CyberAgressionAdo-v1: a Dataset of Annotated Online Aggressions in French Collected through a Role-playing Game

Anaïs Ollagnier[1], Elena Cabrio[1], Serena Villata[1], Catherine Blaya[2]
[1]Université Côte d'Azur, Inria, CNRS, I3S
930 route des Colles, BP 145, 06903 Sophia Antipolis Cedex, France
[2]Université Côte d'Azur, CNRS, Unite de Recherche Migrations et Societe (Urmis)
Pôle Universitaire Saint-Jean d'Angely, 24 avenue des Diables Bleus, F-06357 Nice Cedex 4
{anais.ollagnier,elena.cabrio,serena.villata,catherine.blaya}@univ-cotedazur.fr

Abstract

Over the past decades, the number of episodes of cyber aggression occurring online has grown substantially, especially among teens. Most solutions investigated by the NLP community to curb such online abusive behaviors consist of supervised approaches relying on annotated data extracted from social media. However, recent studies have highlighted that private instant messaging platforms are major mediums of cyber aggression among teens. As such interactions remain invisible due to the app privacy policies, very few datasets collecting aggressive conversations are available for the computational analysis of language. In order to overcome this limitation, in this paper we present the *CyberAgressionAdo-V1* dataset, containing aggressive multiparty chats in French collected through a role-playing game in high-schools, and annotated at different layers. We describe the data collection and annotation phases, carried out in the context of a EU and a national research projects, and provide insightful analysis on the different types of aggression and verbal abuse depending on the targeted victims (individuals or communities) emerging from the collected data.

Keywords: hate speech, cyber aggression, role-playing game

1. Introduction

The Web and social media platforms provide an attractive environment to young people to communicate with their peers, to establish social ties, and bring new opportunities for learning (Scheidt, 2015). At the same time, such environment is also raising concerns about the ethical use of technology and expose people to counterproductive and unsafe interactions that set their mental health and well-being at high risk. *Cyber aggressions* are defined by (Grigg, 2010) as "intentional harm delivered by the use of electronic means to a person or a group of people irrespective of their age, who perceive(s) such acts as offensive, derogatory, harmful, or unwante" (p. 152). Aggressive messages may be published both as public posts on social media, but may also be sent in an insidious way on private instant messaging platforms (Sprugnoli et al., 2018). The possibilities offered by social networks to share privately content among users combined with the increasing digital literacy of teenagers has the paradoxical effect to hinder the possibility to study the actual cyberhate activities (Aizenkot and Kashy-Rosenbaum, 2018).

Within the Natural Language Processing (NLP) community, in the past few years there have been several efforts made to deal with the problem of online hate speech detection, leading to the creation of a number of datasets for hate speech detection in different languages, mainly containing messages publicly posted on Twitter (in which the level of interactivity among users is very limited) (Poletto et al., 2021). On the contrary, due to the private nature of the verbal exchanges on private instant messaging platforms (and to the privacy policies that impede ex-post data collection), very few datasets collecting aggressive conversations targeting a specific (group of) victim(s) are available for the computational analysis of language (Sprugnoli et al., 2018). Collecting such kind of data is of primary importance to study the different kind of cyber aggressions - beside cyberhate interactions and offences - that emerge through instant messaging app conversations, with the final goal to fight and prevent digital harassment.

To contribute filling this gap, in this paper we present the *CyberAgressionAdo-v1* dataset[1], containing aggressive multiparty chats in French collected through a role-playing game in high-schools, and annotated at different layers. More specifically, a role-playing game was proposed to students in three French high schools (16–18 years-old), that has been created in collaboration with a sociologist and expert in education sciences and following a setting similar to (Sprugnoli et al., 2018). It relies on scenarios mimicking cyber aggression situations that may occur among teens, involving topics such as the ethnic origin, the religion or the color of skin. The collected conversations have been annotated considering several layers, as the participant roles, the presence of hate speech, the type of verbal abuse present in the message, and whether utterances use different humour figurative devices (e.g., sarcasm or irony).

Such data collection is ongoing, and is carried out in

[1]The dataset is publicly available: `https://github.com/aollagnier/CyberAgressionAdo-v1`

the context of the UCA IDEX OTESIA project "Artificial Intelligence to prevent cyberviolence, cyberbullying and hate speech online" [2]. This work is being completed in the wake of the CHIST-ERA CREEP (Cyberbullying Effects Prevention) project[3].
The article is organised as follows: Section 2 provides some background and a discussion on the definitions of abusive and aggressive content from a sociological perspective. Section 3 introduces related work on existing resources for aggressive content and hate speech detection as tasks for automated systems. Section 4 describes the data collection phase, while Section 5 presents the released version of the CyberAgressionAdo-v1 dataset. Finally, Section 6 discusses some ethical and epistemic issues, while Conclusions end the paper.

