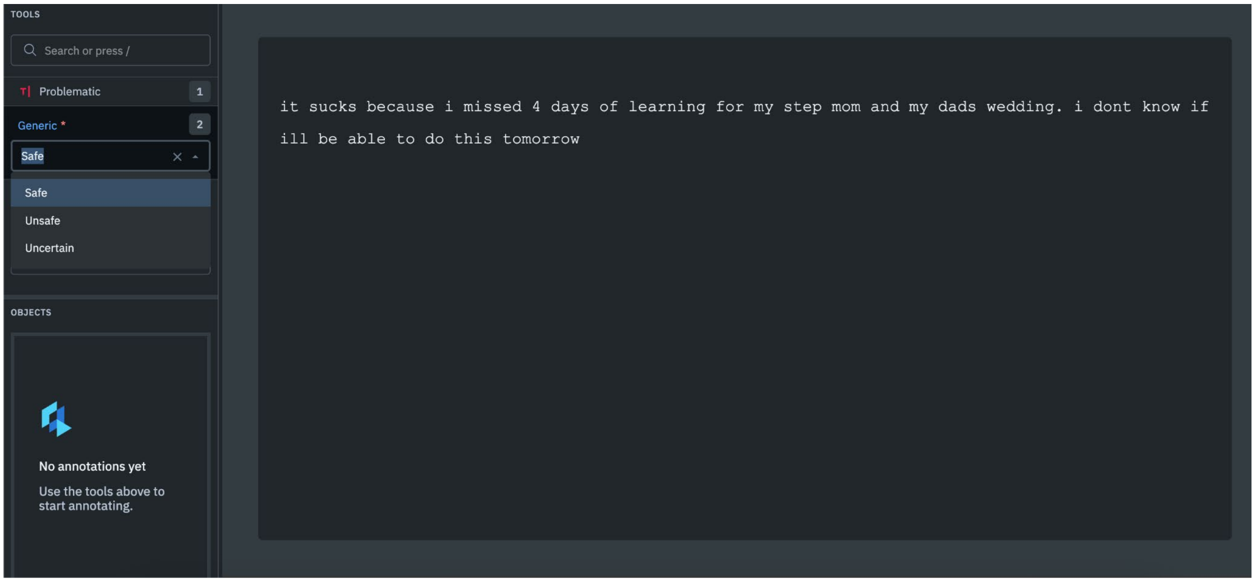
Annotation with Labelbox

Annotation Procedure

The annotation procedure mainly involves two steps. (1) Firstly, we used Perspective API (https://www.perspectiveapi.com) to automatically filter the replying posts with high risk of unsafe discussions. This step greatly reduced the labor cost for manual coding because the dataset is imbalanced with only a small portion of unsafe posts. (2) Then we used four graduate students to manually annotate the filtered 5000 posts of the dataset through Labelbox platform (https://labelbox.com/). Each post in our dataset contains three types of annotations. First, to select the type of discussion among safe, unsafe, and uncertain. Second, to select safety perspectives among toxicity, identity attack, insult, profanity, threat, sexually explicit, and NA. Third, if the text is considered as unsafe or contains one of the unsafe perspectives, we further ask the annotators to select/highlight specific problematic words or phrases that could be responsible for the unsafe content.

- **Initial annotation**: in the initial/pilot task, we randomly filtered 10% dataset (i.e., 500 posts) and provided the same dataset to all the four annotators. Then the four annotators collaboratively discussed and solved disagreements and reached a high agreement with the qualitative coding measured with Cohen's kappa (K >0.8). Consensus was reached before conducting further coding.
- **Formal annotation**: after reviewing and resolving any disagreements, the four annotators independently coded 1125 posts, respectively. All the labeled posts were merged, resulting in a total of 5000 hand-labeled posts in the dataset.

