CyberAgressionAdo-Large: French Multiparty Chat Dataset to Address Online Hate

Anaïs Ollagnier, Elena Cabrio, Serena Villata, Valerio Basile

CyberAgressionAdo-Large: French Multiparty Chat Dataset to Address Online Hate

Anaïs Ollagnier* — **Elena Cabrio*** — **Serena Villata*** — **Valerio Basile****

* *Université Côte d'Azur, Inria, CNRS, I3S, 930 route des Colles, BP 145, 06903 Sophia Antipolis Cedex, France*
** *Department of Computer Science, University of Turin, Corso Svizzera, 185, 10149 Torino, Piemonte, Italy*

ABSTRACT. *This paper presents an extended version of CyberAgressionAdo, a French open-access dataset for online hate detection in multiparty conversations. The annotation process was improved with refined guidelines and a two-phase inter-annotator agreement study. A new adaptation of the Weirdness Index is introduced to analyze annotator disagreements. Now structured as a perspectivist corpus, with annotations provided by multiple annotators, CyberAgressionAdo-Large constitutes an enriched resource for the computational analysis of online hate situations in French.*

KEYWORDS: *Cyber-aggression, Annotation scheme, Multiparty chats.*

TITRE. *CyberAgressionAdo-Large: Jeux de données français de conversations multipartites pour l'étude de la haine en ligne*

RÉSUMÉ. *Cet article présente une version étendue de CyberAgressionAdo, un jeu de données français en accès libre destiné à la détection de la haine en ligne dans des conversations multipartites. Le processus d'annotation a été amélioré grâce à des directives affinées et à une étude en deux phases de l'accord inter-annotateurs. Une nouvelle adaptation de l'indice de « Weirdness » est présentée afin d'analyser les désaccords entre annotateurs. Désormais structuré comme un corpus perspectiviste, avec des annotations réalisées par plusieurs annotateurs, CyberAgressionAdo-Large constitue une ressource enrichie pour l'analyse computationnelle des situations de haine en ligne en français.*

MOTS-CLÉS : *Cyber-agression, Schéma d'annotation, Conversation multipartites.*

1. Introduction

Social media platforms have become essential components of contemporary communication, fostering freedom of expression and enabling the exchange of diverse ideas. This rapid expansion has also led to a rise in harmful, abusive, and degrading content, exposing individuals across various demographics to unsafe and detrimental interactions, thereby posing significant risks to their mental health and overall well-being. In response, automatic detection of online hate has emerged as a prominent research area in Natural Language Processing (NLP), with extensive studies presented at leading conferences, specialized workshops, and shared tasks (Alkomah and Ma, 2022). The majority of research efforts focus on popular social media platforms such as Twitter and Facebook, with an increasing number of techniques specifically developed for identifying harmful content in a monolingual setting, primarily English. However, recent studies indicate that private messaging platforms and chat rooms are significant environments for cyberbullying, particularly among adolescents (Alhashmi *et al.*, 2023). Due to the private nature of exchanges on these platforms and privacy policies that restrict data collection, few datasets capture aggressive interactions suitable for computational analysis (Cecillon *et al.*, 2020). Recently developed resources that simulate online aggression through role-playing games—where participants take on fictional roles to replicate cyber-aggression situations occurring in multiparty conversational settings—have contributed to addressing this gap (Gamal *et al.*, 2023). Notably, two recent French-language datasets (Ollagnier *et al.*, 2022; Ollagnier, 2024) provide valuable resources for addressing multiple online hate-related detection subtasks in conversational contexts, while also contributing to the exploration of linguistic diversity and cultural factors in non-English languages. Despite their significance, these datasets are relatively small in size (19 conversations, 2,921 messages) and exhibit limited topic diversity, resulting in an uneven representation of sensitive themes. Additionally, since participants in these role-playing scenarios have the freedom to influence the direction of their group's story, expanding the scope of scenarios and increasing data collection are essential to better capture the breadth of bullying practices observed in real-world contexts.

In this paper, we present the *CyberAgressionAdo-Large* dataset [1], an extended version of the dataset introduced in Ollagnier (2024). To our knowledge, *CyberAgressionAdo-Large* is currently the largest publicly available French-language dataset of aggressive conversations. It is distinguished by its size, diversity (covering four sensitive topics), and its in-depth linguistic analysis, featuring six layers of annotation designed to computationally address multiple online hate-related detection sub-tasks. Building on both existing materials (Ollagnier *et al.*, 2022; Ollagnier, 2024) and newly developed resources, this paper consolidates all information regarding the experimental process designed to collect conversations mimicking cyber-aggression in schools, including the experimental setup and role-play scenarios. It offers a compre-

1. The dataset is publicly available at: `https://anonymous.4open.science/r/CyberAgression-Large-C71C/`.

hensive presentation of the annotation tagset, supported by inter-annotator agreement experiments, highlighting its scalability and applicability on a larger scale. Additionally, the paper provides a comprehensive overview of the annotation tagset, supported by inter-annotator agreement experiments, to demonstrate its scalability and applicability on a larger scale. Additionally, we adapt the Weirdness Index (Basile, 2020) to facilitate an in-depth analysis of annotator disagreements. Specifically, polarized Weirdness scores are integrated into a visualization method to investigate potential semantic shifts that could explain divergences in human-provided labels. Furthermore, the released dataset includes annotations from three distinct annotators, supporting computational approaches aligned with the perspectivism trend in machine learning (Cabitza *et al.*, 2023). In summary, our contributions are as follows:

– with 5,789 annotated messages and six layers of annotation, *CyberAgressionAdo-Large* serves as a valuable resource for the French online hate-detection community, supporting diverse research and applications beyond the traditional conceptualization of online hate-related sub-tasks;

– we provide a detailed methodology for collecting naturalistic interactions and a comprehensive description of the annotation tagset to ensure reproducibility;

– we conduct (dis)agreement analysis experiments to ensure reliability and scalability, while offering a methodology to explore divergences in annotators' opinions;

– we apply pattern mining to deepen the understanding of complex cyberbullying communication practices commonly observed in multiparty settings;

– we release data annotated by multiple individuals to support perspectivist computational approaches.

2. Related work

From 2016 onward, a high number of resources and benchmark corpora have been developed to address various tasks related to online hate detection. Most of the research has focused on detecting harmful content such as offensive, toxic, abusive, and hateful speech, mainly from social media platforms like Twitter, Facebook, Gab, and Reddit (see Fortuna *et al.* (2020), for definitions). Recently, the diversity of data sources has expanded, with corpora including comments from news media, Wikipedia, chat rooms, forums, as well as messaging services like WhatsApp. Several studies have provided an organized overview of the domain by cataloging existing datasets (Madukwe *et al.*, 2020; Alkomah and Ma, 2022). Despite recent advances, very few datasets collecting aggressive conversations are available for computational analysis of cyberbullying situations. Table 1 lists all datasets related to conversations in which this phenomenon is observed. By "conversation", we refer to interactions involving at least two human speakers who alternate turns, resulting in non-linear and intertwined discourse. Formally, we considered datasets providing at least one sentence with its preceding or following context, as well as a turn of speech (a complete interaction between two speakers). The collected resources are analyzed and com-

pared across six dimensions: language, number of entries, source of collected data, provided annotation layers, number of speakers per conversation, and resource format.