NOTE: This paper contains examples of language which may be offensive to some readers. They do not represent the views of the authors.

2. Background

There are several terms and definitions concerning online peer-to-peer aggressions. The one that is most used is "cyberbullying" and this has an impact on the measurement of the prevalence of the observed phenomena. Cyberbullying is generally defined as "an aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time against a victim who cannot easily defend him or herself" (Smith et al., 2008). This definition is mainly based on the initial definition of traditional bullying by (Olweus, 1978). It is also described as "any behavior performed through electronic media by individuals or groups of individuals that repeatedly communicates hostile or aggressive messages intended to inflict harm or discomfort on others" (Tokunaga, 2010; Blaya and Audrin, 2019). However, all the aspects of traditional bullying do not fit to the online environment and there is no reached consensus up to now, which causes discrepancies in findings and approaches (Ersilia et al., 2012). (Corcoran et al., 2015) but also (Blaya, 2013), argue that cyberbullying is too narrow and that it is relevant to turn to use a broader approach, that is cyber aggression. As stressed by (France et al., 2013), cyberbullying can be considered as a form of cyber aggression. This is also the case for cyberhate. The term refers to hate speech that occurs online. Hate speech has been defined as all forms of expression which spread, incite, promote, or justify hatred, discrimination, xenophobia, and other forms of hatred based on intolerance (Council of Europe, 2018). Cyberhate is understood as "electronic communication initiated by hate groups or individuals, with the purpose to attract new members, build and strengthen group identity; it aims at rejecting others' collective identity" (Blaya and Audrin, 2019). On his side, (Bakalis, 2018) more narrowly defines cyberhate as "any use of technology to express hatred toward a person or persons because of a protected characteristic – namely race, religion, gender, sexual orientation, disability and transgender identity" (p. 87). Whilst there is considerable overlap between cyberhate and the concept of cyberbullying, there are some substantial differences between these two forms of online aggression. They are considered to be distinct theoretical phenomena, to have different operationalization, and to warrant being researched as separate phenomena (Blaya and Audrin, 2019; Wachs and Wright, 2019). Cyberhate does not refer to repetitive aggression and it targets the identity of individuals as well as it affects the community they belong to. This type of online aggression can be triggered by social and political events as shown by (Kaakinen et al., 2018). As it is based on inter-group hostility (Hawdon et al., 2015), it contributes to alter social cohesion and consequences run beyond individuals' well-being. Although cyberhate and cyberbullying are different at the conceptual level, they overlap and can be both part of young people's online negative experiences (Bedrosova et al., 2022). This is why in our present research we have included both general cyber aggression and cyberhate. We could not identify cyberbullying due to the temporal characteristics of role plays students were invited to participate to.

3. Related work

As a part of the same project during which we have started collecting the data presented in this work, i.e., the CREEP project, a corpus of data on cyberbullying interactions gathered through a WhatsApp experimentation with lower secondary school students (three classes of students aged 12-13) in Italy has been collected by our project partners[4]. It has been annotated in terms of cyberbullying roles, cyberbullying type of expressions, the presence of sarcasm or not and whether the expression containing insults is not really offensive but a joke (Sprugnoli et al., 2018).
Beside that corpus, the other available datasets are extracted from social media platforms (mainly Twitter, Facebook and Reddit) and are built for the tasks of hate speech and abusive language detection - often considering specific target communities or subtasks. Several resources for many different languages have been released since 2016 (often as benchmark corpora in shared tasks), showing growing interest of the NLP community around abusive language in social media and hate speech detection in particular. (Poletto et al., 2021) present a systematic and up-to-date review of the resources made available by the NLP community

[2] https://www.actuia.com/english/otesia-launches-its-first-4-ai-projects-in-health-prevention-of-cyber-bullying-education-etc/

[3] http://creep-project.eu/

[4] https://dhsite.fbk.eu/2018/09/whatsapp-dataset-on-cyberbullying/

at large. In particular, they compare existing resources according to five dimensions: the type of the resource, the topical focus, the data source, the annotation framework and the language. The authors highlight that a large proportion of hate speech resources developed in the recent past includes data in languages other than English and with several topical focuses.
Another recent survey paper, (i.e., (Vidgen and Derczynski, 2020)), examine 63 publicly available datasets for training abusive content detection systems, providing critical insight into what the datasets contain (and omit), how they have been annotated, and how tasks have been formulated. They report on the creation of Hatespeechdata[5], a catalogue of abusive language data on multiple languages, that could be used to training automated systems. (Jurgens et al., 2019) call for a paradigm shift in the use of NLP technologies to address abusive language. Authors claim that only some phenomena along the spectrum of abusive content are actually addressed, while others are neglected for being either too subtle or quite rare. With our data collection in schools we aim at carrying out one step in this direction, providing richer and more challenging data on aggressive content beside hate speech, to be investigated from a computational point of view.

