CyberAgressionAdo-v2: Leveraging Pragmatic-Level Information to Decipher Online Hate in French Multiparty Chats

Anaïs Ollagnier

CyberAgressionAdo-v2: Leveraging Pragmatic-Level Information to Decipher Online Hate in French Multiparty Chats

Anaïs Ollagnier
Université Côte d'Azur, Inria, CNRS, I3S
930 route des Colles, BP 145, 06903 Sophia Antipolis Cedex, France
ollagnier@i3s.unice.fr

Abstract

As a part of the release of the *CyberAgressionAdo-V2* dataset, this paper introduces a new tagset that includes tags marking pragmatic-level information occurring in cyberbullying situations. The previous version of this dataset, *CyberAgressionAdo-V1*, consists of aggressive multiparty chats in French annotated using a hierarchical tagset developed to describe bullying narrative events including the participant roles, the presence of hate speech, the type of verbal abuse, among others. In contrast, *CyberAgressionAdo-V2* uses a multi-label, fine-grained tagset marking the discursive role of exchanged messages as well as the context in which they occur –– for instance, attack (ATK), defend (DFN), counterspeech (CNS), abet/instigate (AIN), gaslight (GSL), etc. This paper provides a comprehensive overview of the annotation tagset and presents statistical insights derived from its application. Additionally, we address the challenges encountered when annotating pragmatic-level information in this context, conducting a thorough analysis of annotator disagreements. The resulting dataset comprises 19 conversations that have been manually annotated and is now available to facilitate further research in the field.

Keywords: cyberbullying, pragmatic-level information, annotation tagset

1. Introduction

In today's rapidly expanding era of social media, adolescents are dedicating a substantial amount of their time to various social networking platforms, seeking connection, information sharing, and common interest pursuits. This surge in social media engagement has brought both benefits, such as enhanced interaction and learning opportunities, and challenges, including significant exposure to offensive online content. In the Natural Language Processing (NLP) community, considerable efforts have been made in recent years to address the issue of online hate detection. These efforts have led to the development of various datasets, with a growing focus on instances of cyberbullying occurring within private instant messaging platforms (Sprugnoli et al., 2018; Ollagnier et al., 2022). Cyberbullying within these platforms is especially concerning, as it has witnessed substantial growth, particularly among teens (Aizenkot and Kashy-Rosenbaum, 2018). Recent studies have highlighted the limitations in the coverage of dimension scopes within existing annotation schemes tailored to conversational data (Ollagnier et al., 2023a,b). Indeed, online hate remains a complex and multifaceted phenomenon shaped by a multitude of linguistic, contextual, and social factors (Baider, 2020). Recently, various annotation schemes designed for "easy to access" data (Facebook, YouTube or Instagram) have emerged with the aim of addressing the pragmatic aspects involved in these behaviours (Kumar et al., 2018, 2022). However, their applicability is limited to social media channels exploiting a such structure and dynamic.

In this paper, we present the *CyberAgressionAdo-V2* dataset [1], which builds upon the data introduced by Ollagnier et al. (2022) and incorporates a modified version of the tagset discussed in Kumar et al. (2018). The modifications to this tagset were made with the goal of extending its usability to multi-party settings and ensuring comprehensive coverage of the communication goals associated with the various roles individuals assume in a bullying scenario. The revised tagset consists of six layers, two of which revolve around pragmatic considerations, encompassing intentions and context. These layers are designed to capture the authors' intentions and establish the context in which messages serve as responses. The reliability and interoperability of this tagset have been addressed as part of the analysis of inter-annotator (dis)agreement. Additionally, a statistical analysis of the dataset was conducted using pattern mining techniques, resulting in the discovery of cyberbullying practice patterns. This paper makes a threefold contribution: (1) it contributes to the development of solutions aimed at addressing diversity in digital harassment, (2) it leverages annotators' disagreements in tagset development, and (3) it explores pattern mining to advance the modeling of complex communication schemes prevalent in cyberbullying. In total, the *CyberAgressionAdo-V2* dataset comprises 19 conversations that have been manually annotated, covering 4 different sensitive topics: religion, ethnicity, homophobia,

[1] The dataset is publicly available: https://github.com/aollagnier/CyberAgressionAdo-V2.public

and obesity.

