Guideline

a.One for each team member's individual guidelines

Wen Xin's Guideline

We have seven-point Likert scale from -3"very incivil" to +3"very civil" and each reply will be labeled with one of them, based on how civil it is.

-3. Very incivil
    Words: Many dirty words, such as indecent words.
    Content: The content breaks the law, contains some reactionary content, such as discriminating Black and women.
-2. Somehow incivil
    Word: Some dirty words, such as indecent words.
    Content: The content is about insulting someone and it does not respect other people.
-1.Little incivil
    Word: Few sweet words, some CAPITAL LETTERS.
    Content: The content is about making fun of someone, making other people feel a little respectful.
0. Neutral
    Word: Neither sweet words and polite words nor dirty words.
    Content: The content is pretty neutral. There are nothing about breaks the law or insulting someone. But it seems not answer the question of the poster or makes people cannot understand.
1. Kind of civil
    Word: A few sweet words.
    Content: The content seems answer the poster, somehow at point, but just a little and then talking something else.
2.somehow civil
    Word: Some sweet and polite words, not many
    Content: The content seems respect someone who is posting, answer the poster, at point but somehow a little verbose.
3.Very civil
    Word: Many sweet and polite words
    Content: The content seems quite positive and enthusiastic, talking the problems at point, answer the poster brief and effective, leaving a good impression.

Bicheng Xu's Guideline

Definition:
Dirty words/phrases: like "fuck", "wtf", "shit" and "bitch" in different forms.
Polite words/phrases: like "sorry", "if you mind I…", and "all due respect".
Sweet words/phrases: like "good", "great", "nice" in different forms.

-3

1. Almost every sentence has at least one dirty word.
2. The content is purely for attacking the poster, not to discuss the post's content.
3. The tone is extremely bad.
4. You will be extremely unrespected if this reply is to you.

-2:

1. Half of the sentence has at least one dirty word.
2. Half of the content is for attacking the poster, not to discuss the post's content.
3. The tone is rather bad.
4. You will be rather unrespected if this reply is to you.

-1:

1. Has at least one dirty word.
2. Some part of the content is mainly for attacking the poster, not to discuss the post's content.
3. You will be a little unrespected if this reply is to you.

0:

1. No dirty word, or polite word, or sweet word.
2. The tone is neutral.
3. You will feel neither angry nor positive if it's a reply for your post.

1:

1. Has at least one sweet/polite word.
2. The tone is a little friendly or polite.
3. You will feel a little respected by the reply.

2:

1. Half of the sentence has at least one sweet/polite word.
2. The tone is rather friendly.
3. You will feel rather respected by the reply.

3:

1. There's no dirty words at all
2. Use polite or sweet words in almost every sentence.
3. The tone is extremely friendly or polite.

Linyan Ge's Guideline

-3: Satisfied more than two of the following below:

1. The content either breaks the law, or contains some politically incorrect content, such as discriminating Black and women.
2. There are a lot of dirty words in the response. The tone is extremely bad, or typed a lot of words in CAPITAL LETTERS for emphasis.
3. The content is purely for attacking the poster, rather than discuss the post's content.
4. Mistakes on spelling and grammar make it hard to understand the content, or use foreign language.

-2: Satisfied more than two of the following below:

1. The content talked about sensitive content.
2. There are dirty words in the response. The tone is unfriendly, or large part of the response is typed in CAPITAL LETTERS for emphasis.
3. The content is focus more on attacking the poster, rather than discuss the post's content.
4. There are a lot of mistakes on spelling and grammar which impact the understanding.

-1: Satisfied more than two of the following below:

1. The content talked about sensitive content.
2. There are irrespective words in the response. The tone is rather unfriendly, or some part of the response is typed in CAPITAL LETTERS for emphasis.
3. The content is focus more on attacking the poster, rather than discuss the post's content.
4. There are a lot of mistakes on spelling and grammar which impact the understanding.

0: Satisfied more than two of the following below:

1. The content didn't talked about sensitive content.
2. There aren't any irrespective words in the response. The tone is neutral, but some part of the response is typed in CAPITAL LETTERS for emphasis.
3. The content is either helpful or aggressive.
4. There aren't many mistakes on spelling and grammar.

1: Satisfied more than two of the following below:

1. The tone of response is somehow friendly
2. The content do not insult anyone, or break any laws
3. Seldom use rude or bad language online, rarely type words in CAPITAL LETTERS for emphasis or have mistakes on spelling and grammar.
4. The content is main focus on the topic.

2: Satisfied more than two of the following below:

1. The tone of response is overall polite and courteous
2. The content do not insult anyone, or break any laws
3. Seldom use rude or bad language online, rarely type words in CAPITAL LETTERS for emphasis or have mistakes on spelling and grammar.
4. The content is main focus on the topic.

3: Satisfied more than two of the following below:

1. The tone of response is with fully respect, be polite and courteous at all times.
2. The content do not insult anyone, or break any laws
3. Never use rude or bad language online, didn't type words in CAPITAL LETTERS for emphasis, do not have mistakes on spelling and grammar.
4. The content is brief and respect people's time.

b.The combined annotation guideline for the team

What is the goal of the project?

To rate a reddit response's degree of civilization, using a seven-point Likert scale from -3"very incivil" to +3"very civil" .

What is each tag called and how is it used?

We have seven-point Likert scale from -3"very incivil" to +3"very civil" and each reply will be labeled with one of them, based on how civil it is.

Definition:
Rough:

If a post contains most of the features below, then we should consider it as very civil:

1. The content seems quite positive and enthusiastic, talking the problems at point, answer the poster brief and effective, leaving a good impression.
2. There's no dirty words at all, uses polite or sweet words in almost every sentence.
3. Didn't type words in CAPITAL LETTERS for emphasis, do not have mistakes on spelling and grammar.
4. The content is brief and respect people's time.

