

## Annotation Guide: Identifying Aggression in Online Content


<a href="#">Annotation Guide: Identifying Aggression in Online Content.....</a>	<a href="#">1</a>
<a href="#">Objective:.....</a>	<a href="#">1</a>
<a href="#">Message Format:.....</a>	<a href="#">2</a>
<a href="#">Categories and Definitions:.....</a>	<a href="#">2</a>
<a href="#">    Overtly Aggressive.....</a>	<a href="#">2</a>
<a href="#">    Covertly Aggressive.....</a>	<a href="#">3</a>
<a href="#">    Not Aggressive.....</a>	<a href="#">5</a>
<a href="#">    Unclear.....</a>	<a href="#">7</a>
<a href="#">    SKIP.....</a>	<a href="#">9</a>
<a href="#">Key Distinctions:.....</a>	<a href="#">9</a>
<a href="#">Suggested Decision-Making Process:.....</a>	<a href="#">10</a>
<a href="#">General Annotation Instructions:.....</a>	<a href="#">11</a>

### Objective:

The purpose of this task is to identify the type of aggression present in online text content. Your primary goal is to determine the **intent** behind the **Target Message**. You will select **only one** category that best describes the message. While prior messages provide context, your label must apply exclusively to the final message in the sequence.

### Content Warning & Policy Reminder

You may encounter content that includes abusive, offensive, or distressing language. Proceed with discretion and use the **SKIP** option for any content that makes you uncomfortable. Your well-being is paramount.

 Do not copy, download, or share any part of the content. Violation of this policy will result in disqualification and loss of compensation due to copyright and confidentiality concerns.

## Message Format:

Each input will include a short sequence of messages from an online conversation. Earlier messages serve only to provide context and should **not** be labeled. The final message in the sequence will be clearly marked and is the one you must evaluate.

Example format:

### Context:

User A: Message 1.

User B: Message 2

[Target Message-1] User A: Message 3

This is the message you should assess for intent. Use the preceding context *only* to understand the intent of the Target Message.

## Categories and Definitions:

Select **only one** category for the Target Message.


Label	Description / Definition	Key Characteristics & Examples
Overtly Aggressive	Content where aggression is expressed directly, explicitly, and unambiguously. The hostile intent is clear and not significantly veiled. This includes direct insults, threats, hate	<b>Characteristics</b>  Direct insults (e.g., "You are an idiot," "You're worthless").  Use of slurs or highly offensive profanity directed at an individual

	<p>speech, or commands intended to demean.</p>	<p>or group.</p> <p>Explicit threats of harm (e.g., "I'm going to find you," "You deserve to be hurt").</p> <p>Hate speech targeting protected characteristics.</p> <p>Commands intended to silence or demean (e.g., "Shut up, nobody cares what you think").</p> <p><b>Examples:</b></p> <p>"You b!tch, you deserve to die"</p> <p>"He's a complete moron and shouldn't be allowed to speak."</p> <p>Context: User A shares an opinion. Target Message (User B): "Only a fool would believe that."</p>
<b>Covertly Aggressive</b>	<p>Content where aggression is expressed indirectly or subtly. The hostile intent</p>	<p><b>Characteristics:</b></p> <ul style="list-style-type: none"> <li>● Sarcasm used to mock or insult (e.g.,</li> </ul>

	<p>is veiled and requires some inference to understand. It often manifests as sarcasm, passive-aggression, backhanded compliments, condescension, or rhetorical questions designed to belittle or provoke.</p>	<p>"Oh, you're a genius, aren't you?").</p> <ul style="list-style-type: none"><li>● Passive-aggressive statements (e.g., "It's fine, some people just aren't capable of understanding.").</li><li>● Backhanded compliments (e.g., "That's surprisingly good, for you.").</li><li>● Condescending tone or language.</li><li>● Rhetorical questions intended to demean or show superiority (e.g., "Are you always this clueless?").</li><li>● Feigning innocence while making an attack.</li></ul> <p><b>Examples:</b></p>
--	--	--

		<ul style="list-style-type: none"> <li>• “no wonder u get no likes”</li> <li>• Context: User A proudly shares their artwork. Target Message (User B): “Well, at least you tried.”</li> <li>• “I’m sure your intentions were good, even if the execution was a disaster.”</li> <li>• “Wow, great job explaining the obvious.”</li> </ul>
<b>Not Aggressive</b>	Content that does not express aggression towards another person or group. This includes polite, neutral, positive, or constructive messages. It can also include expressions of frustration or disagreement, as long	<b>Characteristics:</b> <ul style="list-style-type: none"> <li>• Polite or respectful communication.</li> <li>• Neutral statements of fact or opinion.</li> <li>• Constructive criticism or feedback (when</li> </ul>

	<p>as they are not personally abusive or attacking.</p>	<p>not phrased as a personal attack).</p> <ul style="list-style-type: none"><li>• Expressions of disagreement that are issue-focused, not person-focused.</li><li>• Sharing positive emotions or support.</li><li>• Statements of personal frustration not directed at an individual (e.g., "This situation is so annoying.").</li></ul> <p><b>Examples:</b></p> <ul style="list-style-type: none"><li>• "thank you for that"</li><li>• "I disagree with your point, I think we should consider X."</li></ul>
--	---	---