2. Related Work

The NLP community has made substantial progress in developing semantic frameworks that aim to address the complex and multi-faceted nature of hate speech, such as specific targets (groups targeted), nuances (abusive, toxic, dangerous, offensive or aggressive language) or rhetoric devices (slurs, obscenity, offences or sarcasm), among others. Several surveys have provided structured overviews of existing datasets (Fortuna and Nunes, 2018; Vidgen and Derczynski, 2020; Poletto et al., 2021). Most of the available datasets have originated from Twitter and predominantly rely on a binary scheme to categorize the presence or absence of hate speech, as introduced in Poletto et al. (2019) and AbdelHamid et al. (2022). Some datasets have adopted multi-level annotation schemes that account for different hate speech phenomena such as hate speech and offensive language (Martins et al., 2018; Mathur et al., 2018) or sexism and aggressive language (Bhattacharya et al., 2020). The NLP community has also emphasized the importance of multilingual and multimodal datasets to gain insights into the interrelationships between various phenomena as well as to address linguistic and cultural heterogeneity (e.g., bullying role of involvement in aggressive French multiparty chats (Ollagnier et al., 2022) or biases and threats in Meitei, Bangla and Hindi (Kumar et al., 2022). In addition to these developments, numerous open shared tasks at NLP-related conferences have further advanced the field. Notable shared tasks include TRAC[2] (Workshop on Trolling, Aggression, and Cyberbullying), Computational Ethics Tasks at EVALITA'20 & 2023[3], Multilingual Gender Biased and Communal Language Identification in ComMA@ICON[4] at ICON2021, SemEval'19 Tasks 5 (HatEval[5]) & 6 (OffensEval[6]) at NAACL HLT'19, Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC'19, 2020 & 2022[7]), *inter alia*.

While most research has focused on a specific phenomena and its computational processing, little attention has been given to the pragmatic and syntactic structures at play within cyberbullying situations. The initial groundwork for addressing this gap was laid in a prior study (Kumar et al., 2018), where labels describing discursive roles and effects were discussed. This effort was subsequently expanded upon in Kumar et al. (2022), wherein the scheme was restructured and enhanced to more effectively encompass the various communication methods employed in such contexts. Nevertheless, the proposed tagset, while aspiring to offer a versatile tool for annotating aggressive discourse, exhibits certain limitations. Firstly, it has been developed based on observations from social network-based platforms (Facebook and Twitter), which do not entirely align with the conversational dynamics found in multiparty chat environments. Additionally, its representation is limited concerning the discursive roles, which do not capture the complete spectrum of communication goals linked to the roles individuals assume in a bullying scenario (ElSherief et al., 2018). This limitation is particularly critical because it has been identified as a potential source of bias in sub-tasks related to hate speech detection. For instance, participants providing support to a victim might exhibit aggressive behavior in defense, which can lead to misclassification in the task of participant role identification (Ollagnier et al., 2023a,b). In light of these limitations, we propose revising and expanding the initial scheme to broaden its inclusiveness and scope of applicability. The proposed tagset includes a new layer and additional labels aiming at more accurately capturing participants' intentions and deciphering communication practices in such scenarios.

3. Corpus Annotation

The proposed resource is a modified version of the dataset initially introduced in Ollagnier et al. (2022), referred to as the *CyberAgressionAdo-v1 dataset*. This dataset encompasses aggressive multiparty chats conducted in the French language, acquired through a role-playing game designed to mimic online aggression scenarios that teenagers might encounter on private instant messaging platforms. This dataset consists of 19 conversations, annotated using a fine-grained schema that spans five distinct layers including participant roles, the presence of hate speech, the type of verbal abuse, and the identification of figurative language usage. The modified version of this scheme relies on the tagset discussed in Kumar et al. (2022). This tagset, constructed from observations of comments collected from various social media, consists of a hierarchical, fine-grained structure aiming at categorizing different types of aggression and the associated "context" in which they occur. Table 1 presents statistics for the *CyberAgressionAdo-v2* dataset. Appendix 9.1 provides samples of scenarios used to collect the data. The following sub-sections present the

[2] https://sites.google.com/view/trac2022/
[3] https://www.evalita.it/campaigns/evalita-2023/
[4] https://competitions.codalab.org/competitions/35482
[5] https://competitions.codalab.org/competitions/19935
[6] https://sites.google.com/site/offensevalsharedtask/
[7] https://hasocfire.github.io/hasoc/2022

proposed annotation scheme as well as the inter-annotator agreement experiments.