4. Data Collection

In the following, we report on the data collection phase. We provide details on the experimental setting and on the role-playing game.

4.1. Experimental setting

We have carried out the data collection phase in three French high schools. A total of 142 students aged 16-18 volunteered to participate were involved (90 girls, 52 boys). Our intervention in schools was part of a broader framework for raising awareness to cyber aggression and hate speech, and give students additional means to understand the phenomenon first-hand (see Section 6). The first contact with students was to introduce them to AI and how it could play a role in detecting harmful messages online (1.5 hours). Then students were asked to fill in an anonymous questionnaire proposed by the sociologist involved in the study to collect data on their behavior online (e.g., how much time they spend on the Web, on social media), and on their perception of the cyber aggression phenomena (10-15 minutes). Finally, the researchers presented the experimentation to the participants, conceived as a role-playing game, and relying on the scenarios described in Section 4.2. Each student was provided with a computer to work on, and had to connect to an IRC chat[6] with the name of his/her character in the game, so that to ensure a fully anonymous data collection. Each role-play lasted for 45 minutes. Teachers were present in the room, but they never participated in the chat.
A couple of weeks after the experimentation, a two-hour meeting with the students was organised by the sociologist to discuss with them and their teachers about the topic of cyber aggression and hate speech online to raise awareness on teenagers. During this meeting, students could reflect on the experience, highlight with researchers and teachers the most problematic interactions. Students were finally asked to fill a second questionnaire to highlight the benefits and drawbacks of the experimentation. Considering that young people are actors and experts of their own lives we thought it relevant to consult them in order to avoid misinterpretations or to shut them into our own representations (Alderson and Morrow, 2020).

4.2. Scenarios

Created in collaboration with a sociologist and expert in education sciences, the created scenarios address common cyber aggression topics, including cyberhate related to ethnic origin, religion, obesity and homophobia. Table 1 reports on a few examples of scenarios proposed to students. These scenarios were drawn from interviews and case studies collected in French lower and upper secondary school for a previous research on cyberhate among adolescents (Blaya and Audrin, 2019) and thus, they root in genuine negative experiences reported by young people. We included different types of situations: overweight, religion, ethnicity and, homophobia. These situations were selected on the basis that findings show evidence that overweight (Puhl et al., 2017), and LGBT+ students are more at risk to be discriminated and (cyber)bullied (Bucchianeri et al., 2014) and that cyberhate based on origin and religion is one of the types of victimization that increased most these last decades (Blaya and Audrin, 2019; Llorent et al., 2016; Räsänen et al., 2016) and that processes of exclusion and discrimination grounded on weight are similar to racism, sexism, and gender-oriented bullying (Van Amsterdam et al., 2012). Obese and overweight students are at higher odds to be victimized (Kahle and Peguero, 2017).

4.3. Participant roles

Scenarios were introduced to the students as a role-playing simulation of cyber aggression taking place on a private instant messaging platform. The study of roles involved in online aggressions/bullying is an intensively researched topic. Most of existing researches cast participant roles into "victim", "bully" and "bystander" (Musharraf and Anis-ul Haque, 2018). As a part of the proposed role-playing game, we adapted the categorisation introduced in (Sprugnoli et al., 2018) by assigning 5 types of active roles involved in aggressions:

- `bully`: person who initiates the harassment
- `victim`: person who is harassed

[5] `https://hatespeechdata.com/`

[6] We used IRCCloud for the first data collection, and Kiwi IRC for the second one.

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 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 to 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 1: Samples of scenarios adopted in our experimentation

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

An additional person, i.e., a `moderator`, was reading the chat in a passive way, to ensure that the exchanges among the students were following the rules of the role-playing game (this role was played by one of the researchers present during the data collection). Given that the role-playing game represents a protected space to experiment cyber-violence, we avoided students to impersonate the bullied (i.e., victims were always impersonated by young researchers of our team that were not physically present in the experiment room). In order to involve all the students in the role-playing game, some of the roles were duplicated and embodied in the same scenario. The role of `bully` varies between 1 and 2, the role of `bystander-defender` between 1 and 3 and the `bystander-assistant` between 2 and 3. All the other participants involved in an active way, i.e., the `victim` and the `conciliator` were played only by one person per scenario. In general, each scenario was therefore played by 5 to 7 people.
The students were randomly assigned a scenario and a role (independently of their gender). Only in a couple of cases we were advised by the teachers to avoid assigning a certain role to a student taking into account previous class dynamics and the student's behavior or personal characteristics.

