consideration and regard for something or someone. It is possible for a comment to respectfully disagree with the speaker. You should only be concerned about whether or not the comment respects the speaker; it is possible for the comment to be respectful of the speaker and disrespectful of other individuals. Some considerations to keep in mind include: 1. Does the comment show politeness or regard for the talk? 2. Does the comment explicitly offend the speaker? [1]

We also defined 'positive' to reduce ambiguity towards the term:

[1] But how do we know if it offends or not?

Positivity should be understood with respect to the original poster or speaker, as opposed to the content of the TED talk. Some considerations to keep in mind are: 1. Is the comment constructive of the speaker? 2. Do they find the speaker helpful or pleasant?

2.2 Results

Table 9: Fleiss' Kappa Scores for New Annotations by Milton, Lois, and Cole

Category	Fleiss' Kappa
Tone	0.559
Expertise	0.174
Encouraging	0.361
Respectful	0.356

Table 10: Fleiss' Kappa Scores for New Annotations by Milton and Lois

Category	Fleiss' Kappa
Tone	0.511
Expertise	0.042
Encouraging	0.309
Respectful	0.259

Table 11: Fleiss' Kappa Scores for New Annotations by Milton and Cole

Category	Fleiss' Kappa
Tone	0.506
Expertise	-0.074
Encouraging	0.174
Respectful	0.321

2.3 Analysis of results

Compared to problem 1, we see a universal improvement in Fleiss' kappa under the new annotation schema. When compared to the original schema, annotator agreement in "Tone" increased from 0.327 to 0.559, in "Expertise" increased from 0.022 to 0.174, in "Encouraging" increased from 0.3 to 0.361, and in "Respectful" increased from 0.150 to 0.361. It is hard to tell which of our modifications contributed to these performance gains, as each intervention applied affected multiple questions, but we can conclude that the new schema is overall better suited to annotator agreement.

Table 12: Fleiss' Kappa Scores for New Annotations by Lois and Cole

Category	Fleiss' Kappa
Tone	0.656
Expertise	0.419
Encouraging	0.574
Respectful	0.471

2.4 Critique of new annotation scheme

Below we discuss a few problems highlighted by the new annotation scheme. We end with suggestions for improvement.

2.4.1 Ambiguity in context

This task assumes a uniform context across comments, overlooking scenarios where the context inherently demands encouraging feedback. The comparison of "which text is more encouraging " should therefore not be included as a task.

2.4.2 Assessing expertise

Another notable difficulty lies in evaluating the speaker's level of informedness on the topic. This task remains challenging due to the subjective nature of expertise and the diverse ways it can manifest in speech. In practice, we should provide annotators with clear criteria or examples that define varying levels of expertise. However, if this is too time consuming, we believe this should also not be included as a task.

2.4.3 Conflation of respectfulness and positivity

The new annotation scheme highlights an issue with the semantic overlap between being respectful and conveying a positive tone. Perhaps it would be better to have chosen words that are more differentiable or provided better examples of the difference.

3 Advanced Analysis

Upon examining the data, we found 681 female speakers vs 722 male speakers which accounts for a relatively balanced dataset.

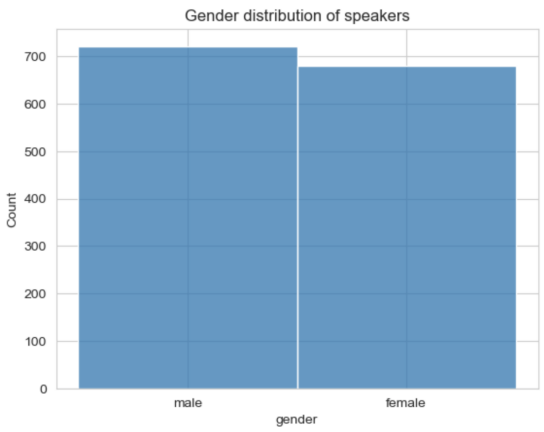

Figure 1: Gender distribution of all speakers

3.0.1 Respectfulness of speaker by gender

We elected to further analyze the respectfulness/disrespectfulness of comments addressed to female and male speakers because this was one of the categories we disagreed the most on as well as the category our preliminary analysis of this dataset showed to have the most variation.

Across all speakers, the dominant category of respectfulness is 'Respectful', followed closely by 'Neutral' and distantly by 'Disrespectful'. It can be noted that more female speakers were given respectful comments than male speakers, while more male speakers were given disrespectful comments than female speakers, and neutral comments seem pretty evenly distributed among female and male speakers.

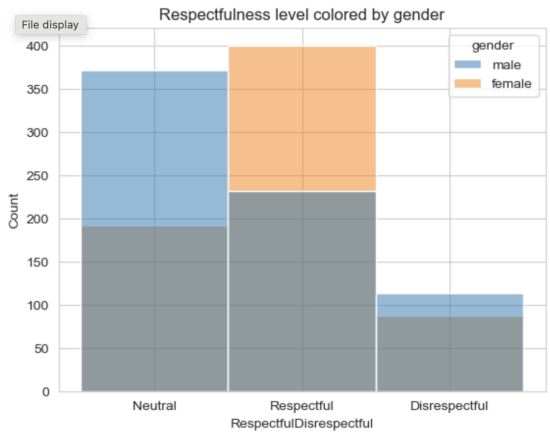

Figure 2: Respectfulness level colored by gender

The dominant category of respect among female speakers is 'respectful' (400 comments), followed by 'neutral' (192 comments), and 'disrespectful' (87 comments) trailing far behind.