Metric	Value
Number of conversations	19
Number of lines	2921
Number of tokens	26289
Average messages per conversations	153.74
Average length of messages (tokens)	9.0

Table 1: Statistics of the *CyberAgressionAdo-v2*

3.1. Annotation

The tagset for both the aggression and discursive levels is given in Table 2. For labels that remain consistent with those used in *CyberAgressionAdo-v1*, they will not be further discussed in the following subsections – Target, Verbal Abuse. The full annotation guidelines is available on the *CyberAgressionAdo-v2* webpage[8].

Aggression		
Code	**Aggression Level**	**TAG**
1.1	Overtly Aggressive	OAG
1.2	Covertly Aggressive	CAG
1.3	Non Aggressive	NAG
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 2: The CyberAgressionAdo-v2 Tagset

[8] https://anonymous.4open.science/r/CyberAgressionAdo-V2public/

3.1.1. Aggression Level

In *CyberAgressionAdo-v1*, this label follows a binary scheme (hate speech or no hate speech). Nevertheless, the use of binary schemes is problematic as it can oversimplify the intricacies of language, potentially impacting the performance of automated computational methods. To address this limitation, we adopt the three broad levels introduced in Kumar et al. (2022) and detailed in Table 2 – OAG, CAG and NAG. However, these definitions diverge from earlier studies (Kumar et al., 2018) by placing an emphasis on interpreting aggression within its contextual framework, which involves considering extralinguistic knowledge (information beyond language) as well as the perspectives of both the author and recipient. In this approach, understanding online hate's perception becomes a multi-faceted process influenced by linguistic, contextual, and social factors – often underappreciated when distinguishing between covert and overt aggression (Benikova et al., 2017; Hartvigsen et al., 2022). These factors play key roles in determining whether aggressive content is perceived as implicit, subtle, or neither (Ocampo et al., 2023). For intance, the context can render overt aggression subtle, even in the absence of explicit derogatory terms, while seemingly subtle expressions may contain underlying explicit aggression. The adopted definitions and corresponding examples for each aggression tag are presented below:

1. **Overt Aggression**: Any communication, whether in the form of speech or text, in which aggressive behavior is explicitly expressed. 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 the aggressiveness becomes apparent when considered alongside extralinguistic knowledge and both author's and recipient's perception.

   The French sentence "woaa !! mate le cachalot" (EN: "woah !! look at the whale.") is overtly aggressive in the context of cyberbullying against obesity. The use of the term "le cachalot" (the whale) in this context is derogatory and offensive, as it is used to mock or insult someone based on their weight. The exclamation "mate" (look at) is used here in a mocking and offensive manner, inviting others to join in ridiculing or making fun of the person. The exclamation marks and the overall tone emphasize the aggressive and mocking nature of the statement.

2. **Covert Aggression**: 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.

   Considering the following 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."). On the surface, it may appear as a compliment, acknowledging the uniqueness of the person's fashion sense. However, the use of "de ouf" (crazy) and "personne peut rivaliser" (nobody can compete) adds a sarcastic and mocking tone to the statement.

3. **Non-Aggression**: 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.

3.1.2. Discursive Level

The initial iteration of *CyberAgressionAdo-v2* lacks comprehensive information pertaining to pragmatic aspects. Here, we introduce two novel layers (intention/context), building upon the framework of discursive analysis previously introduced in Kumar et al. (2022). In the aforementioned study, the discursive level is defined as the overarching function of a given comment within the discourse, encompassing categories such as attack (ATK), defend (DFN), counterspeech (CNS), abet/instigate (AIN), or gaslighting (GSL). As a part of the proposed annotation scheme, we reuse this tagset to classify messages, discerning their discursive function based on their underlying intentions. This information serves a dual purpose: firstly, to unveil the authors' intentions (what they aim to achieve or convey through their messages) and secondly, to establish the context within which these messages are situated as responses. Unlike the prior study, where the focus was on identifying the discursive role of aggressive texts, our annotation scheme extends beyond aggression to encompass all messages. It also includes two new roles (conflict-resolution and empathy) aiming at broadening the previous scheme to encompass non-aggressive messages and gain a deeper understanding of counterspeech. The definitions associated with discursive roles have been reviewed to more accurately capture the complete spectrum of communication goals associated with the roles individuals assume in a bullying scenario. Here are the definitions and examples for each tag:

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 discrediting. 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. **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 not, and may be in retaliation for attacks, whether real or perceived. They are performed exclusively in response to an attack or an abet, and they may employ strategies listed, including (ATK) techniques, or involve challenging and refuting the abusers' messages. This behavior is exclusive to victims, their supporters, or conciliators.

```
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?
```

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's 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
    .
```

4. **Abet/Instigate (AIN)**: Messages supporting, encouraging or validating previous negative messages, inciting aggression either beforehand (instigation) or during/after the act (abetment), and potentially escalate conflicts or foster a hostile atmosphere, typically in reaction to 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.
```

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. It 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. It's initiated by bullies and their supporters.

```
    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.
```

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 and communication. 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.
```

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 even 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.
```

8. **Other (OTH)**: in cases where it is challenging to determine the appropriate tag for a message with unclear intent. This includes situations such as neutral utterances (messages not conveying explicit or implicit harm), non-standard utterances like incomplete sentences, one-word responses, or sentence fragments, as well as the annotation of emoticons and emojis used to convey emotions, attitudes, and reactions.

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

3.2. Analysis of (dis)agreement

3.2.1. Inter-Annotator Agreement

To assess the effectiveness and reliability of our tagset, we carried out an inter-annotator agreement (IAA) experiment involving two experienced annotators with a background in computational linguistics. This experiment encompassed six conversations, representing 36.8% of the dataset. The results, detailed in Table 3, were measured using Cohen's Kappa coefficient. Since the roles of authors and targets were predefined, we did not

Aspect	Phase 1	Phase 2
Aggression	0.71	0.79
Verbal abuse	0.72	0.84
Intention	0.46	0.79
Context	0.53	0.66

Table 3: Inter-annotator agreement Phase 1 & 2

assess the agreement for these labels. However, we observed that agreement for certain labels (context and intention), fell below the desired threshold. To address this, we refined our guidelines and subsequently conducted a second round of agreement experiments, aimed at enhancing the consistency and accuracy of our annotations.

Initially, annotators faced challenges as they were restricted to annotating discursive roles primarily focused on aggressive messages, leading to significant disagreements. Furthermore, the complexity of multi-party chats, with narrative chains spanning multiple messages, required clearer contextual guidelines. Following feedback from annotators, we made two key revisions. First, we refined the assignment of discursive roles based on the role of involvement in bullying, ensuring consistency even if there was a shift in power dynamics (e.g., a victim acting like a bully). On the other hand, we established a set of rules to define the scope of a contextual window within a bullying narrative event in multi-party chats. In these chats, a narrative chain unfolds as a sequence of interconnected events linked through causal relationships. When addressing context, we examine the causal connections between the intentions conveyed in exchanged messages, which can manifest in three distinct ways. First, there's direct causality, where the intention of the previous message directly influences the current one, often indicated by non-ambiguous coreferences like names or pronouns, or through topically-related responses. Second, indirect causality involves messages responding to each other, even when interspersed with contributions from other participants, as long as they are part of the same subsequent bullying event, demonstrating topical relevance and targeting the same participant. Finally, null causality is applied when the analyzed message neither responds to previous messages nor exhibits topical relevance, although it may contribute to escalating bullying without direct association with the subsequent bullying event. These enhancements maintain annotator flexibility, avoiding strict linguistic requirements for each effect. Subsequently, the second round of agreement experiments achieved improved results.

3.2.2. Inter-Annotator Disagreement

Annotators' disagreements in linguistic data have recently gained attention through various initiatives aimed at shedding light on the intricacies of annotating subjective tasks – tasks that may permit diverse valid interpretations of correct data labels (Leonardelli et al., 2021; Plank, 2022). In our work, we draw upon the taxonomy introduced in Sandri et al. (2023), which categorizes potential reasons behind conflicting annotations in subjective tasks. Annotator disagreements have been categorized into four main macro-categories: sloppy annotation, ambiguity, missing information, and subjectivity. Figure 1 illustrates the observed types of disagreements related to the dimensions annotated in our dataset after the IAA phase 2.