5. Dataset description

In total, we have collected 19 conversations including 4 addressing the problem of homophobia, 7 the problem of obesity, 3 the problem of religion and 6 about ethnicity. Following findings about online hate showing that cyberbullying and discrimination against obesity and ethnic origins had increased most in the previous years (Hawdon et al., 2015; Hawdon et al., 2017; Blaya, 2021), we have decided to increase the number of role-playing games on these topics. In the following, we first describe the annotation process (Section 5.1), and we then present the released version of the *CyberAgressionAdo-v1 dataset* (Section 5.2), reporting on some statistics and discussing its main features.

5.1. Annotation phase

We annotated the collected conversations according to 5 independent annotation layers:

- *the role of the message's author*, corresponding to those introduced in Section 4.3. These role labels are used both to annotate the role of the message's author (i.e., layer (i)) and the role of the target(s) (i.e., layer (ii)).

Message	Author role	Target	Hate speech	Verbal abuse	humor
- ecrit francais et tu viendra me parler culture' mdrr **Translation**: write a proper French and you will speak to me about culture lool	bystander-assistant_1	victim	yes	denigration	yes
- dans ton pays y'a même pas accès à l'eau alors l'ecole... et ça se fait ressentir **Translation**: you can tell you don't have water in your country let alone school ... I can tell	bystander-assistant_2	victim	yes	denigration	no

Table 2: Samples of messages illustrating labelling of the dataset

- *the role of the individual(s) targeted by hate speech and/or verbal abuse*;
- *the presence of hate speech*, as a binary categorization (hate speech or no hate speech). Here, we rely on the definition of hate speech as "content that mocks, insults or discriminates against a person or group based on specific characteristics such as color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics" (Zhang and Luo, 2019).
- *the type of verbal abuse*. Cyber aggression can take many forms. Verbal abuse prevails. It can include harassment, that is repetitive, offensive messages sent to a target, cyber-stalking (sending repetitive threatening communications), flaming that is sending messages with abusive and vulgar terms, such as swearing, gossiping, or mocking, denigration (Bauman, 2014; Tokunaga, 2010; Watts et al., 2017). We have identified 5 types frequently occurring in written language: blaming, name-calling, threat, denigration and other. They are defined as follows:
  - *blaming*: making the victim believe they are responsible for the abusive behavior or that they bring the verbal abuse upon themselves. e.g., "*on la traiterai pa de truie si elle avai pas autant de graisse*" (EN: we wouldn't call her sow if she had less fat)
  - *name-calling*: abusive, derogatory language or insults that chip away at the target's self-esteem, sense of self-worth, and self-concept. e.g., '*té qu1 putain de mongol*' (EN: you are a fucking retarded)
  - *threat*: statements meant to frighten, control, and manipulate the victim into compliance. e.g., '*je vais venir en bas de chez toi tu vas voir qui va plus parler*' (EN: I gonna come at your place and you'll see who gonna not talk anymore)
  - *denigration*: harsh remarks that are meant to make the person feel bad about themselves and are not constructive, but deliberate and hurtful. e.g., "*les fille comme toi ca me dégoute*" (EN: girls like you disgust me)
  - *aggression-other*: content corresponding to other types of abusive, derogatory language or insults deliberately harmful not fitting in the other categories. e.g., "*va crevé en enfer*" (EN: go die in hell)
- *whether messages contain some forms of humor*. We do not distinguish the different kinds of humorous devices such as irony, wordplay, metaphor and sarcasm. The annotation relies on a binary categorisation, namely, humor and non-humorous. In this context, it is a tough task to consider content as humorous due to its will to deliberately harm targeted person(s). Here, we consider content as humorous whether they include laughing emojis or interjections such as "ptdr" (EN: laughing my ass off) or "lol". Content initiating in return such humorous markers are also identified as humorous.

hate speech	verbal abuse	humor
98.4%	91.5%	96.3%

Table 3: Results of Inter Annotator Agreement

Table 2 presents a sample of the annotated data. The annotation guidelines have been discussed with the sociologist involved in the data collection, and a first pilot annotation round has been made to agree on some unclear cases. Then, the dataset has been fully annotated by an expert annotator, with background in computational linguistics. A second annotator with background in computational linguistics has annotated four conversations (i.e., debates over four scenarios), to calculate the inter annotator agreement. On average, these scenarios are composed of 172.7 messages and each of them corresponds to one of the cyberhate topics introduced in Section 4.2 (i.e., ethnicity, religion, homo-

phobia and obesity). Results are shown in Table 3 in terms of Cohen's Kappa coefficient. Since that both the role of authors and targets was pre-defined, we did not measure the agreement on the assignment of these labels. Results are satisfactory given that the agreement is above 0.9 for the hate speech, verbal abuse and humor layers. These high scores show that the definitions chosen to describe each annotation layer characterise precisely the different phenomena/aspects we wished to capture and identify, proving the reliability of the dataset for computational purposes.