Dataset	Language	Entries	Source	Annotation	Structure	Type
ConvAbuse	English	4,185	Chatbots	Abusive Type Severity Target Directness	Two-party	Utterance + context (the agent's turn plus the previous turn of both user and agent)
TOXICCHAT	English	10,166	Chatbots	Toxicity Jailbreaking	Two-party	User prompt + agent's turn
Wikipedia Abuse Corpus	English	382,665	Wikipedia comments	Personal attack Aggression Toxicity	Multiparty	Reconstructed conversations
CyberAgression Ado-v1 & -v2	French	3,552	Scripted conversations	v1/v2 Hate v1 Role v1 Target v1 Verbal abuse v1 Humour v2 Intention v2 Context	Multiparty	Full conversations
WhatsApp Dataset	Italian	2,066	Scripted conversations	Cyberbullying Role Type Sarcasm Offensive	Multiparty	Full conversations

Table 1. *Conversational datasets available for tasks related to online hate detection.*

ConvAbuse consists of data from two-party conversations (1 human interlocutor and 1 conversational agent) in English between users and three different conversational AI systems (ELIZA, CarbonBot, and Alana v2) (Curry *et al.*, 2021). Each entry is provided in context, including the target input associated with the system's output and the preceding turns (when available) from both the user and the system. This dataset publicly available provides around 4,000 annotated entries using a hierarchical annotation schema identifying whether the target content is abusive or not, the severity level, type, target, and the presence of implicit or explicit content. *TOXICCHAT* (Lin *et al.*, 2023) is a dataset consisting of 10,166 user prompts, with 5,634 manually annotated for toxicity and jailbreaking. The dataset is based on pre-processed user interactions collected from a demo of the popular open-source chatbot Vicuna, comprising both user prompts and the corresponding agent responses. *Wikipedia Abuse Corpus* (WAC) consists of conversations in English reconstructed from comments posted on Wikipedia talk pages, which are web pages associated with Wikipedia articles where editors can interact. Editors typically write explanatory messages (on average 1,000 characters) about changes made to the articles. The reconstructed conversations are based on discussions in response to these comments. These discussions are relatively short, with the majority consisting of fewer than 20 messages. A dataset of approximately 193,000 conversations, including 383,000 messages annotated as abusive or not, is publicly available (Cecillon *et al.*, 2020). *CyberAgressionAdo V1 & V2* are

two datasets presented respectively in Ollagnier *et al.* (2022) and Ollagnier (2024), consisting of multiparty conversations in French mimicking cyberbullying situations that may occur among adolescents. This data was collected during role-playing games in several high-schools and middle schools and annotated considering various layers, such as participant roles, the presence of hate speech, the type of verbal abuse, the authors' intentions (what they aim to accomplish or convey through their messages), the context in which these messages are situated as responses, etc. The released corpus includes about 3,000 entries, with 19 conversations of approximately 187 messages each exchanged between 5 to 7 adolescents. *WhatsApp Dataset* is a corpus of interactions in Italian, featuring cyberbullying situations, collected during an experiment on WhatsApp with high-school students. It has been annotated in terms of cyberbullying roles, types of cyberbullying, presence of sarcasm, and whether it is an offensive message or simply a joke (Sprugnoli *et al.*, 2018). This corpus includes around 2,000 messages exchanged among about 10 adolescents in 10 different conversations, each comprising approximately 207 messages. Two other corpora similar to conversations, as defined above, have been released; however, for the *Space-Origin* corpus (Papegnies *et al.*, 2017), it is based on proprietary data, and for the *Hateful Messages* corpus (Fillies *et al.*, 2023), no access link is provided, explaining their absence in Table 1. *Ruddit* (Hada *et al.*, 2021), the *Reddit Contextual Abuse Dataset* (Vidgen *et al.*, 2021), and the dataset introduced in Tufa *et al.* (2024) provide conversation threads extracted from Reddit, annotated specifically for online hate-related detection tasks. Similarly, *DeTox* consists of 10,278 annotated German social media tweets, half of which are part of coherent conversation segments (reply trees) annotated for toxicity, criminal relevance, and discrimination types. However, these datasets cannot be considered as conversations due to the dynamics of thread-based interactions on these platforms, which differ substantially from turn-taking, non-linear, and interwoven discourse. On Reddit and Twitter, conversation threads are structured hierarchically around an initial post, with comments grouped beneath it. Each comment can reply to a previous one, forming a tree-like structure. These branches often remain unrelated and involve different users, making it impractical to treat such thread structures as authentic conversations. Literature also reports the use of conversational datasets in other contexts (Ganesh *et al.*, 2023); however, these resources do not include annotations allowing for the development of methods dedicated to online hate detection.

From this analysis, it appears that access to conversational data from real-world applications, curated and annotated for the development of online hate detection tools, is more than limited. Each dataset offers solutions to overcome social media privacy policies that restrict the collection of such data. For example, the *WhatsApp Dataset* and *CyberAgressionAdo V1 & V2* are based on a data-collection methodology that closely resembles natural interactions. Indeed, human-machine interactions, such as those provided in the *ConvAbuse* and *TOXICCHAT* corpora, cannot fully replicate the complexity and dynamics of human-to-human interactions. Similarly, reconstructing conversations based on comments like those in the WAC corpus does not replicate the intrinsic nature and dynamics of such data. Furthermore, conclusions from pre-

vious studies support that role-playing games are a more valid measure of authentic language use than more traditional data collection methods such as interviews or self-assessment questionnaires (Kasper, 1999; Tran, 2006). Beyond collection methodologies, the provided annotations mainly focus on identifying and characterizing abuse, thus limiting the computational analysis of cyberbullying situations in this context. Indeed, online hate remains a complex and multifaceted phenomenon shaped by a multitude of linguistic, contextual, and social factors in general (Baider, 2020). This observation is consistent with the research presented in Kumar *et al.* (2022) and Ollagnier (2024), where the utilization of pragmatic-level information provides descriptors to enhance the understanding of this phenomenon. Moreover, another characteristic of conversational data, especially multiparty ones, is the presence of multiple participants or interlocutors, whose identification is a separate task in addition to identifying abuse. This aspect is even more crucial in cyberbullying situations where participants are involved differently. It may involve the victim, the harasser, and bystanders, which is important to distinguish to nuance interpretations of abuse, as demonstrated in these two studies (Ollagnier *et al.*, 2023a; Ollagnier *et al.*, 2023b). In conclusion, the *CyberAgressionAdo V1 & V2* datasets appear to be the most suitable for the computational analysis of this type of phenomenon occurring within conversations. Additionally, the *V2* version provides annotations allowing for the study of the interplay of different aspects related to the practices underlying the operationalization of cyberbullying situations. While the aforementioned datasets are valuable, their relatively small size and limited topic diversity constrain their capacity to comprehensively capture the breadth of bullying practices, sensitive themes, and participant roles observed in real-world online aggression scenarios. To address this limitation, we build on both existing materials and newly developed resources to introduce *CyberAgressionAdo-Large*, a corpus comprising 36 conversations and a total of 5,789 entries. To our knowledge, this corpus stands out as the largest publicly available dataset of its kind, distinguished by its diversity (covering four sensitive topics) and its in-depth analysis, featuring six layers of annotation addressing multiple analytical dimensions. Additionally, we provide annotations from each annotator to explore the possibilities of developing perspectivist approaches, aiming to preserve the divergence of opinions and integrate them into the process of developing machine-learning methods (Cabitza *et al.*, 2023).

3. CyberAgressionAdo-Large: construction

The *CyberAgressionAdo-Large* dataset was developed following the collection process and annotation scheme introduced in the initial works presented in Ollagnier *et al.* (2022) and Ollagnier (2024). The experimental setting conducted in schools, along with the scenarios used for the role-playing games and the guidelines for applying the multi-label, fine-grained tagset, adhered to the same protocol, as detailed below.