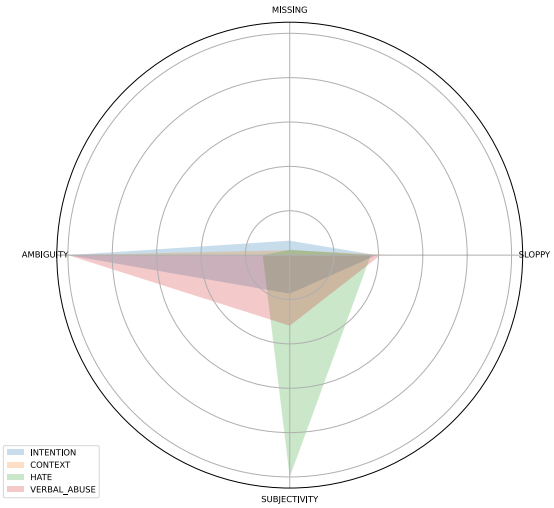

Figure 1: Types of disagreement by annotation layers in *CyberAgressionAdo-v2*

We can observe that certain annotation layers are more prone to specific types of disagreements. For instance, assessing aggression levels often involves subjectivity, while understanding context and intention can introduce ambiguity. Here, subjectivity is closely linked to how annotators perceive a message's overt or covert hatefulness, particularly within sensitive scenarios. Consider the term "gros" in French, meaning "fat" or "overweight." In the context of cyberbullying related to obesity, it can be highly offensive. However, among friends, it may be used casually without harm, similar to "chubby" in English. Refining aggression level definitions with implicitness and subtlety shows that annotator sensitivity plays a key role. The spectrum of aggression isn't always clear-cut, and interpretations vary due to individual perceptions. Concerning ambiguity, it presents an enduring challenge across various annotation lay-

ers, stemming from multiple sources. In the layer dealing with verbal abuse, ambiguity often arises when an instance allows for more than one interpretation, such as an aggression that closely resembles name-calling but may carry additional nuances. In the case of intention and context, annotators must decipher what authors aimed to convey within the ongoing discourse, essentially getting inside the author's mindset. Furthermore, they must establish the context within which a message is situated as responses, a complex task, especially in multi-party settings. These situations demand not just linguistic analysis but also a deep understanding of social dynamics. While missing information is rare in our dataset due to the use of scenarios providing context, instances of ungrammatical and non-standard language may occasionally occur. Concerning sloppy annotation, errors can arise due to annotators' carelessness especially when dealing with multi-layer scheme.

In conclusion, the improvements observed in IAA phase 2 underscore the value of clear guidelines and discussions around challenging situations. 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, integrating information from disagreements into the design of annotation schemes can prove invaluable in mitigating issues like ambiguity and equivocality, ultimately contributing to more precise and effective annotations.

4. Analysis of Cyberbullying Practices

In this section, we present statistical evidence of cyberbullying practices observed in the annotated scenarios. Reported observations are based on frequent patterns at the instance level (i.e., one message) or at the implicature level (i.e., one message and the following reply)

In detail, Table 4 presents the most frequent messages in observed cyberbullying practices, examining individual author utterances (instances). Multiple recurrent cyberbullying practices are observed that align with the roles of involvement. For instance, both bullies and their bystanders tend to target the victims with an intention to deliberately harm (ATK) them. 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 consist of messages in which participants engage in arguments, potentially impacting the conflicts. A clear predominance of aggression is observed in the behavior of bullies and their bystanders, as they utilize both overtly aggressive messages (39.9% of bullies' and 25.6% of bully supporters' posted messages), and covertly aggressive messages (6.5% of bully supporters' utterances). In contrast, victims and their bystanders predominantly use non-aggressive messages either as a part of DFN or OTH intentions. Moreover, conciliators actively strive to resolve conflicts and de-escalate situations, with 30.5% of their messages dedicated to these objectives.

Table 5 provides an overview of cyberbullying practices, emphasizing on utterance pairs (implicatures). Here, "source" refers to the initial message, while "reply" signifies the immediate subsequent message. These pairs embody an implicature relationship as they involve messages generated within the same context, the "reply" often reliant on the previous message for context and meaning. As recurrent patterns observed in all scenarios includes bully supports, it underscores their substantial role of involvement in this context, which align with insights reported in the Educational Research Programme (Xie and yum Ngai, 2020). The presented utterance pairs highlight their collaboration with bullies or other supporters to deliberately harm the victim (ATK). Moreover, they engage with the victims, either as initiators of attacks or participants in arguments (OTH).