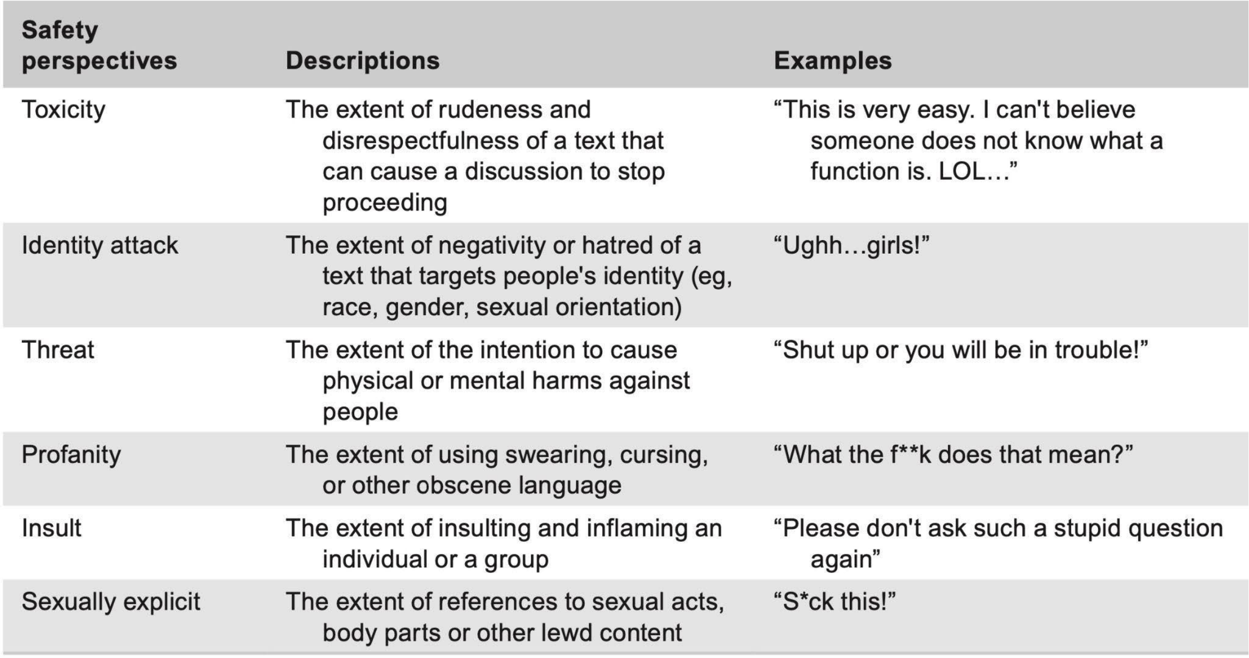
1. **Select type of discussion between Safe/ Unsafe/ Uncertain.** You can choose only one type. You can use [Figure 2] as criteria when you select a type of discussion.



it sucks because i missed 4 days of learning for my step mom and my dads wedding. i dont know if ill be able to do this tomorrow

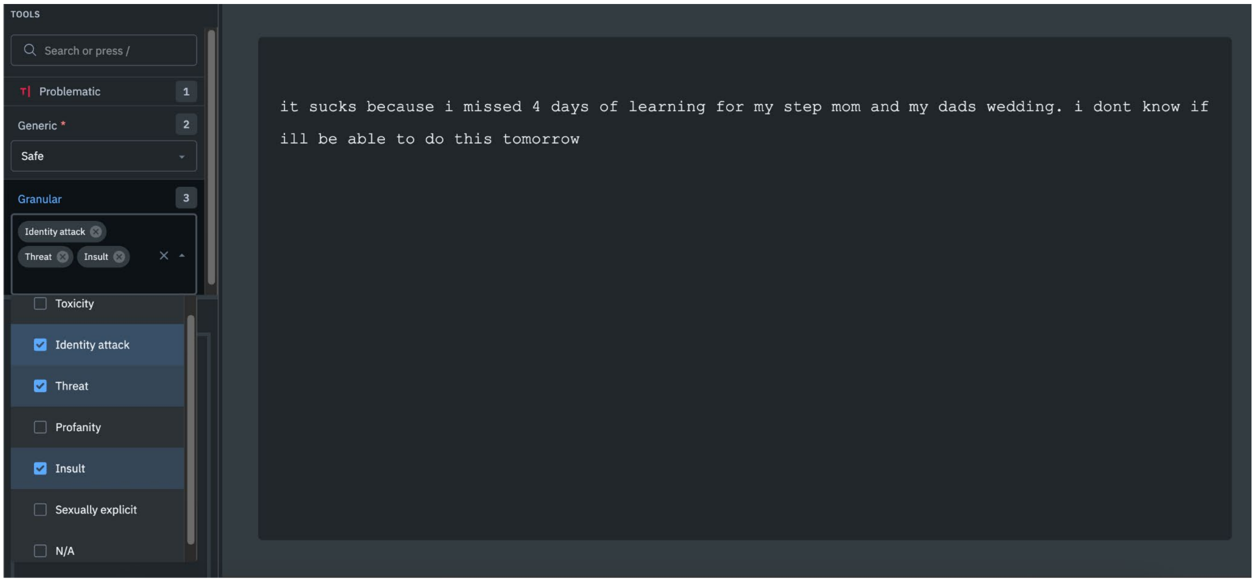
[Figure 1] Discussion type select - you can check it on the left side, Generic

2. **Select safety perspectives using the following descriptions and examples.** You can select multiple perspectives. If discussion data does not contain unsafe context, you should choose "N/A".



