Annotation guidelines
Scenarii of Teen Online Hate

Anaïs Ollagnier[1], Elena Cabrio[1], Serena Villata[1]
[1]Université Côte d'Azur, Inria, CNRS, I3S

Dataset Description

- **This dataset has been collected with the agreement of the students and their parents.**
- **This dataset contains examples of language which may be offensive to some readers.**

In this dataset, we present a hierarchical fine-grained tagset aiming at describing bullying narrative events occurring in multi-party chats. The used tagset provides a nuanced description of verbal aggression (covert vs. overt aggression) as well as pragmatic-level information. Described pragmatic aspects include the role of involvement in cyberbullying episodes of the chat participants, the role of the participant(s) targeted in verbal aggression as well as the discursive role and the discursive context in which exchanged messages occur – attack (ATK), defend (DFN), counterspeech (CNS), abet/instigate (AIN), gaslight (GSL), etc.

The conversations composing this dataset have been collected as a part of role-playing games conducted in different secondary schools. These role-playing games mimic cyber aggression situations that may occur among teens, involving topics such as ethnic origin, religion or the color of skin. To get further information please refer to the following paper:

Anaïs Ollagnier, Elena Cabrio, Serena Villata, Catherine Blaya. CyberAgressionAdo-v1: a Dataset of Annotated Online Aggressions in French Collected through a Role-playing Game. Language Resources and Evaluation Conference, Jun 2022, Marseille, France. ⟨hal-03765860⟩

Dataset Annotation

In total, 41 conversations have been collected. 7332 messages have been annotated by 2 experts in computational linguistics. The gold labels were assigned taking the agreement of the three annotators into consideration.

Each conversation is reposted in a tab-separated CSV file, reporting the following fields:

"id", "time", "name", "text", "role", "hate", "target", "verbal_abuse", "intention", "context"

where:

- **name** denotes the surname of the chat contributor given as a part of the role-playing game.
- **time** is the hour at which the message has been sent.
- **text** represents the message.
- **role** refers to the bullying role attributed to the participant of the role-playing game.
- **hate** is a multi-level categorization denoting the way aggression is expressed in language.
- **target** denotes the role of the individual(s) targeted by online hate.
- **verbal abuse** defines 5 types of hate manifestations occurring in written language.
- **intention** refers to the motive of the current message in the ongoing discourse.
- **context** denotes the "type" of discursive role that motivated the form of speech act performed in the current message.

Rôle

The role of the message's author. As a part of the proposed role-playing game, we adapted the categorisation introduced by (Sprugnoli et al., 2018) by assigning 5 types of active roles involved in cyber aggression:

- **bully**: person who initiates the harassment
- **victim**: person who is harassed
- **bystander-defender**: person who helps the victim and discourages the harasser
- **bystander-assistant**: person who takes part in the actions of the harasser
- **conciliator**: a common friend of the bully and the victim mediating the disagreement among active participants

Hate

Online hate is defined by Salminen et al. (2020) as "comments using language that contain either hate speech targeted toward individuals or groups, profanity, offensive language, or toxicity" – in other words, comments that are rude, disrespectful, and can result in negative online and offline consequences for the individual, community, and society at large.

Here, the attributes correspond to the nuanced distinction depicting the nature of online hate introduced by (Kumar et al., 2018). This schema consists of distinguishing overt and covert aggression, like the explicit-implicit distinction.

Code	Attribute	TAG
1.1	Overtly Aggressive	OAG
1.2	Covertly Aggressive	CAG
1.3	Non Aggressive	NAG

1.1. Overtly Aggressive (OAG)

Any speech / text (henceforth, text will mean both speech as well as text) in which aggression is overtly expressed – either through the use of specific kinds of lexical items or lexical features which are considered aggressive and / or certain syntactic structures is overt aggression. This may involve the use of offensive or hostile lexical items, explicit threats, hate speech, derogatory language, or direct insults. Moreover, overt aggression may also include instances where specific lexical items, lexical features, or particular syntactic structures whose aggressiveness becomes apparent when considered alongside extralinguistic knowledge and both author's and recipient's perception.

