Error Analysis

Automated Metrics:

1. We select the baseline models - Vaswani et al and Rushkin et al. Any other baselines that you find useful are okay too.
2. Fine-tune these models onto our dataset.
3. Compare the BLEU scores of these models with the BLEU score of our model.
4. We can also compare the perplexity.

(Other evaluations I think can be used are BLEURT, BARTScore.)
Let me know if you have anything to add.

Human Evaluation:

1. Important Part
2. Human-centric evaluations that I feel we can use:
    a. Krippendorff's alpha - this is something different - evaluates the disagreement between 2 annotators.
    b. Cohen's kappa - you guys know
    c. Fleiss's kappa - used to check the agreement between multiple annotators. We can use this if all of us are annotating. All the above mentioned can be imported.
3. This scenario works for only 1 annotation too. There is a Likert Scale which we can use to measure the following attributes:
    a. Readability
    b. Coherence
    c. We introduce a new parameter called Empathetic Appropriateness/Empathy Score.
4. Basically, this works better if there are at least two annotators is what I think.

Note:
I am okay with doing all the error analysis and also majorly drafting the submission file if you guys can take care of the rest :)
I will start writing the Background/Related work, abstract, introduction, and References sections once the model is finalized. Will take time to get references and if you guys read the final submission template, it mentions that we need to cite at least 10 papers in related work, and explain at least 2-3 paper approaches, so that will also take a lot of time.
I will also start on the Data Collection and preprocessing section if Subramanya can give me the statistics. (That day we discussed remember?)