3.1. *Data collection*

CyberAgressionAdo-Large was created from multiple data collection efforts conducted at four French high-schools and one middle school, involving approximately 243 participants. Our intervention in schools was part of a broader effort to raise awareness about cyberbullying and hate speech, aiming to provide students with additional means to understand and better address this phenomenon. The initial contact with students involved introducing them to artificial intelligence and its potential role in detecting harmful online messages (1.5 hours). Then, students were asked to complete an anonymous questionnaire elaborated by the sociologist involved in the study, aiming to collect data on their online behavior (e.g., time spent on the web, on social media) and their perception of cyberbullying phenomena (10 to 15 minutes). Researchers then introduced the practical phase, during which students participated in a role-playing game that mimicked cyberbullying situations occuring on instant messaging platforms. Each student had a computer to work with and had to log into an Internet Relay Chat (IRC) with a pseudonym provided at their discretion during each game, ensuring fully anonymous data collection. Each role-playing game lasted on average 45 minutes. Teachers were present in the room but were in no way involved in the role-plays. A few weeks after the experimentation, based on feedback from the sociologist's survey, a two-hour meeting with the students was organized to discuss cyberbullying issues and online hate speech with them and their teachers. During this meeting, students could exchange ideas and share their personal experiences and feelings about conducting this experiment. Regarding the latter point, students were asked to fill out a second questionnaire to share their perceptions of the advantages and disadvantages of this experimental method. Since young people are the actors and experts of their own lives, we deemed it relevant to consult them to avoid misinterpretations or to confine them to our own representations, which, in the context of developing tools for detecting and preventing this phenomenon, could lead to biases (Alderson and Morrow, 2011).

3.2. *Scenarios*

Created in collaboration with a sociologist and an expert in education sciences, the scenarios address topics commonly reported during cyberbullying incidents, including cyberhate related to ethnic origin, religion, obesity, and homophobia. Table 2 presents some examples of scenarios proposed to students. These scenarios were developed based on interviews and case studies conducted in French secondary schools reported in Blaya and Audrin (2019), thus relying on authentic negative experiences encountered by young people. We included different types of situations: obesity, religion, ethnic origin, and homophobia. These situations were selected based on research showing that overweight students (Puhl *et al.*, 2017) and LGBT+ individuals are more likely to be discriminated against and harassed (online) (Bucchianeri *et al.*, 2014), and that cyberhate based on origin and religion is one of the types of victimization that has increased the most in recent decades (Blaya and Audrin, 2019; Llorent

et al., 2016; Räsänen *et al.*, 2016), and that the processes of exclusion and discrimination related to weight are similar to racism, sexism, and gender-based harassment (Van Amsterdam *et al.*, 2012). Obese and overweight students are more likely to be victims of bullying (Kahle and Peguero, 2017).

In these role-playing scenarios, participants were assigned specific active roles reflecting varying levels of involvement in cyberbullying situations. These roles included: the bully, who initiates the harassment; the victim, who is the target of the harassment; the victim supporter, who defends the victim; the bully supporter, who assists or encourages the bully's actions; and the conciliator, a mutual friend of both the bully and the victim who intervenes to mediate and resolve the conflict. Additionally, a moderator role was introduced to ensure that the interactions adhered to the rules of the role-playing game. This role, which remained passive and observational, was fulfilled by one of the researchers present during the data collection process. Since the role-playing game represents a protected space to experiment with cyber violence, we avoided having students play the victims, with the victims always being represented by researchers from our team who were not physically present in the experimentation room. In order to involve all students in the role-playing game, some roles were duplicated and embodied in the same scenario. The number of bullies could vary between 1 and 2, the victim supporter role between 1 and 3, and the bully supporter role between 2 and 3. All other actively involved participants, i.e., the victim and the conciliator, were played by one person per scenario. In general, each scenario was played by 5 to 7 people. Students were randomly assigned to a scenario and a role (regardless of their gender). In a few cases, teachers advised us to avoid assigning a certain role to a student considering previous class dynamics and the student's behavior or personal characteristics.

4. CyberAgressionAdo-Large: annotation

The annotation scheme utilized in this study builds on the schema introduced in Ollagnier (2024). This multi-label, fine-grained tagset encompasses six distinct annotation layers, including participant roles, the presence of hate speech and the type of verbal abuse. Furthermore, it incorporates a detailed hierarchical structure aimed at capturing the communicative intentions behind each message and the contextual factors influencing its production. Table 3 presents the statistical properties of the *CyberAgressionAdo-Large* dataset, while Table 4 provides a detailed description of the annotation schema. The complete annotation guidelines are publicly accessible on the *CyberAgressionAdo-Large* project webpage [2].

2. `Https://github.com/aollagnier/CyberAgression-Large/`.

Scenario	Topic
Julie and Léa use to hang out together and are walking in the schoolyard holding hands. Emilie, jealous of Julie, posts their photo on Snapchat and makes mean comments about their relationship, insinuating that they are lesbians. Marie tries to intervene to defend Julie and Léa, but Emilie brings her best friends, Elodie and Anna, with her, and they try to exclude them from their friend group in class and on social media. Arthur, who is friend with both Julie, Léa, and Emilie, tries to intervene by explaining to them that it's pointless and that they should stop arguing.	homophobia
In the cafeteria, Paul, who is a bit overweight, has his dessert stolen by his table neighbor, Brice, who is also in his class. Brice tells him he's already fat enough and doesn't need to eat, while he eats the dessert. Meanwhile, Julien films the scene and shares it on social media, commenting on Paul's appearance, his gluttony, and his lack of control, which makes everyone laugh. Justine and Thibaut try to defend him, and Pierre, a friend of Paul's but also of Brice and Julien, tries to stop the teasing.	obesity
Justine is Jewish. On her profile, she posts a picture of her little brother's Bar Mitzvah. Léo and Guillaume, Justine's classmates, share the photo with harmful comments against Jews, including caricatures. Aurélie and Isabelle, while looking at the photo, also laugh. Léa and Anna, friends of Justine, try to defend her in the chat with the help of Amine to put an end to the harassment against Justine and her religion.	religion
Sophie and Lucas have been together for a few months and attend the same school. During a school trip, taking advantage of Sophie's absence, who stayed home with the flu, Lucas secretly kisses Silvia, a classmate of Sophie. Sophie discovers Lucas's betrayal through her friend Adrien, who witnessed the scene. Thinking that Silvia had flirted with Lucas, Adrien starts insulting Silvia on the WhatsApp chat, aided by Théo, Diana, and Camille: "She's here because they didn't want her at home! She has no business being here. She came to steal other people's boyfriends. Besides, they're all thieves." Soan, a classmate, decides to defend Silvia by blaming Lucas. Herbert, a friend of both Adrien and Silvia, intervenes to put an end to the harassment between Adrien and Silvia.	ethnic origin

Table 2. *Examples of role-playing scenarios on each sensitive topics proposed to students.*

Metric	Value
Number of conversations	36
Number of lines	5,789
Number of tokens	36,299
Average messages per conversations	156.45
Average length of messages (tokens)	6.47