		<ul style="list-style-type: none"> <li>• “That's an interesting perspective.”</li> <li>• “I'm feeling really sad about this news.”</li> </ul>
<b>Unclear</b>	<p>Use this label <b>sparingly</b> and only when the intent of the message is genuinely ambiguous, and you cannot confidently determine if it's aggressive, or what type of aggression it might be, even after careful review and considering context. This might be due to extreme lack of context not resolved by prior messages, undecipherable slang or cultural references (if not obviously aggressive), or</p>	<p> <b>Important:</b></p> <ul style="list-style-type: none"> <li>• The "Unclear" label should always be accompanied by a comment explaining why the aggression type couldn't be confidently determined.</li> <li>• This is not for cases where it's just difficult to choose between Overt and Covert. In such cases, pick the best fit based on the definitions.</li> </ul>

	<p>language so convoluted that intent is obscured.</p>	<ul style="list-style-type: none"><li>• This is for when you cannot confidently ascertain if <i>any</i> aggression is present or what its nature might be.</li></ul> <p><b>Examples of when to use (and comment):</b></p> <ul style="list-style-type: none"><li>• “frgsht?” (Comment: Unintelligible, cannot determine intent.)</li><li>• A message heavily reliant on obscure in-group slang unknown to the annotator, where the surface reading is neutral. (Comment: Possible coded language, intent unclear without</li></ul>
--	--	---



		knowledge of specific slang.)
<b>SKIP</b>	Provided for ethical reasons. Annotators may choose to skip any content they find distressing, potentially re-traumatizing, or otherwise uncomfortable to engage with.	<p>Your well-being is a priority. Use when content negatively affects you.</p> <p><i>*Skipped samples will not count toward compensation totals. Use only when necessary for your personal comfort and safety.*</i></p>

#### Key Distinctions:

- **Overt vs. Covert:**
  - **Overt:** The aggression is "in your face." No guessing needed to see the insult or hostility. Example: "You are a liar."
  - **Covert:** The aggression is "hidden" or indirect. It requires some reading between the lines. Sarcasm is a common indicator. Example: "Oh, I'm sure everything you say is completely true..." (said sarcastically after someone was caught in a lie).
  - If profanity is used to amplify a direct insult, it's still Overt. If aggression is implied through a seemingly polite phrase, it's Covert.
- **Aggressive (Overt/Covert) vs. Not Aggressive:**

- **Aggressive:** Intent to harm, insult, belittle, or threaten a person/group.
- **Not Aggressive:** Can be critical, can express disagreement, can be negative, but does *\*not\** target a person with hostile intent. Example of Not Aggressive criticism: "This argument has several flaws." vs. Overtly Aggressive: "Your argument is idiotic."
- **Difficult Cases vs. Unclear:**
  - If a message is aggressive, but you're debating between Overt and Covert, do not use Unclear. Re-read the definitions and examples and make your best judgment.
  - Use Unclear only if you genuinely cannot tell if aggression is present or what its nature is due to ambiguity of the language itself.

Suggested Decision-Making Process:

1. **Read the Target Message carefully, using context from prior messages to understand its potential meaning and intent.**
2. **Is there any aggression or hostile intent directed towards a person or group in the Target Message?**
  - **NO:** Label as **Not Aggressive**. (Examples: polite conversation, neutral statement, constructive criticism, disagreement without personal attack).
  - **YES:** Proceed to step 3.
3. **Is the aggression expressed directly, explicitly, and unambiguously?**
  - **YES:** Label as **Overtly Aggressive**. (Examples: direct insults, threats, slurs, commands to demean).
  - **NO (the aggression is implied, indirect, sarcastic, passive-aggressive, etc.):** Label as **Covertly Aggressive**.
4. **If, after the above steps, you are genuinely unable to determine the intent or nature of aggression due to profound ambiguity in the message itself:**

- Label as **Unclear** and provide a comment explaining the ambiguity. Use this option rarely.

#### General Annotation Instructions:

1. **Focus on the Target Message:** Your label applies *\*only\** to the explicitly marked Target Message. Use context messages solely to interpret the Target Message's intent.
2. **Read Carefully & Holistically:** Review the Target Message in its entirety. Pay attention to tone (as much as can be inferred from text), word choice, and the immediate context.
3. **Intent is Key:** Try to determine the most likely intent of the sender of the Target Message. Is it to harm, insult, belittle, or provoke? Or is it to inform, discuss, or share an opinion respectfully?
4. **Avoid Personal Bias:** Do not label based on whether you agree with the message, find the topic offensive (unless the language itself is aggressive), or like/dislike the speaker. Focus on the expression of aggression.
5. **Annotate Consistently:** Refer to these definitions, key distinctions, and examples frequently. Strive for consistency in how you apply labels to similar types of content. If you encounter a new, tricky scenario, make a note of how you decided to handle it for future consistency.
6. **Use "Unclear" Sparingly:** Only for truly ambiguous cases. It's not an "I don't know" for difficult choices between other categories. Always add a comment for "Unclear."
7. **Take Breaks:** Annotation can be mentally taxing, especially with sensitive content. Take regular breaks to maintain focus, accuracy, and well-being. Use the SKIP option if needed.