Figure 2[9] illustrates the distribution of exchanged messages based on their combined intention (on the left side) and context (on the right side). Here, the "MISC" label is used when the analyzed message is not responsive to previous messages and/or lacks topical relevance. It can encompass messages contributing to the escalation of bullying but not directly related to the subsequent bullying event. From this diagram, we observe that while attacks often occur in response to a defensive posture, they mainly take place in response to an OTH intention, often as part of arguments. Notably, defensive postures are not exclusively reactive to an ATK setting; they are reactive in various scenarios such as OTH, GSL or MISC contexts. Additionally, it's interesting to observe that conflict resolution is primarily carried out as part of the argumentative segment (OTH) rather than as a direct response to ATK or AIN. Other observations reveal that GSL and EMP are employed respectively as means of supporting ATK/AIN and DFN.

Presented patterns in tables consistently occur in all scenarios (support measure of 1.0), providing generalizable and reliable insights that are valuable for studying this complex behavioral phe-

[9] A dynamic version of this diagram is publicly available: https://github.com/aollagnier/CyberAgressionAdo-V2.public/blob/main/sankey_diagram.html

ROLE	HATE	TARGET	INTENTION	FREQ. (%)
bully	OAG	victim	ATK	39.9
bully_support	OAG	victim	ATK	25.6
	CAG	victim	ATK	6.5
conciliator	NAG	-	CR	30.5
victim	NAG	-	OTH	37.9
	OAG	bully	DFN	14.9
	NAG	-	DFN	6.9
victim_support	NAG	-	OTH	41.1

Table 4: Cyberbullying patterns at the instance level (i.e., one message). The percentages relate to the frequency of the given pattern regarding all the messages sent by the corresponding authors overall scenarios.

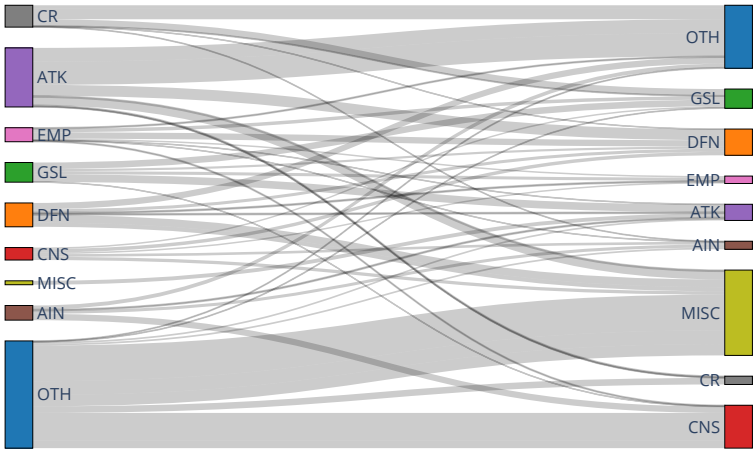

Figure 2: Distribution of intention (left side) and context (right side) aspects in exchanged messages. Colors identify each individual Intention/Context label introduced in Table 2.

(source ↪ reply)	HATE	TARGET	INTENTION
bully	OAG	victim	ATK
↪ bully_support	OAG	victim	ATK
bully_support	OAG	victim	ATK
↪ bully_support	OAG	victim	ATK
bully_support	NAG	-	OTH
↪ victim	NAG	-	OTH
bully_support	OAG	victim	ATK
↪ victim	NAG	-	OTH

Table 5: Cyberbullying patterns at the implicature level (i.e., one message and the following reply).

nomenon. In contrast, all the observed patterns related to both intention and context aspects are utilized for visualization, providing valuable insights from a statistical perspective. Overall, this study supports the notion that cyberbullying is a multi-faceted and complex phenomenon involving intricate communication practices. However, patterns emerge and the use of pragmatic-level information proves beneficial in gaining a better understanding and contrasting the aggressive behaviors observed among bullies, victims, and their respective bystanders. These observations lay the foundation for exploring the complex proactive-reactive schemes involved in such contexts. Additionally, it becomes evident that non-aggressive utterances, which are seldom analyzed in cyberbullying episodes, play a crucial role in the unfolding of events. For instance, victims and their supporters engage in non-aggressive exchanges that should be analyzed, as they contribute to either the escalation or de-escalation of the situation.

5. Conclusion

In this paper, we introduced the *CyberAgressionAdo-V2* dataset, which has been annotated with a hierarchical, fine-grained

tagset designed to describe bullying narrative events in multi-party chat conversations. Presently, the dataset includes 19 conversations in French, simulating online aggression scenarios that can commonly occur among teenagers on private instant messaging platforms. We are actively working on expanding the dataset by including more data from both French and Italian languages. Additionally, we plan to enhance the proposed tagset by incorporating labels that enable the capture of multimodal aspects, thereby revealing valuable pragmatic information.