Table 3. *Statistics of the CyberAgressionAdo-Large.*

Aggression		
Code	**Aggression Level**	**TAG**
1.1	Overtly Aggressive	OAG
1.2	Covertly Aggressive	CAG
1.3	Non-Aggressive	NAG
Role/Target		
Code	**Attribute**	**TAG**
1.A 1.1	victim	victim
1.A 1.2	victim support	victim_support
1.A 1.3	bully	bully
1.A 1.4	bully support	bully_support
1.A 1.5	conciliator	conciliator
Verbal Abuse		
Code	**Attribute**	**TAG**
1.B 1.1	Blaming	BLM
1.B 1.2	Name-calling	NCG
1.B 1.3	Threat / Coercion	THR
1.B 1.4	Denigration	DNG
1.B 1.5	Aggression-other	OTH
Discursive Level		
Code	**Intention/Context**	**TAG**
2.1	Attack	ATK
2.2	Defend	DFN
2.3	Counterspeech	CNS
2.4	Abet and Instigate	AIN
2.5	Gaslighting	GSL
2.6	Conflict-resolution	CR
2.7	Empathy	EMP
2.8	Other	OTH

Table 4. *The CyberAgressionAdo-Large tagset.*

4.1. *Aggression level*

This label is based on a multiclass schema comprising the categories OAG, CAG, and NAG. Label assignment is performed by interpreting aggression within its context, requiring annotators to consider extralinguistic knowledge and the perspectives of both the author and the recipient, including their roles and discursive postures. Detailed definitions and corresponding examples for each aggression label are provided below.

1.1 **Overt Aggression (OAG)**: This refers to communication, whether in speech or text, where aggressive behavior is explicitly expressed. It often involves offensive

or hostile language, explicit threats, hate speech, derogatory terms, or direct insults. Overt aggression may also arise from specific lexical items, features, or syntactic structures whose aggressive nature becomes apparent when contextualized with extralinguistic knowledge and the perspectives of both the author and recipient.

Example: The French sentence “woaa!! mate le cachalot” (EN: “woah !! look at the whale”) demonstrates overt aggression in the context of cyberbullying related to obesity. Here, “le cachalot” (the whale) is derogatory and offensive, mocking someone based on their weight. The phrase “mate” (look at) adds a mocking tone, inviting others to ridicule the individual. The exclamation marks and overall tone further emphasize the aggressive nature of the statement.

1.2 **Covert Aggression (CAG)**: This form of communication employs linguistic strategies to mask aggression beneath subtle or indirect expressions, avoiding explicit threats or derogatory language. While covert aggression is often subtle, it can also include non-subtle expressions that still convey aggressive intent despite an attempt to conceal it. Common strategies include figurative language (e.g., sarcasm, irony, black humor, exaggeration, metaphor), rhetorical questions, euphemisms, fallacies, or circumlocution.

Example: The sentence “T’as vraiment des fringues de ouf, mec, personne peut rivaliser avec ton style” (EN: “You’ve got some crazy clothes, dude, nobody can compete with your style”) appears to be a compliment. However, the phrase “des fringues de ouf” (crazy clothes) and “personne peut rivaliser” (nobody can compete) carry a sarcastic and mocking tone, revealing covert aggression.

1.3 **Non-Aggression (NAG)**: This category includes any text or speech devoid of hostile or harmful intent. It excludes explicit derogatory language, threats, or expressions of harm towards individuals or groups, as well as linguistic strategies that might subtly imply aggression or intimidation.

4.2. *Role/target*

Five specific active roles are used to represent varying levels of involvement in cyberbullying situations, depicting both the fictional roles embodied by participants during the scenarios and the target(s) of online hate. Target annotations are applied exclusively to messages identified as OAG or CAG. As described in Section 3.2, these roles include: (1.A 1.1) the victim, who is the individual being harassed; (1.A 1.2) the supporter of the victim, who defends him; (1.A 1.3) the bully, who initiates the harassment; (1.A 1.4) the supporter of the bully, who collaborates in or supports the bully’s actions; and (1.A 1.5) the conciliator, a mutual friend of the bully and the victim who intervenes to mediate and resolve the conflict.

4.3. *Verbal abuse*

Cyberbullying can take many forms, with verbal abuse being prevalent among them. It may include harassment, which involves sending repetitive and offensive messages to a target, cyberstalking (sending repetitive threatening communications), flaming, which entails sending messages containing abusive and vulgar terms such as insults, gossip, or mockery, and denigration (Bauman, 2014; Tokunaga, 2010; Watts *et al.*, 2017). Five types commonly encountered in written language are annotated here, and these are exclusively assigned to messages identified as OAG or CAG:

1.B 1.1 **Blaming (BLM)**: This involves making the individual believe they are responsible for the abuse they are experiencing, attributing it to their actions, words, or behavior. **Example:** "*on la traiterait pas de truie si elle avait pas autant de graisse*" ("she wouldn't be called a pig if she didn't have so much fat").

1.B 1.2 **Name-calling (NCG)**: Refers to abusive, insulting, or derogatory language aimed at undermining the self-esteem, personal worth, and self-perception of the targeted individual. **Example:** "*té qu1 putain de mongol*" ("you're such a fucking retard").

1.B 1.3 **Threat (THR)**: These statements are intended to intimidate, control, or manipulate the victim, coercing them into submission. **Example:** "*je vais venir en bas de chez toi, tu vas voir qui va plus parler*" ("I'm going to come to your house, and you'll see who won't be talking anymore").

1.B 1.4 **Denigration (DNG)**: Disparaging remarks aimed at attacking the reputation of the targeted person, belittling, discrediting, and tarnishing their image. These remarks are deliberately hurtful, non-constructive, and malicious. **Example:** "*les filles comme toi, ça me dégoûte*" ("girls like you disgust me").

1.B 1.5 **Other aggression (OTH)**: Covers content that includes deliberately harmful, abusive, insulting, or derogatory language that does not align with the other defined categories. **Example:** "*va crevé en enfer*" ("go die in hell").

4.3.1. *Discursive level*

The intention and context categories form two distinct layers, encompassing classifications such as attack (ATK), defense (DFN), counter-speech (CNS), instigation (AIN), gaslighting (GSL), conflict resolution (CR), and empathy (EMP). The purpose of label assignment is to decipher the discursive function of exchanged messages based on their underlying intentions, covering both aggressive and non-aggressive utterances. This annotation serves a dual purpose: first, to uncover the authors' intentions (what they aim to achieve or convey through their messages), and second, to establish the contextual framework in which these messages function as responses. Below, we provide the definitions and examples for each label.

2.1 **Attack (ATK)**: Any form of communication that intentionally exhibits overt or covert aggression towards victims, their supporters, or even conciliators. Such communication may involve insults, threats, mockery, exclusion, taunting, and dis-

crediting. This behavior is exclusive to bullies and their supporters and can manifest either as a deliberate act aimed at inflicting harm or as a means to escalate the level of violence.

```
User1: [ATK] ALLEZ MANIFESTE TOI GROS PORCS. / (EN) GO
    ON, SHOW YOURSELF, YOU FAT PIGS.
User2: [ATK] User3 le cachalot. / (EN) User3 the sperm
    whale.
```

2.2 **Defend (DFN)**: Any text/speech aiming to protect oneself or others from perceived attacks. It is characterized as an impulsive and non-deliberate response, which can be either aggressive or non-aggressive, and may be in retaliation for real or perceived attacks. This behavior is exclusive to victims, their supporters, or conciliators and may involve strategies such as challenging and refuting the abuser's messages.

```
User1: [ATK] jalouse de quoi mon pote tu me dégoute.
/ (EN) jealous of what my friend you disgust me.
User2: [DFN] t'es blanche comme un c*l tu crois t mieux
    User1? / (EN) you're as pale as an *ss do you think
    you're better User1?
```