5.2. The CyberAgressionAdo-v1 dataset

Table 4 reports statistics on the different annotation layers performed on the full dataset.

As we can observe, the `bystander-assistants` and the `bully` are the roles putting forward most of the messages in the scenarios with respectively 27.5% and 22.3% of the exchanged messages, followed by the `victim` role (21.0% of the messages) and the `bystander-defender` role (19.3% of the interactions). The roles of `conciliator` is the less active representing 9.8% of the total of exchanged messages. `Victim` is the role mainly targeted by either hate speech and verbal abuse representing more than the half of the reported cyber aggression behaviors (52.6%). Figure 1 reports the proportion of interactions between participants showing the direction of the exchanges, i.e. from the writer to the target (V : B means from `victim` to `bully`). We can observe a greater asymmetry in the proportion of interactions between both bullies and their assistants toward the victims. Conversely, the whole exchanges between `bystander-defender` and both the bullies and their assistants is balanced. This Figure contributes to better understand participant roles of involvement in online aggression situations. Here, the victims are clearly overwhelmed by bullies/bystander assistants. In contrast, bystander defenders react accordingly to aggression perpetrated against themselves or the victims.

We identified a total of 1210 messages corresponding to the different types of verbal abuse, meaning that the phenomenon we are investigating covers 37.0% of the whole dataset. Verbal abuse of type *Denigration* occurs the most representing 38.0% of the total of verbal abuse encountered, followed by *Aggression-other* and *Name-calling*, covering respectively 32.9% and 18.7%. Concerning the use of hate speech, it constitutes 7.5% of the whole dataset. In detail, scenarios about homophobia report the highest percentage of hate speech (11.9% considering the whole scenarios involving this topic), followed by those about ethnicity in which it represents 11.1%, and religion (6.2%). Concerning the obesity scenarios, we report 2.8% of hateful comments.

In the whole dataset, messages using humorous devices constitute 4.6%. Scenarios about obesity report the highest presence of humor with 5.9%, followed by religion (5.3%). In ethnicity and homophobia scenarios, it constitutes respectively 4.1% and 2%.

Figure 2 presents the proportion of verbal abuse targeting both `victim` and `bystander-defender` across topics. According to the topic we can observe different victimization practices relying on the use of specific or combinations of types of verbal abuse. For instance, aggression against individual(s) based on their ethnicity or their weight rely mainly on the use of *Denigration* (respectively, 41% and 53%). Concerning the conversations involving aggression against individual(s) based on their religion or their sexual orientation, there is proportionate distribution between *Denigration* and *Aggression-other* as well as *Name-calling* in homophobia situations. The proportion of *Threat* is also significant in some scenarios, especially those about ethnicity and obesity. In detail, `bully` and `bystander-assistant` mainly target `victim` using *Denigration* and `bystander-defender` using *Aggression-other*. Denigration against victims consists of using harsh remarks related to the topic of the given scenario, e.g. obesity: "espece de boule jte pousse tu roule" (EN: you're a ball I'll push you around). Concerning aggression against `bystander-defender`, they correspond to abusive, derogatory language or insults in response to their support to the victim, e.g. "tg" (EN: shut up) or "toi creve" (EN: you die). On the other hand, `victim` and `bystander-defender` use mainly *Denigration* and *Name-calling* against `bully` such as "t tro con" (EN: you're stupid), "tu pues la merde" (EN: you smell like shit). Concerning `bystander-assistant`, victims tend to use the same types of verbal abuse than the ones used against bullies, while `bystander-defender` use mainly *Aggression-other* (e.g., "va deterrer tes mort" (EN: go dig up your dead) or "ferme ta gueule toi meme" (EN: shut the fuck up yourself)). Concerning the use of hate speech, it occurs mostly combined with *Denigration* and *Name-calling* utterances, e.g. denigration: "moi jalouse ? de ta peau mi jaune mi marron mdrr" (EN: me jealous? of your half yellow half brown skin lool), name-calling: "SALE MIGRANTS DE MERDE" (EN: FUCKING MIGRANTS).