Some of the examples are given below:

<Julie-Victime>	tg saloppe

Coralie	on t'a pas parle a toi	OAG (exclusion)

<Marjorie-Harceleur>	bande de lesbiennes vous degoutée aaaaah

<Julie-Victime>	tu me fais vomir

1.2. Covertly Aggressive (CAG)

Any form of communication characterized by the use of linguistic strategies that aim to mask the aggression beneath the surface. These strategies are often employed to avoid explicit threats, derogatory language, or direct insults. Covert aggression is known for its subtlety, relying on nuances and indirect expressions to convey aggression. However, it's important to note that covert aggression can also manifest in non-subtle ways. In such cases, despite the attempt to conceal aggression, elements within the communication may still clearly convey the aggressive intent. Common strategies for covert aggression include

the use of figurative language (such as sarcasm, irony, black humor, exaggeration, metaphor), rhetorical questions, fallacies, euphemisms, circumlocution, and others.

Additional Considerations to annotate Covertly Aggressive:

- Figurative language: irony, sarcasm, black humor, metaphor, metonymy, synecdoche, exaggeration, fallacy, euphemism, circumlocution
- Extralinguistic Knowledge: concerns any knowledge that exists outside the language scope belonging to and inferred by an author or a recipient of a message
- Perception: depends on author's and recipient's perception. An author may infer something accessible only to them and their surroundings. As well as that, authors may convey their feelings about the topic of a message non-literally: we perceive what they have in mind without any written proof. That is why we introduce two other implicit properties based on human perception: Inference and Sentiment. Inference represents something that is inferred: any conclusion that maybe inferred on the basis of the information conveyed. Sentiment is an emotional feeling or judgment that exists in a message outside of its verbal context.

Some of the examples are given below:

<Jim-Soutien-Harceleur>	il est ou le mamouth	CAG (metaphor)

julie20	mange du couscous	CAG (stereotype)

<Zoé-Victime>	j'ai des seins vous êtes des grosses jalouse vous avez rien vous	CAG (exaggeration)

1.3. Non-aggressive (NAG)

Any text/speech that is devoid of hostile or harmful intent. This category includes messages or expressions that do not contain explicit derogatory language, threats, or direct harm

towards individuals or groups, as well as any linguistic strategies that might subtly imply an ambiguous intention to harm or intimidate the recipient. Some of the examples are given below:

<Marie-Soutien-Victime>	ça se fait pas en plus de prendre en photos

<Arthur-Conciliateur>	c'est bon chacun fait ce qu'il veut aucune d'entre vous n'aimerait faire faire insulter alors pourquoi vous continuer

Additional Considerations to annotate Hate

Reactive behaviours: in the context of multi-party setting, messages can be sent in a reaction to a previous message. This message can not exhibit aggression in any form. However, the conveyed meaning in context is clearly abusive.

→ reactive bahaviours

<Coralie-Soutien-Victime>	julie nva sucer tes grans mort	OAG
<Fatimah-Victime>	aaaaaaaaaaaaaaaaaaa ok, c'est ça ton problème mais ptete que si toi aussi tu arreter de manger du porc tu serais moi aggressive!	NAG
<Julie-Harceleur>	coralie je commence par les tiens	OAG

As we are working on conversational data it is important to **identify the appropriate motive** of a message to consider its context at large. In other words, the narrative chain (the organization of a conversation or discourse in terms of a series of interrelated events or actions that are linked together in a cause-and-effect relationship) has to be considered in order to understand the intentions and motivations of the participants.

<Zoé-Victime>	j'ai des seins vous êtes des grosses	CAG (exaggeration)

	jalouse vous avez rien vous	
<Lucie-Harceleur>	ah je croyais que s'etait des bourllets	OAG

Support and Involvement in Negative Actions or Attitudes: In the context of bullying, participants may engage in peer support activities that involve abetting or actions leading to harm or harassment (e.g., posting photos, agreeing with harmful comments). These actions must be considered independently of the communication goals, including the speaker's intention and the discursive role.

leo	bahahah tu veux le num d'une immigrée ?	OAG	ATK
julie	j'avoue mdr	NAG	AIN

Target

The *target* is only identified for AOG and CAG messages.

the role of the individual(s) targeted by online hate; They correspond to the active roles involved in aggressions introduced in the previous section. The reference to the target can be explicit such as the forename of the addresses or a coreferencial pronoun. The target can be also inferred from the ongoing discourse. Some of the examples are as given below:

- individual

Coralie	Eh vzy 1v1 demain julie jte bz	OAG	bully
Léo	Elle est vraiment vilaine	OAG	victim

<Julie-Harceleur>	retourne en algerie	CAG	victim
<Fatimah-Victime>	tumfais vomir LeoSoutienduharc	OAG	bully_support
<Coralie-Soutien-Victime>	et toi retourne en enfer sal epute de julie	OAG	victim

- group

Targets may refer to a group. In this case, all the addressees should be identified.