2.3 **Counterspeech (CNS)**: Any non-aggressive response to harmful speech, aiming to undermine it. It employs strategies like presenting facts, highlighting contradictions, warning of consequences, and denouncing hate. It is initiated by victims, supporters, or conciliators.

```
User1: [DFN] tu sais dire d'autres choses à part ça ? /
    (EN) Do you know how to say anything else apart from
     that?
User2: [CNS] ça se fait pas en plus de prendre en photos
     / (EN) It's not right in addition to taking
    pictures.
```

2.4 **Abet/instigate (AIN)**: Messages supporting, encouraging, or validating previous negative messages, inciting aggression either beforehand (instigation) or during/after the act (abetment). These messages typically escalate conflicts or foster a hostile atmosphere, often initiated by bullies and their supporters.

```
User1: [ATK] qui les supp du groupe la / (EN) Who
    removes them from the group there?
User2: [AIN] je vais les supprimer / (EN) I am going to
    delete them.
```

2.5 **Gaslighting (GSL)**: Any text/speech minimizing or distorting another person's trauma or memory, aiming to manipulate their perception of reality and exert control. This includes tactics like denying or downplaying harm, blaming the victim, questioning their memory, invalidating their feelings, and using group consensus to make them doubt themselves.

```
User1: [ATK] wsh tu parle pas comme ca je vais te
    dechire / (EN) Hey don't talk like that I'm going to
     tear you apart.
User2: [GSL] User1 t es changer wsh / (EN) User1 you've
    changed seriously.
```

2.6 **Conflict-Resolution (CR)**: Any communication aiming to resolve conflicts and de-escalate situations without resorting to aggression. This includes mediation to resolve conflicts, mitigation to lessen the impact of cyberbullying, and education to promote appropriate online behavior. CR messages are consistently non-aggressive and are typically initiated by victim supporters and conciliators.

```
User1: [GSL] c toi ta un problème grosse p*te / (EN) You
    're the one with a problem you big sl*t.
User2: [CR] mais calmez-vous chaqu'un s'est préférence /
     (EN) calm down, everyone has their preferences.
```

2.7 **Empathy (EMP)**: Messages that demonstrate understanding, compassion, and support for those affected by cyberbullying. These messages may express sympathy, offer assistance or resources, validate emotions, or include self-empathy when victims acknowledge their own distress. This behavior is exclusive to victims, their supporters, or conciliators.

```
User1: [DFN] Elles sont juste immature de faire ca,
    preuve que c'est des gamines / (EN) They are just
    immature to do this, proof that they are kids.
User1: [EMP] User3 tu vaux mieux que sa / (EN) User3 you
    're worth more than this.
```

2.8 **Other (OTH)**: This category applies to cases where the appropriate tag for a message is unclear. It includes neutral utterances (messages without explicit or implicit harm), non-standard utterances such as incomplete sentences, one-word responses, sentence fragments, or emoticons and emojis used to convey emotions, attitudes, or reactions.

```
User1: [CR] Ca sert a rien de se prendre la tete
    franchement / (EN) There's no point in getting
    worked up, honestly.
User2: [OTH] quelle sexplique / (EN) What does it mean?
```

5. Disagreement vs. perspectives

A thorough analysis of the causes of disagreement among annotators, established in Ollagnier (2024), revealed that the various sources of disagreement stemmed from

(a) the clarity of annotation labels (i.e., their applicative scope), (b) text ambiguity, and (c) differences among annotators (i.e., their individual viewpoints), with the latter two being the most frequent causes. Following these findings, the annotation guidelines were improved by providing a precise description of the application cases for each annotation layer and corresponding labels. Based on these new guidelines, *CyberAgressionAdo-Large* was manually annotated by three experts from a text annotation specialized company. Table 5 presents the results of inter-annotator agreement obtained through the measurement of Krippendorff's Alpha across all conversations.

Label	**Score**
Hate	0.83
Target	0.88
Verbal abuse	0.82
Intention	0.87
Context	0.81

Table 5. *Measurement of Inter-Annotator Agreement on CyberAgressionAdo-Large.*

The obtained scores demonstrate a significant increase in inter-annotator agreement across all labels compared to those presented in Ollagnier (2024), and this on a dataset twice as large. This underscores the value of clear guidelines and discussions around challenging situations. Feedback from the annotators highlights that the main source of the remaining disagreement primarily revolves around varied interpretations arising from individual perceptions. Due to this finding we do not conduct here an analysis of annotator disagreements consisting in categorizing potential reasons behind conflicting annotations (Sandri *et al.*, 2023), such as sloppy annotation, ambiguity, missing information, and subjectivity. Supported by the perspectivist paradigm introduced in Cabitza *et al.* (2023), we decided to experimentally use residual disagreement to reflect individual viewpoints that may arise in the interpretation of online hate detection in a multiparty setting. We specifically investigate the application of the Weirdness Index (Ahmad *et al.*, 1999). In its original formulation, the W-index is used to extract domain-specific terms by comparing the relative frequencies of words in a domain-specific corpus vs. a generic corpus. The index was later applied to annotated corpora in order to rank the words according to their association to a specific human-provided label (Basile, 2020). We further adapt the method to automatically compute the association between each word and the disagreement between a pair of annotators. Given a pair of annotators a, b, the dataset is divided in two parts: $A_{a,b}$, i.e. the set of messages on which a and b agree on a specific label, and $D_{a,b}$, i.e. the set of messages on which a and b disagree. The Agreement Weirdness (AW) index for a word w is therefore defined as:

$$AW(w, a, b) = \sigma\left(\frac{P(w|D_{a,b})}{P(w|A_{a,b})}\right)$$

where $P(w|D)$ and $P(w|A)$ are the relative frequencies of w in D and A respectively, and σ is the standard logistic function. In essence, $AW(w)$ will be a number in $[0, 1]$ close to 1 if it occurs more often in texts on which a and b disagree, and close to 0 if it occurs more often in texts on which a and b agree.

In order to explore the results of the AW-index analysis, we introduce an ad-hoc visualization method where the level of disagreement associated with a word is correlated to its distance from fixed points. In each figure, the three blue dots represent the three annotators. Starting from the center of the triangle, each word is moved toward the line connecting two annotators based on $AW(w, a, b)$, which quantifies the degree of pairwise disagreement between annotators a and b for the word w. As a consequence, we may observe three main patterns:

– a words stays close to the center, if its disagreement levels are balanced across all three annotators;

– a word is close to an edge, if its disagreement is observed between a specific pair of annotators only. We call this **bilateral** disagreement;

– a word is close to a corner, if its disagreement is observed between a specific annotator and both the others, but not between the other two. We call this **multilateral** disagreement.