6. Ethical and epistemic issues

Involving students in research about sensitive topics such as cyber aggression and cyberhate raises ethical implications. All students under the age of 18 were asked to participate providing they submitted a parental consent form. Parents were sent an explanation letter informing them on the objectives of the research, the use and management of data and the associated potential risks. All students participating in the role-playing game were informed about the objectives of the study, the potential benefits of using AI to detect online hostile messages and the relevance of basing our research on their participation. Before starting, they had an introduction on cyber aggression as well as a general pre-

ROLES	
Bystander assistant	897 (27.5%)
Bully	727 (22.3%)
Victim	685 (21.0%)
Bystander defender	630 (19.3%)
Conciliator	321 (9.8%)

VERBAL ABUSE	
Denigration	461 (38.0%)
Aggression-other	399 (32.9%)
Name-calling	227 (18.7%)
Threat	100 (8.2%)
Blaming	23 (1.9%)
Total	**1210**

TARGETS	
Victim	600 (52.6%)
Bully	185 (16.2%)
Bystander assistant	120 (10.5%)
Bystander defender	116 (10.1%)
Bully/Bystander assistant	54 (4.7%)
Conciliator	42 (3.6%)
Victim/Bystander defender	19 (1.6%)
Bully/Bystander assistant/Victim/Bystander defender	2 (0.1%)

humor	
humor	4.6%

HATE SPEECH	
hate speech	7.5%

Table 4: Statistics on the different annotation layers performed on the *CyberAgressionAdo-v1* dataset

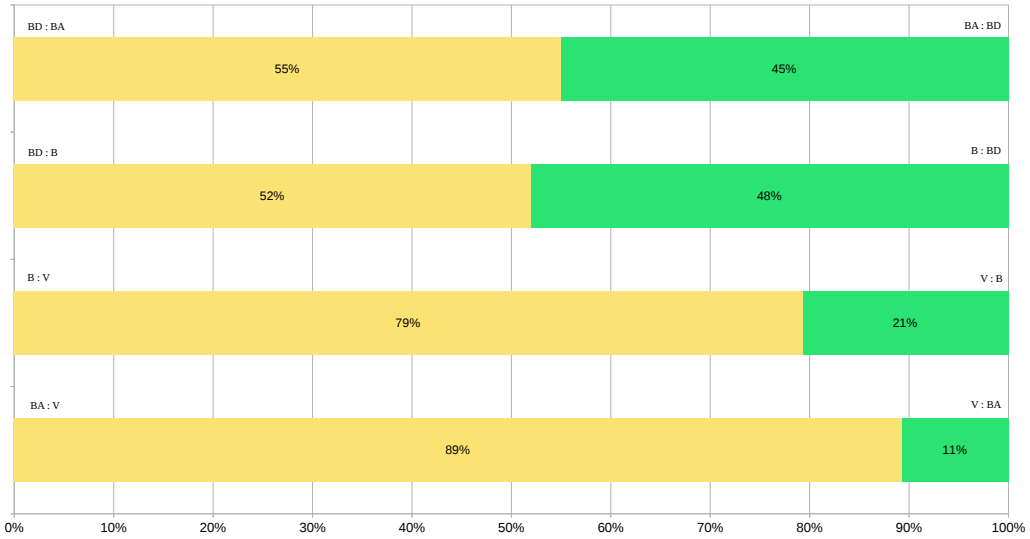

Figure 1: Proportion of exchanged messages between participants. As for participant names, "V" = `victim`, "BD" = `bystander-defender`, "BA" = `bystander-assistant`, "B" = `bully`.

sentation of AI and its potential uses. As contributing to research should be linked to appropriate benefit, following data analyses, students and school staff were invited to participate to a feedback session introducing the findings in terms of prevalence of aggressive and hateful messages. This presentation was followed by a training session on cyber aggression, cyberbullying and cyberhate, and their consequences both at the victims' and perpetrators' level. No student was asked to impersonate the victim and the whole process was performed under the supervision of the researchers who could provide help if needed and observe the students' reactions. The researchers met the European ethical requirements as well as their university's. The schools involved in the data collection have a psychologist in charge of catering for students in need and that in case of problematic situation due to the survey the participant could have been referred to her/him. The administration board of each school examined the study request and protocol and gave their assent. Students were also provided some information about the agencies or resource adults able to provide assistance in case of trouble. Autonomy, confidentiality and anonymity were respected throughout the whole process. The observation of the performance leads us to the conclusion that although some students expressed surprise about the aggressivity of some comments, no one left the classroom feeling upset. Some students stayed behind to chat about the experience in a free and relaxed way. Although we cannot certify the realisticness of the collected interactions in the wake of (Sprugnoli et al., 2018), several factors lead us to be confident with the validity of the process: the proposed scenario were based on previous real experience reported by young people of the same age; the spontaneity of multiparty chats does not allow much time to extrapolate and the role-play data are