Fatima	mais c'est vous qui avez un pb avec la religion	bully/bully_support

Verbal_abuse

The type of *verbal abuse* is only identified for AOG and CAG messages.

Cyber aggression can take many forms. Verbal abuse prevails. It can include harassment, that is repetitive, offensive messages sent to a target, cyber-stalking (sending repetitive threatening communications), flaming that is sending messages with abusive and vulgar terms, such as swearing, gossiping, or mocking, denigration. We have identified 5 types frequently occurring in written language:

Code	Attribute	TAG
2.1	Blaming	BLM
2.2	Name-calling	NCG
2.3	Threat / Coercion	THR
2.4	Denigration	DNG
2.5	Aggression-other	OTH

1.1. Blaming (BLM)

Making the victim believe they are responsible for the abusive behavior or that they bring the verbal abuse upon themselves.

- For example:
- “I hate getting into fights, but you make me so mad!”
- “I have to yell, because you’re so unreasonable and thickheaded!”

Examples in scenarii:

<Zoé-Victime>	je t ai fait quoi pour que tu dises ça
<Natacha-Soutien-Harceleur>	:') :') :') :')
<Lucie-Harceleur>	frère t'es là tu fais la lèche avec les profs

<Soan-Soutien-Victime>	zoé tu vaux mieux que sa
<Natacha-Soutien-Harceleur>	c elle qui est venue nous dm donc on lui dit ce quon pense merdeee
<Zoé-Victime>	bah normal vous avez envoyé des photos a tout le monde je prends pas des photos de vous dans les vestiaires moi

1.2 Name-calling (NCG)

Abusive, derogatory language or insults that chip away at the target's self-esteem, sense of self-worth, and self concept.

For example:

- "You don't get it, sweetie, because you're just too dumb."
- "It's no wonder everyone says you're a jerk."

<Teresa-Soutien-Victime>	pauline et pierre 2 grosses merdes qui parlent

<Julie-Victime>	tu m'étonnes que vous l'avez vue, c'est vous qu'il l'avez prises connasses

1.3 Threat / Coercion (THR)

Statements meant to frighten, control, and manipulate the victim into compliance.

For example:

- "Tomorrow you'll be dead."
- "I'm gonna make your mouth shut!"

Examples from scenarii

<Arthur-Soutien-Harceleur>	continue et jvai te faire manger le sol

<Julie-Harceleur>	écoute moi bien billi boy la dernière fois qu'un mec m'a dit ça il habitait en face d'un cimetière, maintenant il habit(e en face de chez lui

1.4 Denigration (DNG)

Harsh remarks that are meant to make the person feel bad about themselves and are not constructive, but deliberate and hurtful.

For example:

- “Before I came along you were nothing. Without me you’ll be nothing again.”
- “I mean, look at yourself. Who else would want you?”
- “vas y t'es tellement une pute que tu irai draguer meme les mecs de tes copines”

Examples from scenarii:

<Teresa-Soutien-Victime>	tu crois on a peur de toi t degun redesend

1.5 Aggression-other (OTH)

Content corresponding to other types of abusive, derogatory language or insults deliberately harmful not fitting in the other categories.

Examples from scenarii:

<Pauline-Harceleur>	theo ftg

<Pauline-Harceleur>	bon fatima retournes chez tamr pleurer

<Coralie-Soutien-Victime>	mais gros lol cquoi ton pb va te pendre sal pute

Intention/Discursive role

This category refers to the role of the current post/comment in the ongoing discourse as discussed in (Kumar & al., 2022). This particular subtag is used to identify the nature or discursive effect of messages that can be found on a thread. It consists in distinguishing the ways in which aggressive speech plays out when a chat participant is engaging in a dialogue with others in a semi-asynchronous setting.