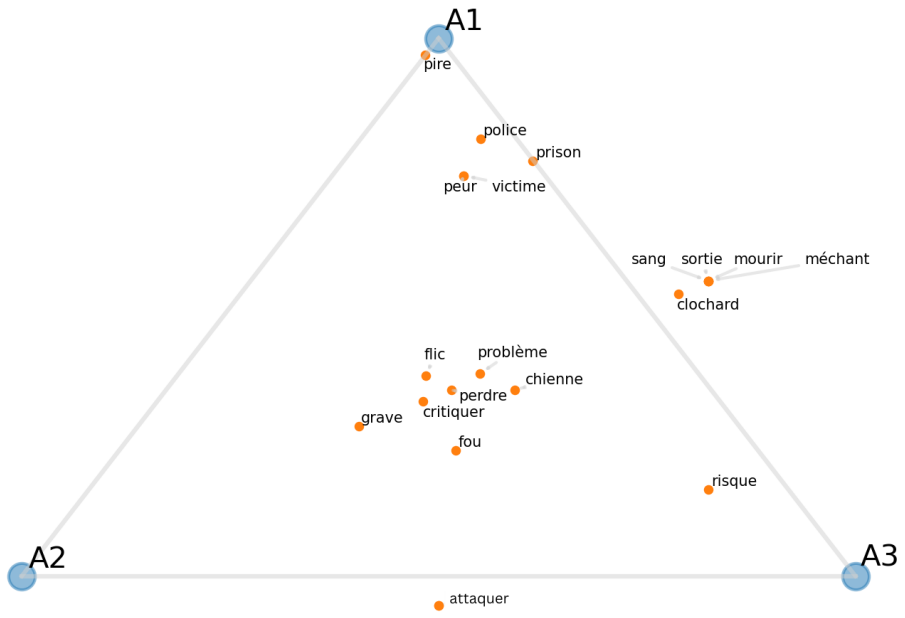

Figure 1. *A sample of visualization obtained for the emotion fear on the topic of obesity.*

Figure 1 presents a generated visualization using the AW-index on the full *CyberAgressionAdo-Large* dataset. Here, we can observe a bilateral disagreement

among the annotators A1 and A3 concerning utterances containing the reported words, as well as a multilateral disagreement for A1. The analysis of the visualizations obtained for the labels HATE and VERBAL_ABUSE, considering all the emotions, confirms the benefits of such a visualization method in unveiling factors of divergence of opinions, all the while facilitating the discovery of meaningful patterns and causal relationships among the systematic disagreements [3]. For instance, repeated bilateral disagreements among annotators could unveil divergence of opinions from a community perspective, while multilateral ones would refer to divergences influenced by individual perspectives. Moreover, this visualization method facilitates tailored research and the exploration of various analytical perspectives. For instance, when paired with a word affect lexicon such as *NRC VAD* (Mohammad, 2018), it highlights the potential interplay between the affective connotations of words and their interpretation in conveying hate.

In conclusion, observations reported in annotators' disagreements confirm that capturing pragmatic depictive social dynamics and interactions shaping conversations is achievable through the incorporation of annotation layers. Additionally, we believe that providing scenarios to annotators has influenced their interpretations, a factor that warrants further study to fully understand its impact. However, it remains undeniable that preserving annotations provided by different individuals is necessary in this context to access multiple potential interpretations of conversational data. This diversity of annotations allows for a comprehensive understanding of real-world scenarios and human values, thereby empowering the development of NLP systems to more accurately reflect and respect the intricacies of human communication and interaction. This is particularly crucial for addressing tasks related to online hate detection.

6. Analysis of cyberbullying practices

In this section, we present statistical evidence of cyberbullying practices observed in the annotated scenarios. The reported observations are based on frequent patterns identified at the instance level (i.e., a single message) or at the implicature level (i.e., a message and its subsequent reply). The patterns presented at the instance level are derived from observations that consider each annotator's perspective individually. In contrast, the observations reported at the implicature level are based on frequent patterns identified collectively among all the annotators.

In detail, Table 6 presents the most prevalent patterns observed in cyberbullying practices by analyzing individual author utterances (instances). Multiple recurrent cyberbullying behaviors are identified, which coincide with the roles of involvement concerning the type of hate expressed, the role of the individual(s) targeted, as well as the authors' intentions behind the posted message. Across all annotators, both bullies and their bystanders tend to target victims and their bystanders with the intention

3. The visualizations are available here: https://github.com/aollagnier/CyberAgression-Large/viz/.

ROLE	HATE	TARGET	INTENTION	FREQ. (%)
bully	OAG	victim	ATK	34.0
	NAG	-	OTH	19.5
	OAG	victim_support	ATK	13.2
bully_support	OAG	victim	ATK	29.9
	NAG	-	OTH	19.8
	CAG	victim	ATK	11.7
conciliator	NAG	-	OTH	30.9
	NAG	-	CR	24.1
victim	NAG	-	OTH	26.9
	NAG	-	DFN	17.3
	NAG	-	CNS	13.2
victim_support	NAG	-	OTH	22.9
	OAG	bully	DFN	17.0
	OAG	bully_support	DFN	15.0

(a) *Annotator 1*

ROLE	HATE	TARGET	INTENTION	FREQ. (%)
bully	OAG	victim	ATK	34.3
	NAG	-	OTH	20.1
	OAG	victim_support	ATK	13.4
	CAG	victim	ATK	11.4
bully_support	OAG	victim	ATK	29.1
	NAG	-	OTH	21.1
	OAG	victim_support	ATK	10.9
conciliator	NAG	-	OTH	29.7
	NAG	-	CR	22.8
victim	NAG	-	OTH	29.5
	NAG	-	DFN	15.6
	NAG	-	CNS	13.7
victim_support	NAG	-	OTH	24.5
	OAG	bully	DFN	16.9
	OAG	bully_support	DFN	14.0

(b) *Annotator 2*

ROLE	HATE	TARGET	INTENTION	FREQ. (%)
bully	OAG	victim	ATK	34.7
	NAG	-	OTH	20.6
	OAG	victim_support	ATK	13.4
	CAG	victim	ATK	10.8
bully_support	OAG	victim	ATK	28.8
	NAG	-	OTH	21.3
	OAG	victim_support	ATK	10.7
conciliator	NAG	-	OTH	31.1
	NAG	-	CR	22.1
victim	NAG	-	OTH	30.7
	NAG	-	DFN	16.0
	NAG	-	CNS	12.8
victim_support	NAG	-	OTH	24.0
	OAG	bully	DFN	16.8
	OAG	bully_support	DFN	14.5

(c) *Annotator 3*

Table 6. *The patterns of cyberbullying practices observed for each annotator at the instance level (i.e., a single message). The percentages indicate the frequency of each pattern relative to all messages sent by the corresponding authors across all scenarios annotated by the same annotator.*

of deliberately harming (ATK) them. The primary observed intention typically falls within a proactive aggression scheme characterized by repeated attacks intended to deliberately inflict harm or escalate the level of violence. Conversely, the intentions of victims and their supporters are predominantly characterized by OTH, which typically corresponds to neutral utterances (messages not conveying explicit or implicit harm). According to annotators' observations, neutral utterances often involve participants engaging in arguments, potentially impacting conflicts. DFN and CNS are also among the primary discursive devices used, representing behaviors aligned with a reactive aggression scheme describing impulsive aggressive responses to provocation. Conciliators are mainly non-aggressive, employing neutral utterances (29.7%-31.1%) or actively striving to resolve conflicts and de-escalate situations, with between 22.1%-24.1% of their messages dedicated to these objectives. While Annotators 2 and 3 noted the presence of covertly aggressive (CAG) messages directed at victims (10.8%-11.4%), it appears that the use of figurative devices is less common in this setting compared to other social media platforms (Ocampo *et al.*, 2023).