6. Ethics Statement

NLP research focusing on online aggression and harassment detection inevitably gives rise to ethical considerations. In our work, we place significant emphasis on the importance of considering diversity in resources, the intricate nature of understanding online hate communication, and our support for methods that promote inclusivity.

Firstly, it is important to notice that many studies often use "easy to access" data like Facebook, YouTube, or Instagram. This approach, while convenient, limits the scope of online detection systems to the dynamics and structures specific to these social media channels. As a result, addressing online manifestations occurring on private instant messaging platforms is often overlooked. To address this, it is imperative to develop resources that encompass these platforms and to craft tools and methodologies capable of effectively combating a wide spectrum of harassment and aggression in various online spaces.

Secondly, it's crucial to acknowledge that the interpretation of online hate communication is contingent upon individual viewpoints, personal experiences, and sensitivity levels. Depending on the recipient, what may appear non-aggressive on the surface could still contain elements that are harmful or offensive. Researchers must remain attentive to potential biases, individual perspectives, and the author's intent, as these factors significantly influence the true nature of communication. Furthermore, the annotation process, especially when identifying hate speech, inherently carries subjectivity. Hate speech lacks universal definitions and can vary based on cultural, societal, and contextual factors. Consequently, annotators should be mindful of their own potential biases and preconceptions, while embracing diverse perspectives within the annotation process to enhance accuracy and fairness.

In conclusion, our work advocates for the development of resources that encompass a broader range of languages and communication methods. Additionally, we endorse methodologies that assess the quality of annotation guidelines by analyzing disagreements among annotators. By addressing these ethical concerns and fostering inclusivity, the NLP community can make substantial strides in the development of more effective tools and approaches to curb online hate within diverse online environments.

7. Acknowledgements

This work has been supported by the French government through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002, and by national funding hosted by the French National Research Agency through the EFELIA Côte d'Azur project with the reference number ANR-22-CMAS-0004.

8. Bibliographical References

Medyan AbdelHamid, Assef Jafar, and Yasser Rahal. 2022. Levantine hate speech detection in twitter. *Soc. Netw. Anal. Min.*, 12(1):121.

Dana Aizenkot and Gabriela Kashy-Rosenbaum. 2018. Cyberbullying in whatsapp classmates' groups: Evaluation of an intervention program implemented in israeli elementary and middle schools. *New Media Soc.*, 20(12).

Fabienne Baider. 2020. Pragmatics lost?: Overview, synthesis and proposition in defining online hate speech. *Pragmatics and Society*, 11(2):196–218.

Darina Benikova, Michael Wojatzki, and Torsten Zesch. 2017. What does this imply? examining the impact of implicitness on the perception of hate speech. In *Language Technologies for the Challenges of the Digital Age - 27th International Conference, GSCL 2017, Berlin, Germany, September 13-14, 2017, Proceedings*, volume 10713 of *Lecture Notes in Computer Science*, pages 171–179. Springer.

Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr. Ojha. 2020. Developing a multilingual annotated corpus of misogyny and aggression. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, TRAC@LREC 2020, Marseille, France, May 2020*, pages 158–168. European Language Resources Association (ELRA).

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of hindi-english code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop*

on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, PEOPLES@NAACL-HTL 2018, New Orleans, Louisiana, USA, June 6, 2018, pages 36–41. Association for Computational Linguistics.

Tommy K. H. Chan, Christy M. K. Cheung, and Zach W. Y. Lee. 2021. Cyberbullying on social networking sites: A literature review and future research directions. *Inf. Manag.*, 58(2):103411.

Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth M. Belding. 2018. Peer to peer hate: Hate speech instigators and their targets. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, pages 52–61. AAAI Press.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4):85:1–85:30.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3309–3326. Association for Computational Linguistics.

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, LREC 2022, Marseille, France, 20-25 June 2022*, pages 4149–4161. European Language Resources Association.

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, May 7-12, 2018*. European Language Resources Association (ELRA).

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 10528–10539. Association for Computational Linguistics.

Ricardo Martins, Marco Gomes, José João Almeida, Paulo Novais, and Pedro Rangel Henriques. 2018. Hate speech classification in social media using emotional analysis. In *7th Brazilian Conference on Intelligent Systems, BRACIS 2018, São Paulo, Brazil, October 22-25, 2018*, pages 61–66. IEEE Computer Society.