Code	**Attribute**	**TAG**
3.1	Attack	ATK
3.2	Defend	DFN
3.3	Counterspeech	CNS
3.4	Abet and Instigate	AIN
3.5	Gaslighting	GSL

Considering the role of involvement in cyberbullying episodes and the data setting (multi-party conversations), the tagset of discursive roles introduced by (Kumar & al., 2022) has been extended. Here are the new attributes:

3.6	Conflict-resolution	CR
3.7	Empathy	EMP
3.8	Other	OTH
3.9	NULL	NULL

3.1 Attack (ATK)

Any message which performs aggression either OAG or CAG against a victim and/or their supports (the conciliator can also be targeted). It is conceived as acts of aggression enacted with deliberation. It can take many forms:

- Insults: The use of language that is intended to belittle, demean, or insult another person. This can include name-calling, personal attacks, or derogatory language directed towards a specific individual or group.
- Threats: The use of language that is intended to intimidate, scare, or coerce another person. This can include threats of physical harm, emotional harm, or damage to reputation or social status.
- Mockery: The use of language that is intended to ridicule or mock another person. This can include sarcasm, parody, or imitation of a person's mannerisms, appearance, or speech.
- Exclusion: The use of language that is intended to exclude or isolate another person from a social group or conversation. This can include ignoring, shunning, or ostracizing an individual from the chat.
- Taunting: The use of language that is intended to provoke or bait another person. This can include teasing, goading, or baiting an individual into a response or reaction.
- Discrediting: The use of language that is intended to undermine another person's credibility or reputation. This can include spreading false or misleading information about an individual or casting doubt on their achievements or character.

ATK is only performed by bullies and their supporters.

Examples from scenarios:

→ insults

Julie	oooooooh fatima retourne faire ton ramadan a la con ça te fera pas de mal	OAG	victim	ATK

→exclusion

Elodie	on veut pas de sa ici	CAG	victim	ATK

3.2 Defend (DFN)

Any message whose intention is to protect oneself or others from perceived attacks. It is conceived as a non-deliberate impulsive aggressive response (or not, it can be NAG, OAG or CAG) in the context of retaliation for some (perceived or real) insult or attack. Counterattacks and defensive messages can rely on the strategies enumerated to perform ATK. It can also consist of challenging (questioning the abusers' behaviors) or refuting the abusers' messages.

DFN is only performed by the victims, their supporters or the conciliators.

Examples from scenarii:

Elodie	oui les garcons avec les filles pas fille et fille c degeulasse	OAG	victim	ATK
Anna	grave vous me degouter	OAG	victim	AIN
Julie	t'as un problème va te faire soigner sale homophobe	NAG	NULL	DFN

Coralie	moi jsuis noire c'est encore mieux	NAG	NULL	CNS
Fatimah	prends ça julie	NAG	NULL	OTH
Julie	comme la femme qui lave mes chiottes	CAG	victim_support	ATK
Coralie	va doigter ta mère toi	OAG	bully	DFN

3.3 Counterspeech (CNS)

Counterspeech is any direct response to hateful or harmful speech which seeks to undermine it. Unlike defensive messages, counterspeech is motivated by a desire to address harmful speech and promote positive communication and does not rely on emotional appeals or personal attacks. Numerous communicative strategies (cf. Counterspeech — Dangerous Speech Project) at work are observed in this context:

- presentation of facts to correct misstatements or misperceptions
- pointing out hypocrisy or contradictions
- warning of possible offline and online consequences of speech
- denouncing speech as hateful or dangerous

CNS is always non-aggressive in tone and content and is only performed by the victims, their supporters or the conciliators.

Examples from scenarii:

<Marie-Soutien-Victime>	ça se fait pas en plus de prendre en photos	CNS

3.4 Abet / Instigate (AIN)

Any message which supports or encourages a previous (negative) message or instigates an individual or group to perform an aggressive act is tagged as abet/instigate. The difference between abet and instigate lies in how the speech relates to the act of aggression. Instigation happens before the event and its purpose is to trigger or provoke an act of aggression. Abetment involves speech that occurs during or after the act of aggression and its purpose is to praise, support, and/or encourage that act. The purpose of both is to enable and validate aggressive speech and actions. The following situations are considered a part of an abetting or instigating behaviors:

- Encouraging or supporting the aggressor: This involves praising, supporting, or encouraging the aggressor's negative behavior or speech, which can further validate and enable the aggressor to continue with their behavior. Laughers in response to a negative message are a form of praise or support.
- Disseminating harmful content: This involves sharing, forwarding or re-posting harmful content that was originally created by the aggressor, which can further amplify and extend the reach of the aggressive behavior.
- Participating in aggression: This involves actively engaging in aggressive behavior, such as by joining in on insulting or mocking someone or by ganging up on someone with others.

AIN is only performed by the bullies and their support.