(source ↪ reply)	HATE	TARGET	INTENTION
bully	OAG	victim	ATK
↪ victim_support	OAG	bully	DFN
victim_support	OAG	bully	DFN
↪ victim	OAG	bully	DFN
bully_support	OAG	victim	ATK
↪ bully	OAG	victim	ATK
victim	OAG	bully	DFN
↪ victim_support	OAG	bully	DFN

Table 7. *Patterns of cyber-aggressions observed at the implicature level (i.e., one message and the subsequent reply), common to all annotators.*

Table 7 offers a comprehensive overview of cyberbullying practices observed across pairs of utterances (implicatures), taking into account all annotators' perspectives. In this context, "source" refers to the initial message, while "reply" denotes the immediate subsequent message. These pairs denote an implicature relationship as they comprise messages generated within the same context, with the "reply" often reliant on the preceding message for context and meaning. It's noteworthy that each recurring pattern involves distinct roles, shedding light on the intricate dynamics of cyberbullying situations. Additionally, these patterns consistently emerge among all annotators across all scenarios (with a support measure of 1.0), providing generalizable and reliable insights essential for studying this complex behavioral phenomenon. In detail, bystanders of the victim frequently intervene in bullying episodes (ATK setting) by directly assisting victims against the bullies. Victims and their bystanders tend to support each other against the bullies, while the bullies and their bystanders unite with the aim of jointly attacking the victims.

Overall, the observations derived from these tables offer an initial depiction of cyberbullying practices in this specific multiparty context, providing valuable insights into the complex nature of cyberbullying phenomena. Firstly, despite variances in annotators' perceptions, common practices being topic-agnostic emerge and recur in each scenario. Secondly, non-aggressive exchanges are prevalent and should be analyzed, as they can contribute to either the escalation or de-escalation of situations. A recent study presented in Kaliampos *et al.* (2022) confirms this finding by examining potential behaviors of bystanders in bullying episodes. It reports that neutral utterances by victims' supporters can aim to de-escalate tension, seek clarification, maintain normalcy, or subtly intervene without provoking further hostility. Lastly, while the proactive-reactive aggression scheme has been extensively studied in the operationalization of cyberbullying, it appears that peer support schemes play a crucial role in the unfolding of events (Cowie, 2014). Under the umbrella of peer support, activities such as befriending, peer counseling, conflict resolution, or mediation, as well as interventions in bullying situations, are included. These activities should be con-

sidered with differing intentions, depending on whether peer support is offered by the bullies or the victims.

7. Conclusion

In this paper, we introduce the *CyberAgressionAdo-Large* dataset, which applies a hierarchical, fine-grained tagset designed for annotating bullying narrative events in multi-party chat conversations. Currently, the dataset comprises 36 conversations in French, mimicking online aggression commonly observed among teenagers on private instant messaging platforms. Our data collection efforts are ongoing, with additional sessions planned in French high schools over the coming months to expand both the size and diversity of the dataset. Given that participants in the role-playing game have the freedom to influence their group's storyline, it is crucial to conduct more scenarios and gather additional data from schools to ensure comprehensive coverage of real-world bullying practices. Furthermore, we intend to enhance the existing tagset by incorporating labels that facilitate computational modeling of multi-party dialogues. These enhancements aim to support tasks such as identifying participant roles, managing initiative and turn-taking, and analyzing discourse relations, which are essential for detecting online hate and related phenomena effectively within this context.

8. Ethics statement

NLP research focusing on online aggression and harassment detection inevitably raises ethical considerations. In our work, we place significant emphasis on the importance of ensuring that students involved are fully informed, that the data collected replicate naturalistic interactions, and our support for an annotation methodology promoting diverse opinions and perspectives.

Firstly, all students under 18 participated with parental consent, receiving comprehensive explanations about research objectives, data usage, and associated risks. Transparency was paramount as both parents and students were informed about AI's potential benefits in detecting hostile online messages. Prior to participation, students underwent education on cyber aggression and AI to foster informed consent. Our research protocol underwent rigorous review and approval by each participating school, adhering to European ethical standards and university guidelines. Throughout the study, we maintained strict confidentiality, anonymity, and respect for participants' autonomy. To ensure a positive experience, we provided support during role-playing sessions and conducted post-session feedback and training on cyber aggression's impact on victims and perpetrators.

Secondly, the validity of our data collection process was validated by a sociologist and an expert in education sciences. The scenario designs were based on real experiences shared by young people, ensuring authenticity and relevance to actual online interactions. The spontaneous nature of multi-party chats minimized scripted

responses, aligning with research that shows role-plays provide a more genuine portrayal of natural language use compared to methods such as interviews, questionnaires, human-machine interactions, or reconstructing conversations from threads.

Finally, despite the lack of comprehensive sociodemographic information about the annotators provided by the company, our work underscores the importance of acknowledging and incorporating annotator subjectivity in NLP applications. Indeed, diverse annotator viewpoints can be utilized to mitigate biases and reflect real-world human values. Moreover, our corpus serves as a foundation for exploring perspectivist computational approaches to address subjective tasks in conversational data.

In conclusion, by addressing these ethical concerns and promoting diversity, the NLP community can significantly advance in combating online hate across diverse digital environments, relying on more effective, fairer, and transparent NLP models.

Acknowledgements

This work is funded under the IDEX UCA OTESIA "L'intelligence artificielle au service de la prévention de la cyberviolence, du cyberharcèlement et de la haine en ligne", and by the UCA Academy 1 project with the reference number C870A021 – D103 – ACAD1_FIN_17_20Y. It has also been supported by the French government, through the 3IA Cote d'Azur Investments in the project managed by the National Research Agency (ANR) with the reference number ANR-23-IACL-0001.

9. References

Ahmad K., Gillam L., Tostevin L. *et al.*, "University of Surrey Participation in TREC 8: Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER)", p. 1-8, 1999.

Alderson P., Morrow V., *The ethics of research with children and young people: A practical handbook*, Sage, 2011.

Alhashmi A. A., KM A. K., Eid A. A., Mansouri W. A., Othmen S., Miled A. B., Darem A. A., "TAXONOMY OF CYBERBULLYING: AN EXPLORATION OF THE DIGITAL MENACE", *Journal of Intelligent Systems and Applied Data Science*, 2023.

Alkomah F., Ma X., "A Literature Review of Textual Hate Speech Detection Methods and Datasets", *Inf.*, vol. 13, n° 6, p. 273, 2022.

Baider F., "Pragmatics lost?: Overview, synthesis and proposition in defining online hate speech", *Pragmatics and Society*, vol. 11, n° 2, p. 196-218, 2020.

Basile V., "Domain Adaptation for Text Classification with Weird Embeddings", *in* J. Monti, F. Dell'Orletta, F. Tamburini (eds), *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021*, vol. 2769 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020.

Bauman S., *Cyberbullying: What counselors need to know*, John Wiley & Sons, 2014.

Blaya C., Audrin C., "Toward an Understanding of the Characteristics of Secondary School Cyberhate Perpetrators", *Frontiers in Education*, vol. 4, p. 46, 2019.

Bucchianeri M. M., Eisenberg M. E., Wall M. M., Piran N., Neumark-Sztainer D., "Multiple types of harassment: Associations with emotional well-being and unhealthy behaviors in adolescents", *Journal of Adolescent Health*, vol. 54, n° 6, p. 724-729, 2014.

Cabitza F., Campagner A., Basile V., "Toward a Perspectivist Turn in Ground Truthing for Predictive Computing", *in* B. Williams, Y. Chen, J. Neville (eds), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, AAAI Press, p. 6860-6868, 2023.

Cecillon N., Labatut V., Dufour R., Linarès G., "WAC: A Corpus of Wikipedia Conversations for Online Abuse Detection", *in* N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (eds), *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, European Language Resources Association, p. 1382-1390, 2020.

Cowie H., "Understanding the role of bystanders and peer support in school bullying.", *International journal of emotional education*, vol. 6, n° 1, p. 26-32, 2014.