Ariadna Matamoros-Fernández and Johan Farkas. 2021. Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*, 22(2):205–224.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14867–14875. AAAI Press.

Puneet Mathur, Rajiv Ratn Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting offensive tweets in hindi-english code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media, SocialNLP@ACL 2018, Melbourne, Australia, July 20, 2018*, pages 18–26. Association for Computational Linguistics.

Nicolas Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An in-depth analysis of implicit and subtle hate speech messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1989–2005. Association for Computational Linguistics.

Anaïs Ollagnier, Elena Cabrio, and Serena Villata. 2023a. Harnessing bullying traces to enhance bullying participant role identification in multi-party chats. In *Proceedings of the Thirty-Sixth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2023, Clearwater Beach, FL, USA, May 14-17, 2023*. AAAI Press.

Anaïs Ollagnier, Elena Cabrio, Serena Villata, and Catherine Blaya. 2022. Cyberagressionado-

v1: a dataset of annotated online aggressions in french collected through a role-playing game. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 867–875. European Language Resources Association.

Anaïs Ollagnier, Elena Cabrio, Serena Villata, and Sara Tonelli. 2023b. Birdy: Bullying role detection in multi-party chats. In *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*, pages 16464–16466. AAAI Press.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4674–4683. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 10671–10682. Association for Computational Linguistics.

Fabio Poletto, Valerio Basile, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2019. Annotating hate speech: Three schemes at comparison. In *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, volume 2481 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Lang. Resour. Evaluation*, 55(2):477–523.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2420–2433. Association for Computational Linguistics.

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, ALW@EMNLP 2018, Brussels, Belgium, October 31, 2018*, pages 51–59. Association for Computational Linguistics.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data: Garbage in, garbage out. *CoRR*, abs/2004.01670.

Han Xie and Steven Sek yum Ngai. 2020. Participant roles of peer bystanders in school bullying situations: Evidence from wuhan, china. *Children and Youth Services Review*, 110:104762.

Ziqi Zhang and Lei Luo. 2019. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10(5):925–945.

9. Appendices

9.1. Scenarios

Developed collaboratively with a sociologist and an expert in education sciences, the scenarios aim to address prevalent cyber aggression themes, such as cyberhate related to ethnicity, religion, obesity, and homophobia. Table 6 showcases a selection of scenarios provided to students, drawn from interviews and case studies conducted in French lower and upper secondary schools as part of prior research on adolescent cyberhate (Ollagnier et al., 2022).

Scenario	Type of addressed problem
Julie and Léa use to hang out together during breaks at school holding hands. Emilie, who is jealous of Julie, shares their photo on Snapchat and comments maliciously on their relationship, saying that this situation is suspicious and they are probably homosexual. Marie tries to stand up for Julie and Léa, but Emilie involves her best friends Elodie and Anna, then they try to exclude them from their social group in class and on social networks. Arthur, who is both a friend with Julie and Léa but also Emilie, tries to intervene by explaining to them that it is silly and that it would be better to stop arguing.	Homophobia
Zoe is overweight. After the gym class, Marjorie and Lucie, who are jealous of her good academic results, take a picture of her in a posture that highlights her extra pounds. They share it with the whole class with harmful comments. Natacha, a friend of Zoe, tries to defend her. She is helped by Pauline, who also has a few extra pounds and is a friend of Marjorie. Julien, who was obese when he was younger, tries to intervene with Marjorie and Lucie as well as Zoe to stop the situation.	Obesity
Justine is Jewish. On her profile, she posts a picture of her younger brother's Bar Mitzvah. Léo and Guillaume, Justine's classmates, share the photo with harmful comments against Jews, including caricatures. Aurélie and Isabelle, when they look at the photo, also laugh. Léa and Anna, friends of Justine, try to defend Justine on the chat with the help of Amine to end the harassment against Justine and her religion.	Religion
Fatima is a new student and is very pretty. During a school trip by the sea, she goes swimming with her classmates. Among them are Pauline (jealous of Fatima), Teresa, Julie, and Theo, the best friends of Pauline and Fatima. Pauline takes a picture of Fatima and shares it with the whole class by making fun of her because she has dark skin. Pierre and Nicolas on top of that make harmful comments about Arabic people. Teresa and Julie defend Fatima, and Theo tries to stop the incident.	Ethnicity

Table 6: Samples of scenarios adopted in our experimentation