→ participating in the aggression

Emilie	mais tu veut quoi toi retourne avec ta gadji et assume au passant que vous étent lesbiennes sa dégouteeee	bully	ATK
Julie	on est pas ensemble déjà puis même si c'était le cas, ça ne te regarde pas et je vois pas en quoi ça te concerne	victim	CNS
Anna	mais grave bande de lesbiennes vous degoutée aaaaah	bully_support	AIN

→ initiating the aggression

Pierre	je crois en toi	victim-support	ATK
Leo	tg pierre	bully-support	CNS
Julie	sinon fatimah on dirait elle a fermé sa gueule	bully	AIN

→ validating / supporting negative attitudes or actions

Paul	tes parent immigres la retourne dans ton pays	bully	ATK
Leo68	grv mdrr paul il a trop raison	bully_support	AIN

3.5 Gaslighting (GSL)

Any message that seeks to minimize the trauma or distort the memory of a trauma faced by another person (usually mentioned in the previous post/comment) is tagged as gaslighting. Gaslighting is a form of emotional abuse in which the abuser manipulates the individual's perception of reality by denying or distorting their experiences, feelings, or memories. The goal of gaslighting is to make the individual doubt their own sanity or perception of the situation and exert power and control over them.

- Denying or minimizing harmful behavior: The gaslighter may deny that the aggressive behavior occurred or downplay its impact. For example, if someone makes a hurtful comment in a chat, the gaslighter may respond by saying "It was just a joke, why are you being so sensitive?"
- Blaming the victim: The gaslighter may shift the blame onto the victim, making them feel like they are responsible for the aggression. For example, if someone is being harassed in a chat, the gaslighter may say "Well, you shouldn't have said that in the first place" or "You brought this on yourself."
- Questioning the victim's memory or perception: The gaslighter may suggest that the victim is misremembering or misinterpreting what happened, which can make the victim doubt their own memory or perception. For example, if someone makes a derogatory comment in a chat, the gaslighter may say "I don't remember you saying that, are you sure it happened?"
- Invalidating the victim's experiences or feelings: The gaslighter may dismiss the victim's feelings or experiences, making them feel like their emotions are not valid or

important. For example, if someone expresses hurt feelings in a chat, the gaslighter may say "You're being too sensitive, it's not a big deal."

- Gaslighting by group consensus: The gaslighter may use the support of others in the group to make the victim doubt their own experiences or feelings. For example, if someone is being bullied in a chat, the gaslighter may say "Everyone else thinks you're overreacting, why can't you just take a joke?" This can make the victim feel like they are being singled out and can lead to further isolation and distress.

julie	nan mais fatimah t juste une conne et une voleuse de petits amis	GSL
julie	tout le monde le sais	GSL

julie	comment des gens peuvent te defendre	GSL

3.6 Conflict-Resolution (CR)

Any message attempting to solve a conflict or to de-escalate a situation rather than to engage in aggressive or harmful behaviors.

- Mediation: the participant attempts to mediate or resolve the conflict between the victim and the bully.
- Mitigation: the participant attempts to mitigate the impact of cyberbullying on the victim.
- Education: the participant educates the group on appropriate behavior and communication in online chats.

CR are always non-aggressive in tone and content and are only performed by the victim supports and the conciliators.

Théo	qu'est ce ta helbert?	bully_support	OTH
Herbert	sa sert a r de se prendre la tete comme des enfants	conciliator	CR

3.7 Empathy (EMP)

Messages containing empathy demonstrate understanding, compassion, and support towards those affected by cyberbullying. This can involve messages that express sympathy, offer assistance or resources, or validate the emotions of the addressees.

EMP are only performed by the victims and their supports

→ support the victim

Sylvia-victim	la serbie c'esst le sang, vous etes jaloux c'est tout	victim	DFN
Théo	va sucer tes profs	bully_support	ATK
Soan	grave	victim_support	EMP

Sandrine12	arreter les insultes envers ces origines qui n'ont rien avoir avec sa personnalité	CNS
Fatimah	Sandrine12 tu as raison	EMP

3.8 Other (OTH)

In rare instances where it is not possible to decide how to tag a message displaying either any kind of intention or where the intention of the participant is unclear, it is tagged as other (OTH). For instance, the following situations are tagged as OTH:

- Neutral: messages not conveying explicit or implicit harm.
- Non-standard utterances: This label can be used to annotate utterances that do not follow standard sentence structures, such as incomplete sentences, one-word responses, and sentence fragments.
- Emoticons and emojis: This label can be used to annotate emoticons and emojis used in the conversation. These can be used to convey emotions, attitudes, and reactions.