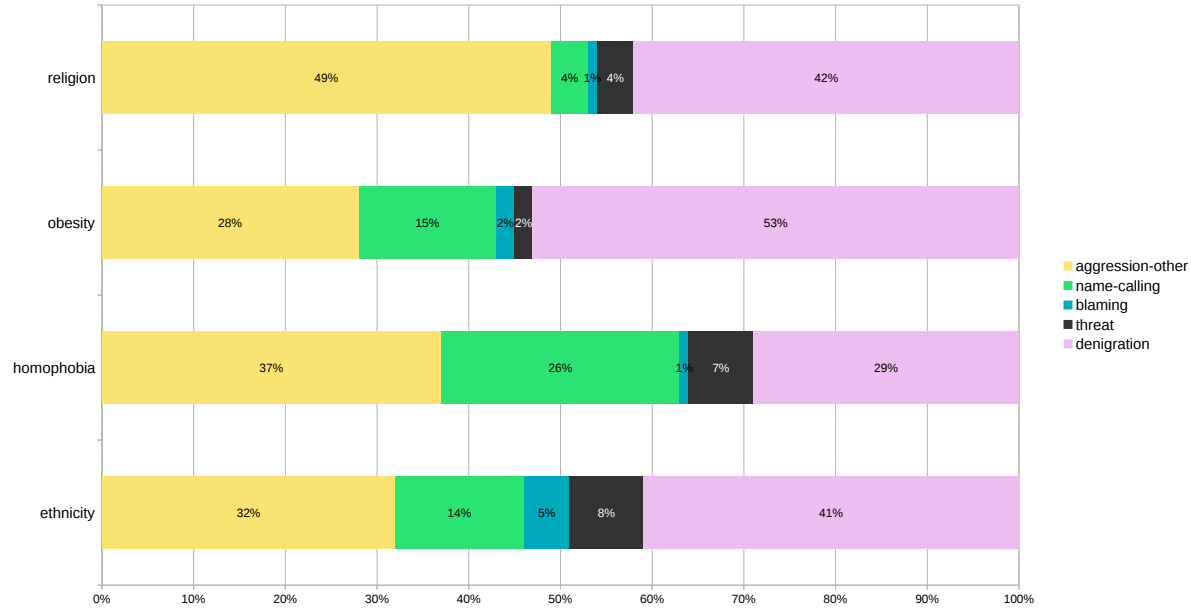

Figure 2: Distribution of verbal abuse targeting `victim` and `bystander-defender` across topics.

similar to the naturalistic interactions. This last consideration meets findings from previous studies according to which role-plays are a more valid measure of authentic language use than more traditional ways of collecting data such as interviews or self-report questionnaires (Kasper, 1999; Tran, 2006).

7. Conclusion

In this paper we have presented the data collection and annotation of the *CyberAgressionAdo-v1* dataset.

To sum up our analysis on the released dataset, we have observed frequent/common patterns in cyber aggression practices emerging from this work. This study highlights the existing links between the use of certain aggression/bullying methods and different aspects (the topic on which the victim is targeted or who is involved in the interactions). Indeed, we have noticed a propensity in using hate speech to target individual(s) based on their sexual orientation while aggression against overweight people are spread using humorous devices. Concerning religion both hate speech and humor are used equally. Interactions involving both victim-bully are characterized by the use of denigration and name-calling utterances. Furthermore, aggression-other is the most frequent type of verbal abuse used by bystanders from both sides. These outcomes open multiple research directions for research in NLP leveraging the potential of automating the identification of aggression/bullying methods in order to better understand and curb online cyber aggression among teens (especially in the context of private instant messaging platforms in which cyber aggressions occur the most).

Our data collection is still ongoing, and additional sessions in French high-schools are scheduled in the next months, allowing us to increase the size (and variety) of the released dataset. As players involved in the role-playing game are free to make choices that impact the direction of their group's story, it is important to run more scenarios and collect additional data in schools, to cover the different types of aggression and verbal abuse according to the targeted victims (individuals or communities) in real-world settings.

8. 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 Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002.

9. Bibliographical References

Aizenkot, D. and Kashy-Rosenbaum, G. (2018). Cyberbullying in whatsapp classmates' groups: Evaluation of an intervention program implemented in israeli elementary and middle schools. *New Media & Society*, 20(12):4709–4727.

Alderson, P. and Morrow, V. (2020). *The ethics of research with children and young people: A practical handbook*. Sage.

Bakalis, C. (2018). Rethinking cyberhate laws. *Information & Communications Technology Law*, 27(1).