Curry A. C., Abercrombie G., Rieser V., "ConvAbuse: Data, Analysis, and Benchmarks for Nuanced Detection in Conversational AI", *in* M. Moens, X. Huang, L. Specia, S. W. Yih (eds), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Association for Computational Linguistics, p. 7388-7403, 2021.

Fillies J., Peikert S., Paschke A., "Hateful Messages: A Conversational Data Set of Hate Speech produced by Adolescents on Discord", *CoRR*, 2023.

Fortuna P., Soler Company J., Wanner L., "Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets", p. 6786-6794, 2020.

Gamal D., Alfonse M., Jiménez-Zafra S. M., Aref M., "Intelligent Multi-Lingual Cyber-Hate Detection in Online Social Networks: Taxonomy, Approaches, Datasets, and Open Challenges", *Big Data Cogn. Comput.*, vol. 7, n° 2, p. 58, 2023.

Ganesh A., Palmer M., Kann K., "A Survey of Challenges and Methods in the Computational Modeling of Multi-Party Dialog", *in* Y.-N. Chen, A. Rastogi (eds), *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, Association for Computational Linguistics, Toronto, Canada, p. 140-154, July, 2023.

Hada R., Sudhir S., Mishra P., Yannakoudakis H., Mohammad S. M., Shutova E., "Ruddit: Norms of Offensiveness for English Reddit Comments", *in* C. Zong, F. Xia, W. Li, R. Navigli (eds), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Association for Computational Linguistics, p. 2700-2717, 2021.

Kahle L., Peguero A. A., "Bodies and bullying: The interaction of gender, race, ethnicity, weight, and inequality with school victimization", *Victims & Offenders*, vol. 12, n° 2, p. 323-345, 2017.

Kaliampos G., Katsigiannis K., Fantzikou X., "Aggression and bullying: a literature review examining their relationship and effective anti-bullying practice in schools", *International Journal of Educational Innovation and Research*, vol. 1, n° 2, p. 89–98, Jul., 2022.

Kasper G., "Data collection in pragmatics research", *University of Hawai'i Working Papers in English as a Second Language 18 (1)*, 1999.

Kumar R., Ratan S., Singh S., Nandi E., Devi L. N., Bhagat A., Dawer Y., Lahiri B., Bansal A., Ojha A. K., "The ComMA Dataset V0.2: Annotating Aggression and Bias in Multilingual Social Media Discourse", *in* N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (eds), *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, European Language Resources Association, p. 4149-4161, 2022.

Lin Z., Wang Z., Tong Y., Wang Y., Guo Y., Wang Y., Shang J., "ToxicChat: Unveiling Hidden Challenges of Toxicity Detection in Real-World User-AI Conversation", *in* H. Bouamor, J. Pino, K. Bali (eds), *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Association for Computational Linguistics, p. 4694-4702, 2023.

Llorent V. J., Ortega-Ruiz R., Zych I., "Bullying and cyberbullying in minorities: Are they more vulnerable than the majority group?", *Frontiers in psychology*, vol. 7, p. 1507, 2016.

Madukwe K. J., Gao X., Xue B., "In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets", *in* S. Akiwowo, B. Vidgen, V. Prabhakaran, Z. Waseem (eds), *Proceedings of the Fourth Workshop on Online Abuse and Harms, WOAH 2020, Online, November 20, 2020*, Association for Computational Linguistics, p. 150-161, 2020.

Mohammad S. M., "Word Affect Intensities", *in* N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (eds), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*, European Language Resources Association (ELRA), 2018.

Ocampo N., Sviridova E., Cabrio E., Villata S., "An In-depth Analysis of Implicit and Subtle Hate Speech Messages", *in* A. Vlachos, I. Augenstein (eds), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, Association for Computational Linguistics, p. 1989-2005, 2023.

Ollagnier A., "CyberAgressionAdo-v2: Leveraging Pragmatic-Level Information to Decipher Online Hate in French Multiparty Chats", *in* N. Calzolari, M. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (eds), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, ELRA and ICCL, p. 4287-4298, 2024.

Ollagnier A., Cabrio E., Villata S., "Harnessing Bullying Traces to Enhance Bullying Participant Role Identification in Multi-Party Chats", *in* M. Franklin, S. A. Chun (eds), *Proceedings of the Thirty-Sixth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2023, Clearwater Beach, FL, USA, May 14-17, 2023*, AAAI Press, 2023a.

Ollagnier A., Cabrio E., Villata S., Blaya C., "CyberAgressionAdo-v1: a Dataset of Annotated Online Aggressions in French Collected through a Role-playing Game", *in* N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (eds), *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, European Language Resources Association, p. 867-875, 2022.

Ollagnier A., Cabrio E., Villata S., Tonelli S., "BiRDy: Bullying Role Detection in Multi-Party Chats", *in* B. Williams, Y. Chen, J. Neville (eds), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, AAAI Press, p. 16464-16466, 2023b.

Papegnies E., Labatut V., Dufour R., Linarès G., "Impact of Content Features for Automatic Online Abuse Detection", *in* A. F. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing - 18th International Conference, CICLing 2017, Budapest, Hungary, April 17-23, 2017, Revised Selected Papers, Part II*, vol. 10762 of *Lecture Notes in Computer Science*, Springer, p. 404-419, 2017.

Puhl R. M., Wall M. M., Chen C., Austin S. B., Eisenberg M. E., Neumark-Sztainer D., "Experiences of weight teasing in adolescence and weight-related outcomes in adulthood: A 15-year longitudinal study", *Preventive medicine*, 2017.

Räsänen P., Hawdon J., Holkeri E., Keipi T., Näsi M., Oksanen A., "Targets of online hate: Examining determinants of victimization among young Finnish Facebook users", *Violence and victims*, vol. 31, n° 4, p. 708-725, 2016.

Sandri M., Leonardelli E., Tonelli S., Jezek E., "Why Don't You Do It Right? Analysing Annotators' Disagreement in Subjective Tasks", *in* A. Vlachos, I. Augenstein (eds), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, Association for Computational Linguistics, p. 2420-2433, 2023.

Sprugnoli R., Menini S., Tonelli S., Oncini F., Piras E., "Creating a WhatsApp Dataset to Study Pre-teen Cyberbullying", *in* D. Fiser, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, J. Wernimont (eds), *Proceedings of the 2nd Workshop on Abusive Language Online, ALW@EMNLP 2018, Brussels, Belgium, October 31, 2018*, Association for Computational Linguistics, p. 51-59, 2018.

Tokunaga R. S., "Following you home from school: A critical review and synthesis of research on cyberbullying victimization", *Computers in human behavior*, vol. 26, n° 3, p. 277-287, 2010.

Tran G. Q., "The naturalized role-play: An innovative methodology in cross-cultural and interlanguage pragmatics research", vol. 5, Reflections on English Language Teaching, p. 1-24, 2006.

Tufa W. T., Markov I., Vossen P., "The Constant in HATE: Analyzing Toxicity in Reddit across Topics and Languages", *CoRR*, 2024.

Van Amsterdam N., Knoppers A., Claringbould I., Jongmans M., "A picture is worth a thousand words: Constructing (non-)athletic bodies", *Journal of Youth Studies*, vol. 15, n° 3, p. 293-309, 2012.

Vidgen B., Nguyen D., Margetts H. Z., Rossini P. G. C., Tromble R., "Introducing CAD: the Contextual Abuse Dataset", *in* K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (eds), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, Association for Computational Linguistics, p. 2289-2303, 2021.

Watts L. K., Wagner J., Velasquez B., Behrens P. I., "Cyberbullying in higher education: A literature review", *Comp. in Human Behavior*, 2017.