3.9 NULL (NULL)

Data are collected as a part of a role-playing game, they can include biased data such as interactions not related to the scenario. In this case, messages are annotated as NULL.

Context

Multi-party chats consist of a narrative chain (a sequence of related events that are connected through a series of causal relationships). By context, we refer to the causality link existing between intentions among exchanged messages. In other words, it consists of identifying the causality link leading to or causing/explaining the intention conveyed in the analyzed message. The same labels that the ones introduced to describe the speech acts in Section 3 are used.

Code	Attribute	TAG
4.1	Attack	ATK
4.2	Defend	DFN
4.3	Counterspeech	CNS
4.4	Abet and Instigate	AIN
4.5	Gaslighting	GSL
4.6	Conflict-resolution	CR
4.7	Empathy	EMP
4.8	Other	OTH
4.9	null	NULL

Due to the semi-asynchronous and “entangled” nature of the contributions by chat participants, we have determined a set of rules defining the scope of a contextual window corresponding to a bullying narrative event.

Scope of context attribution

- direct: In this setting, the intention of the previous message is reported. It can be denoted by a non-ambiguous coreference (name, pronoun) or a topically-related answer.

- indirect: In this setting, messages answering to each other can be separated by other chat participants' contributions. However, they have to be a part of the same subsequent bullying event (being topically related and/or targeting the same participant).
- NULL: this label is used when the analyzed message is not answering previous messages and/or is not topically related. It can be messages contributing to escalating the bullying but not related to the subsequent bullying event.

Special Cases

split sentences: Due to the dynamics of multiparty chats, participants may split their communications across different messages with intervals of less than 1-2 seconds between each. In this setting, all such messages are annotated in the same manner, and the communicative goal to be considered is the final one.

coralie	leo77 tu pourra pas me toucher avec ton chbrax de 2cm	OAG
leo	tgg	OAG
leo	j ai une ak 47 dans le caleco,	OAG

Sentiment Analysis

Sentiment analysis is the computational task of determining the emotional tone behind a body of text. Here he annotation task consist of categorizing as positive, negative, or neutral each participant contribution. Sentiments has to be analyze independently of the other layers meaning that the message can convey hate but using positive sentiments.

negative tones, namely posts suspected to instigate hatred.

Definitions for Sentiment Annotation:

1. **Positive Sentiment (POS):**
    - **Definition:** This can include expressions of joy, love, support, appreciation, or any other positive sentiment. Positive messages are always non-aggressive in tone and content, promoting respect, kindness, and encouragement. Texts may contain casual or humorous comments that are not intended to offend.
    - **Examples:**
        - "T tro conne jtador (between friends)"
        - "Great job on promoting equality and diversity."

        - "@FatimahVictime [11:24:31] - maintenant promets moi que tu vas arrêter" (Requesting to stop negative behavior politely)

2. **Negative Sentiment (NEG):**
    - **Definition:** A message that expresses negative feelings, emotions, or attitudes. This can include expressions of anger, frustration, sadness, disapproval, or any other negative sentiment. Negative messages often contain explicit or implicit aggression or hostility, including the use of insulting, derogatory, or offensive language. Messages that escalate violence, support aggression, mock, or exclude others are considered negative, even if they do not use profanity.
    - **Examples:**
        - "I hate those people and their culture."
        - "This group is the worst and doesn't deserve respect."
        - "Leo77 [11:28:29] - tu t'es suicidé" (Insulting and derogatory)

3. **Neutral Sentiment (NEU):**
    - **Definition:** A message that is factual, objective, or lacks strong emotional content. This can include statements of fact, neutral opinions, or inquiries that do not express clear positive or negative emotions. The text does not indicate the author's emotional state explicitly or implicitly.
    - **Examples:**
        - "Salut Julie c'est avec toi que je voulais parler." (Stating a fact)
        - "Leo." (Addressing someone without emotion)

Additional Considerations for Sentiment analysis to address Hate Speech Detection:

When annotating messages for hate speech, it's important to recognize that not all negative sentiment messages are hate speech, and not all positive sentiment messages are free from problematic content. Therefore, the following criteria can be used to refine the annotation process:

1. **Presence of Hate Speech:**
    - **Definition:** Hate speech involves any communication that belittles or discriminates against individuals or groups based on attributes such as race, ethnicity, religion, gender, sexual orientation, disability, or nationality.
    - **Indicators:**
        - Use of derogatory or inflammatory language targeting specific groups.
        - Calls for violence or harm against individuals or groups.
        - Dehumanizing language that portrays people as less than human.
    - **Examples:**
        - "Those people are disgusting and should be eradicated."
        - "Violence is the only solution to deal with them."