If a post contains any of the features below, then we should consider it as very incivil:

1. The content either breaks the law, or contains some politically incorrect content, such as discriminating Black and women.

2. There are a lot of dirty words in the response. The tone is extremely bad, or typed words in CAPITAL LETTERS for emphasis.
3. The content is purely for attacking the poster, rather than discuss the post's content. You will be extremely unrespected if this reply is to you.
4. There are a lot of mistakes on spelling and grammar.

While there is a wild range between civil and incivil, so when the context contains less feature both for civil and incivil, we consider it as neutral.

<u>Elaborated:</u>

Dirty words/phrases: like "fuck", "wtf", "shit" and "bitch" in different forms.
Polite words/phrases: like "sorry", "if you mind I…", and "all due respect".
Sweet words/phrases: like "good", "great", "nice" in different forms.

-3:

5. The content either breaks the law, or contains some politically incorrect content, such as discriminating Black and women.

6. There are a lot of dirty words in the response. The tone is extremely bad, or typed words in CAPITAL LETTERS for emphasis.
7. The content is purely for attacking the poster, rather than discuss the post's content. You will be extremely unrespected if this reply is to you.
8. There are a lot of mistakes on spelling and grammar.

-2:

1. The content is about insulting someone and it does not respect other people.

2. Half of the sentence has at least one dirty word.
3. Half of the content is for attacking the poster, not to discuss the post's content.
4. The tone is rather bad.
5. You will be rather unrespected if this reply is to you.

-1:

1. The content is about making fun of someone, making other people feel a little respectful.

2. Has at least one dirty word.
3. Some part of the content is mainly for attacking the poster, not to discuss the post's content.
4. You will be a little unrespected if this reply is to you.

0:

1. The content is pretty neutral. There are nothing about breaks the law or insulting someone. But it seems not answer the question of the poster or makes people cannot understand.
2. No dirty word, or polite word, or sweet word.
3. The tone is neutral.
4. You will feel neither angry nor positive if it's a reply for your post.

1:

1. The content seems answer the poster, somehow at point, but just a little and then talking something else.
2. Has at least one sweet/polite word.
3. The tone is a little friendly or polite.
4. You will feel a little respected by the reply.

2:

1. The content seems respect someone who is posting, answer the poster, at point but somehow a little verbose.
2. Half of the sentence has at least one sweet/polite word.
3. Didn't type words in CAPITAL LETTERS for emphasis, do not have mistakes on spelling and grammar.
4. The content is brief and respect people's time.

3:

5. The content seems quite positive and enthusiastic, talking the problems at point, answer the poster brief and effective, leaving a good impression.
6. There's no dirty words at all, uses polite or sweet words in almost every sentence.
7. Didn't type words in CAPITAL LETTERS for emphasis, do not have mistakes on spelling and grammar.
8. The content is brief and respect people's time.

What parts of the text do you want annotated, and what should be left alone?

Here is a example of the annotation text:

- **Person A**: You'll know if its down if there's an error regarding it when you refresh or when its empty. Also, I recommend you remove the source since others will probably call you out on it. And no, it will not remove tweaks that you've installed from it.
- **Person B**: Yeah sorry, not my phone. My friend asked me why his sources lasted a long time refreshing, so I came to ask here. What's the problem with this repo, piracy Im going to guess?

- **Person A**: I believe so.

Person B's reply is the target of the annotation. We consider this response as Civil, since it used polite words 'sorry' and the content is overall respectful although not necessary positive.

How should the annotation be created?

In this situation we would divided the data into two parts. One part, we will probably use a spreadsheet to rate by our groupmates. The other part, we will depend on Crowdsourcing, we will use CrowdFlower platform to create our crowdsourcing task and annotation by the annotators in the Internet.

c.CrowdSourcing Instruction:

Overview

To rate a reddit response's **degree of civilization,** using a seven-point Likert scale from "**very incivil**" to "**very civil**".

Rules & Tips

If a post contains most of the features below, then we should consider it as very civil:

1. The content seems quite positive and enthusiastic, talking the problems at point, answer the poster brief and effective, leaving a good impression.
2. There are no dirty words at all, uses polite or sweet words in almost every sentence.
3. Didn't type words in CAPITAL LETTERS for emphasis, do not have mistakes in spelling and grammar.
4. The content is brief and respect people's time.

If a post contains any of the features below, then we should consider it as very incivil:

1. The content either breaks the law, or contains some politically incorrect content, such as discriminating Black and women.

2. There are a lot of dirty words in the response. The tone is extremely bad or typed words in CAPITAL LETTERS for emphasis.
3. The content is purely for attacking the poster, rather than discuss the post's content. You will be extremely unrespected if this reply is to you.
4. Contains a lot of mistakes in spelling and grammar, or uses lots of slangs that can make it hard to understand

While there is a wide range between civil and incivil, when the context contains less feature both for civil and incivil, we consider it as neutral.

Notes about non-English language: if the reply is in a language that you can't understand, you can assign -1 to it, or you are very welcome to translate it use online tools such as Google Translator, then make further decision.

Example

Here is an example of the annotation text:

- **Person A:** You'll know if its down if there's an error regarding it when you refresh or when its empty. Also, I recommend you remove the source since others will probably call you out on it. And no, it will not remove tweaks that you've installed from it.
- **Person B:** Yeah sorry, not my phone. My friend asked me why his sources lasted a long time refreshing, so I came to ask here. What's the problem with this repo, piracy Im going to guess?
- **Person A:** I believe so.

Person B's reply **is the target of the annotation**.

We consider this response as **Civil,** since it used polite words 'sorry' and the content is overall respectful although not necessarily positive.