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

Bedrosova, M., Machackova, H., Šerek, J., Smahel, D., and Blaya, C. (2022). The relation between the

cyberhate and cyberbullying experiences of adolescents in the czech republic, poland, and slovakia. *Computers in Human Behavior*, 126:107013.

Blaya, C. and Audrin, C. (2019). Toward an understanding of the characteristics of secondary school cyberhate perpetrators. *Frontiers in Education*, 4:46.

Blaya, C. (2013). *Les ados dans le cyberespace: prises de risque et cyberviolence*. De Boeck Sup.

Blaya, C. (2021). Bias bullying- problems among school children: Prevalence, and intersectional considerations. *Handbook of Bullying, Characteristics, risks and outcomes*, 1.

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

Corcoran, L., Guckin, C. M., and Prentice, G. (2015). Cyberbullying or cyber aggression?: A review of existing definitions of cyber-based peer-to-peer aggression. *Societies*, 5(2):245–255.

Council of Europe, R. (2018). Hate speech. `https://www.coe.int/en/web/freedom-expression/hate-speech`.

Ersilia, M., Calussi, P., and Nocentini, A. (2012). Cyberbullying and traditional bullying: Unique, additive, and synergistic effects on psychological health symptoms. *Wiley Blackwell*, pages 245—-262.

France, K., Danesh, A., and Jirard, S. (2013). Informing aggression–prevention efforts by comparing perpetrators of brief vs. extended cyber aggression. *Computers in Human Behavior*, 29(6):2143–2149.

Grigg, D. W. (2010). Cyber-aggression: Definition and concept of cyberbullying. *Journal of Psychologists and Counsellors in Schools*, 20(2):143–156.

Hawdon, J., Oksanen, A., and Räsänen, P. (2015). Online extremism and online hate. *NORDICOM*.

Hawdon, J., Oksanen, A., and Räsänen, P. (2017). Exposure to online hate in four nations: A cross-national consideration. *Deviant Behavior*, 38(3).

Jurgens, D., Hemphill, L., and Chandrasekharan, E. (2019). A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666. ACL.

Kaakinen, M., Oksanen, A., and Räsänen, P. (2018). Did the risk of exposure to online hate increase after the november 2015 paris attacks? a group relations approach. *Computers in Human Behavior*, 78.

Kahle, L. and Peguero, A. A. (2017). Bodies and bullying: The interaction of gender, race, ethnicity, weight, and inequality with school victimization. *Victims & Offenders*, 12(2):323–345.

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

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

Musharraf, S. and Anis-ul Haque, M. (2018). Cyberbullying in different participant roles: exploring differences in psychopathology and well-being in university students. *Pakistan journal of medical research*, 57(1).

Olweus, D. (1978). *Aggression in the schools: Bullies and whipping boys.* Hemisphere.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Lang. Resour. Evaluation*, 55(2):477–523.

Puhl, R. M., Wall, M. M., Chen, C., Austin, S. B., Eisenberg, M. E., and Neumark-Sztainer, D. (2017). Experiences of weight teasing in adolescence and weight-related outcomes in adulthood: A 15-year longitudinal study. *Preventive medicine*, 100.

Räsänen, P., Hawdon, J., Holkeri, E., Keipi, T., Näsi, M., and Oksanen, A. (2016). Targets of online hate: Examining determinants of victimization among young finnish facebook users. *Violence and victims*, 31(4):708–725.

Scheidt, L. A. (2015). It's complicated: The social lives of networked teens. *New Media Soc.*, 17(2).

Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., and Tippett, N. (2008). Cyberbullying: Its nature and impact in secondary school pupils. *Journal of child psychology and psychiatry*, 49(4).

Sprugnoli, R., Menini, S., Tonelli, S., Oncini, F., and Piras, E. (2018). Creating a whatsapp dataset to study pre-teen cyberbullying. In Darja Fiser, et al., editors, *Proceedings of the 2nd Workshop on Abusive Language Online, ALW@EMNLP*. ACL.

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

Tran, G. Q. (2006). The naturalized role-play: An innovative methodology in cross-cultural and interlanguage pragmatics research. 5(2):1–24.

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

Vidgen, B. and Derczynski, L. (2020). Directions in abusive language training data: Garbage in, garbage out. *PLoS ONE 15(12): e0243300*, abs/2004.01670.

Wachs, S. and Wright, M. F. (2019). The moderation of online disinhibition and sex on the relationship between online hate victimization and perpetration. *Cyberpsychology, Behavior, and Social Netw.*, 22(5).

Watts, L. K., Wagner, J., Velasquez, B., and Behrens, P. I. (2017). Cyberbullying in higher education: A literature review. *Comp. in Human Behavior*, 69.

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