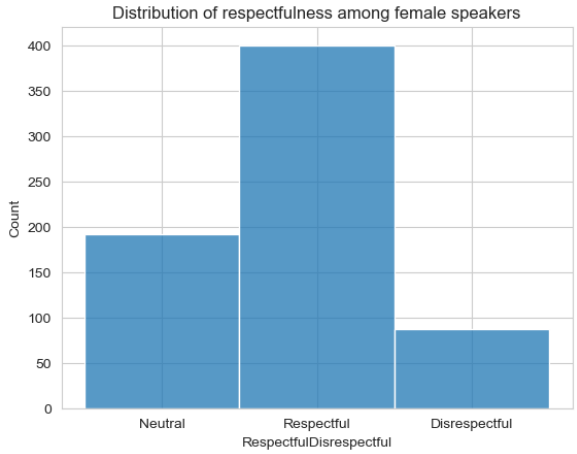

Figure 3: Distribution of respectfulness among female speakers

Among male speakers, the dominant category of respect is 'neutral' (372 comments), followed by 'respectful' (232 comments), and 'disrespectful' (113 comments).

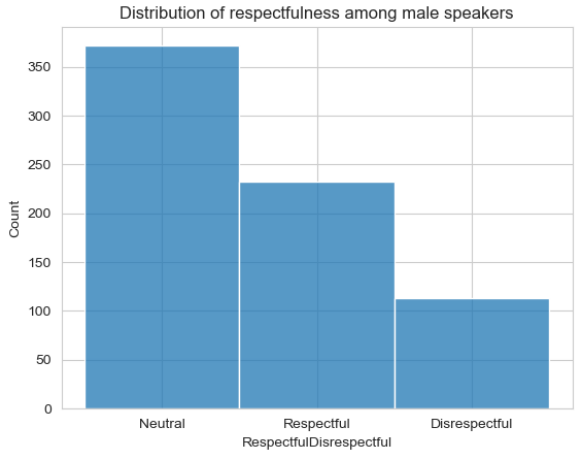

Figure 4: Distribution of respectfulness among male speakers

3.0.2 Expertise of speaker by gender

Across all speakers, the dominant category of perceived expertise is No degree, followed closely by Non-STEM degree and STEM degree. It can be noted that more female speakers were classified as having No degree than male speakers, while more male speakers were classified as having both Non-STEM and STEM degrees than female speakers.

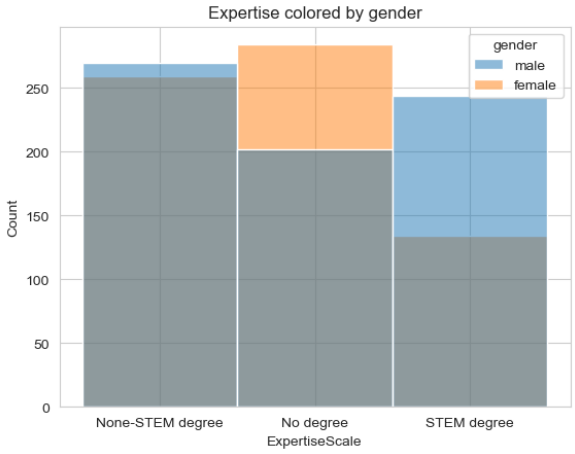

Figure 5: Expertise colored by gender

3.0.3 Tone of speaker by gender

Across all speakers, the dominant category of tone is None, followed by Positive, Strongly Positive, Negative, Neutral, and Strongly Negative in that order. While more males were given a comment classified as having the tones 'None', "Neutral', 'Strongly Negative', and 'Negative', more females were given comments classified as having 'Positive' and 'Strongly Positive' tones.

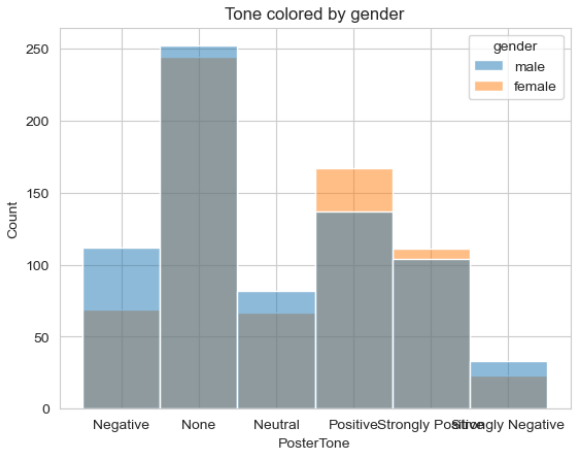

Figure 6: Tone colored by gender

3.0.4 Encouragement towards comment speaker by gender

Across all speakers, the dominant category of Encouragement in the comments is 'Encouraging', having 2x as many counts as 'Discouraging.' It can be noted that more female speakers were given encouraging comments than male speakers, and more male speakers were given discouraging comments than female speakers.

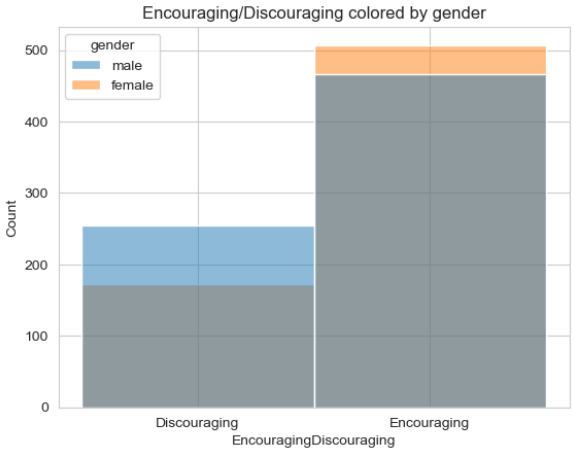

Figure 7: Encouragement colored by speaker

References

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.