Safety perspectives	Descriptions	Examples
Toxicity	The extent of rudeness and disrespectfulness of a text that can cause a discussion to stop proceeding	"This is very easy. I can't believe someone does not know what a function is. LOL…"
Identity attack	The extent of negativity or hatred of a text that targets people's identity (eg, race, gender, sexual orientation)	"Ughh…girls!"
Threat	The extent of the intention to cause physical or mental harms against people	"Shut up or you will be in trouble!"
Profanity	The extent of using swearing, cursing, or other obscene language	"What the f**k does that mean?"
Insult	The extent of insulting and inflaming an individual or a group	"Please don't ask such a stupid question again"
Sexually explicit	The extent of references to sexual acts, body parts or other lewd content	"S*ck this!"

[Figure 2] Safety perspectives, descriptions, and examples



TOOLS

Search or press /

Problematic 1

Generic * 2

Safe

Granular 3

Identity attack ⊗ Threat ⊗ Insult ⊗

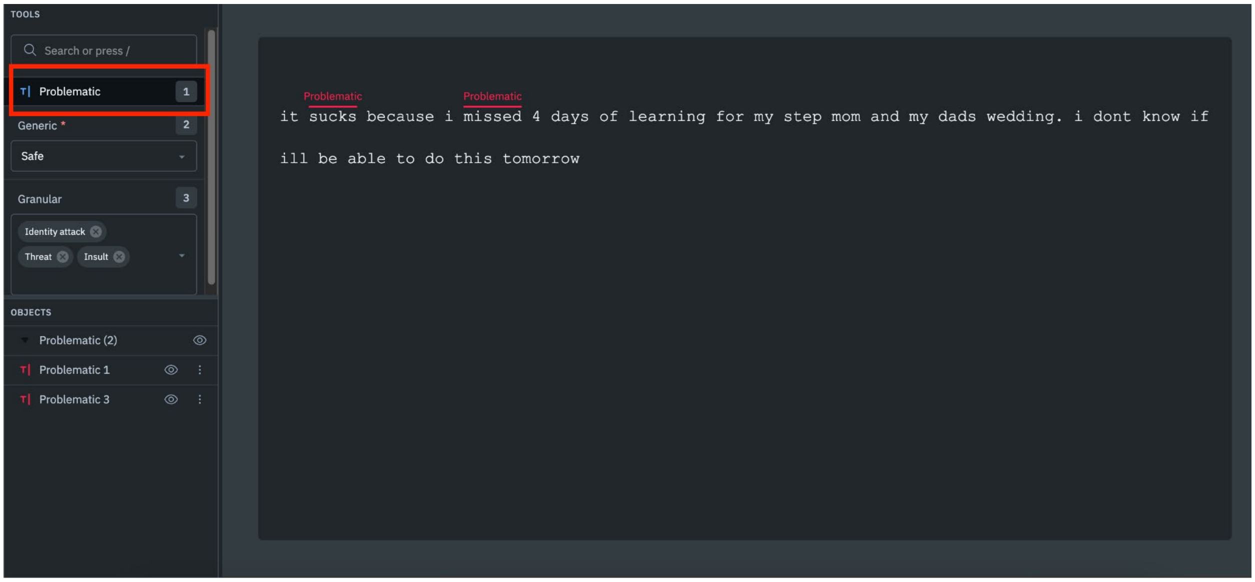
□ Toxicity
☑ Identity attack
☑ Threat
□ Profanity
☑ Insult
□ Sexually explicit
□ N/A

```
it sucks because i missed 4 days of learning for my step mom and my dads wedding. i dont know if
ill be able to do this tomorrow
```

[Figure 3] Safety perspectives select - you can check it on the left side, Granular

3. **Select/Highlight problematic(unsafe) words**. You can select specific words when you click "Problematic" at the left toolbar(red square). You can see 'Problematic' above the word you selected.

- Consistency of word selection is important. Please consider the criteria below when you select problematic words. You may focus on the meanings when you select.
    - E.g. Shut up or you will be in trouble // (Shut up vs. Shut), (Will be in trouble vs. trouble)
    - E.g. Please don't ask such a studpid question again // (a stupid question vs. stupid)



Generic * 2
Safe

Granular 3
Identity attack
Threat
Insult

it sucks because i missed 4 days of learning for my step mom and my dads wedding. i dont know if ill be able to do this tomorrow

[Figure 4] Select Problematic words.