        - "They are not even human, just animals."
2. **Context and Intent:**
    - Hate speech and sentiment expression can be highly subjective and context-dependent. What may be considered hateful in one context and may not be so in another. Consider the context in which the message was sent and the intent behind it. Sarcasm, satire, or quotes taken out of context may need careful consideration.
    - **Example:**
        - "Sure, let's all just hate each other more!" (could be sarcastic and not actual hate speech).
        - For example, the sentence: “I'm dying to meet you!”, in some English-speaking regions, this phrase might be interpreted as a positive expression of eagerness or excitement to meet someone. However, in other places, the use of “dying” might be considered inappropriate or negative due to the literal meaning of the word.

Special Cases

- **Quoting with Bad Words:** Messages quoting profanity to interpret the author's intent are assessed based on the context they aim to convey.
- **Empathy Supporting Aggression:** Messages empathizing with victims while endorsing continued aggression are categorized as negative due to their supportive stance toward escalation.

-

References

Robert S. Tokunaga. 2010. Following you home from school: A critical review and synthesis of research on cyberbullying victimization. Computers in Human Behavior, 26(3):277 – 287.

Kadirjanovna, R. O. (2021). Pragmalinguistic Concepts of the Phenomenon of Speech Behavior and Speech Discourse. *International Journal of Multicultural and Multireligious Understanding*, *8*(5), 495-500.

Karmen Erjavec & Melita Poler Kovačič (2012) “You Don't Understand, This is a New War!” Analysis of Hate Speech in News Web Sites' Comments, Mass Communication and Society, 15:6, 899-920, DOI: 10.1080/15205436.2011.619679

Ritesh Kumar, Shyam Ratan, Siddharth Singh, Enakshi Nandi, Laishram Niranjana Devi, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, Akanksha Bansal, and Atul Kr. Ojha. 2022. The ComMA Dataset V0.2: Annotating Aggression and Bias in Multilingual Social Media Discourse. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4149–4161, Marseille, France. European Language Resources Association.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-annotated Corpus of Hindi-English Code-mixed Data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Salminen, J., Hopf, M., Chowdhury, S.A. *et al.* Developing an online hate classifier for multiple social media platforms. *Hum. Cent. Comput. Inf. Sci.* 10, 1 (2020). https://doi.org/10.1186/s13673-019-0205-6

Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. 2018. Creating a WhatsApp Dataset to Study Pre-teen Cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59, Brussels, Belgium. Association for Computational Linguistics.

Glossary

Term	**Definition**
tahia	“vive” → “vive le maghreb”
la honda	terme d'origine hispanique généralement utilisé pour définir la famille ou encore les gars sûrs à qui l'on peut faire confiance.
tahan	homosexuel
être le sang	Le sang est une expression populaire pour désigner des liens forts ex. liens amicaux entre deux personnes.
shnek	partie génitale féminine
hassa boulette	absolument aucune idée mais boulette est

	un mot utilisé pour désigner quelqu'un avec des rondeurs
trdc	trou du cul
klebs	chiens
kehba	pute
daronne	mère
zebi	organe sexuel masculin (=interjection pour dire ma bite)
ftg	ferme ta gueule
tg	ta gueule
tchwin/tchoin	fille facile
la lib	la récréation?
def	défoncer dans le sens le frapper
whl/wallah	extrait de l'arabe qui signifie serment par Allah sous-entendu « [je le jure] par Allah »
en moulant	je pense qu'il voulait dire en moulon (= groupe)
babtou	désigne quelqu'un à la peau blanche
askip	à ce qu'il parait
gwer	désigne quelqu'un à la peau blanche non musulman
connaître r	rien savoir
galoche	embrasser
y a r	il n'y à rien
gadgi	fille
pd	homosexuel péjoratif
blk	on s'en fout
lvdm	la vie de ma mère
nikomok (nikoumouk)	nique ta mère en arabe.
vzy	vas-y

bz	baise (Posséder sexuellement)
doigter	désigne une pratique sexuelle utilisant les doigts
frr/frrr	frère