Creating a WhatsApp Dataset to Study Pre-teen Cyberbullying

Rachele Sprugnoli[1], Stefano Menini[1], Sara Tonelli[1], Filippo Oncini[1,2], Enrico Maria Piras[1,3]
[1]Fondazione Bruno Kessler, Via Sommarive 18, Trento, Italy
[2]Department of Sociology and Social Research, University of Trento, Via Giuseppe Verdi 26, Trento
[3]School of Medicine and Surgery, University of Verona, Piazzale L. A. Scuro 10, Verona, Italy
{sprugnoli;menini;satonelli;piras}@fbk.eu
filippo.oncini@unitn.it

Abstract

Although WhatsApp is used by teenagers as one major channel of cyberbullying, such interactions remain invisible due to the app privacy policies that do not allow ex-post data collection. Indeed, most of the information on these phenomena rely on surveys regarding self-reported data.

In order to overcome this limitation, we describe in this paper the activities that led to the creation of a WhatsApp dataset to study cyberbullying among Italian students aged 12-13. We present not only the collected chats with annotations about user role and type of offense, but also the living lab created in a collaboration between researchers and schools to monitor and analyse cyberbullying. Finally, we discuss some open issues, dealing with ethical, operational and epistemic aspects.

1 Introduction

Due to the profound changes in ICT technologies over the last decades, teenagers communication has been subjected to a major shift. According to the last report by the Italian Statistical Institute (ISTAT, 2014) in Italy 82.6% of children aged 11-17 use the mobile phone every day and 56.9% access the web on a daily basis. Despite being of fundamental importance for teenagers' social life, the use of these new technologies paved the way to undesirable side effects, among which the digitalisation of traditional forms of harassment. We refer to these form of harassment as "cyberbullying". Should we adopt a narrow definition, cyberbullying refers only to actions repeated over time with the aim to hurt someone. By definition, cyberbullying is in fact '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). In everyday life, however, the notion of cyberbullying indicates each episode of online activity aimed at offending, menacing, harassing or stalking another person. The latter definition of cyberbullying is the one currently adopted in most of the research conducted on this phenomenon and in the present work. The phenomenon has been recognised as a ubiquitous public health issue, as the literature clearly highlights negative consequences for teenagers: studies show that victims are more likely to suffer from psychosocial difficulties, affective disorders and lower school performance (Tokunaga, 2010). Since the difference between cyberbullying and traditional forms of bullying lies in the intentional use of electronic forms of contact against the designed victim, data on verbal harassment and cyberbullying stances are particularly useful to analyse the phenomenon. However, due to the private nature of these verbal exchanges, very few datasets are available for the computational analysis of language. It should be noted that the possibilities offered by social networking platform to share privately content among users combined with the increasing digital literacy of teenagers has the paradoxical effect to hinder the possibility to scrutinize and study the actual cyberbullying activities. For instance, although WhatsApp is used by teenagers as one major channel of cyberbullying (Aizenkot and Kashy-Rosenbaum, 2018), their interactions remain invisible due to the privacy policies that impede ex-post data collection. Most of the information on these phenomena, though, relies on surveys regarding self-reported data. Yet, specifically for this reason, the possibility to study how cyberbullying interactions and offenses emerge through instant messaging app conversations is crucial to fight and prevent digital harassment.

In this light, this paper presents an innovative corpus of data on cyberbullying interaction gath-

Proceedings of the Second Workshop on Abusive Language Online (ALW2), pages 51–59
Brussels, Belgium, October 31, 2018.

ered through a WhatsApp experimentation with lower secondary school students. After outlining the CREEP project[1] (CybeRbullying EffEcts Prevention) during which the activities were carried out, and the living lab approach that led to the creation of the corpus, we present the provisional results of the computational analysis.

In the discussion, we address the main ethical concerns raised by the experimentation and we discuss the implications of such a methodology as a tool for both research and cyberbullying prevention programs. Annotated data, in a stand-off XML format, and annotation guidelines are available online[2].

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

2 Related Work

In this Section we provide an overview of datasets created to study cyberbullying, highlighting the differences with respect to our Whatsapp corpus.

The sources used to build cyberbullying datasets are several and cover many different websites and social networks. However, the main differences are not in the source of the data but in the granularity and detail of the annotations. Reynolds et al. (2011) propose a dataset of questions and answers from Formspring.me, a website with a high amount of cyberbullying content. It consists of 12,851 posts annotated for the presence of cyberbullying and severity. Another resource developed by Bayzick et al. (2011) consists of conversation transcripts (thread-style) extracted from MySpace.com. The conversations are divided into groups of 10 posts and annotated for presence and typology of cyberbullying (e.g. denigration or harassment, flaming, trolling). Other studies cover more popular social networks such as Instagram and Twitter. For example, the dataset collected from Instagram by Hosseinmardi et al. (2015), consists of 2,218 media sessions (groups of 15+ messages associated to a media such as a video or a photo), with a single annotation for each media session indicating the presence of cyberagressive behavior. The work by Sui (2015) is instead on Twitter data, presenting a corpus of 7,321 tweets manually annotated for the presence of cyberbullying. In this dataset, other than cyberbullying, the authors provide different layers of associated information, for example the role of the writer, the typology of attack and the emotion. The aforementioned works have two main differences with respect to the data presented in this paper. First, each annotation refers to an entire message (or a group of messages) and not to specific expressions and portions of text. Second, the categories used for the annotation are more generic than the ones adopted in our guidelines.

A more detailed investigation on the cyberbullying dynamics can be found in Van Hee et al. (2015b,a) presenting the work carried out within the project AMiCA[3], which aims at monitoring web resources and automatically tracing harmful content in Dutch. The authors propose a detailed analysis of cyberbullying on a dataset of 85,485 posts collected from Ask.fm and manually annotated following a fine-grained annotation scheme that covers roles, typology of harassment and level of harmfulness. This scheme has been recently applied also to English data (Van Hee et al., 2018). This approach is different from the previous ones since its annotations are not necessarily related to entire messages but can be limited to shorter strings (e.g., single words or short sequences of words). We found this approach the more suitable for our data, allowing us to obtain a detailed view of pre-teens use of offensive language and of the strategies involved in a group with ongoing cyberbullying. The details on how we use the guidelines from the AMiCA project (Van Hee et al., 2015c) as the starting point to define our annotation scheme can be found in Section 5.

Another important difference between our corpus and the datasets previously discussed is the source of our data. Indeed, among the many available instant messaging and social media platforms, WhatsApp is not much investigated because it is private, thus an explicit donation from the chat participants is required to collect data (Verheijen and Stoop, 2016). Although in the last few years WhatsApp corpora have been released in several languages including Italian (see for example the multilingual corpus built in the context of the "What's up, Switzerland?" project[4] (Ue-

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

[2] http://dh.fbk.eu/technologies/whatsapp-dataset

[3] https://www.lt3.ugent.be/projects/amica

[4] https://www.whatsup-switzerland.ch/index.php/en/

berwasser and Stark, 2017), these corpora have been collected with the goal to investigate specific linguistic traits such as code-switching, or to study whether the informal language used in the chats affects the writing skills of students (Dorantes et al., 2018; Verheijen and Spooren, 2017). With respect to the aforementioned works focused on the analysis of linguistic phenomena in WhatsApp, our aim is to study a social phenomenon, that is cyberbullying, by analysing how it is linguistically encoded in WhatsApp messages. Even if the relation between cyberbullying and WhatsApp is strong, no annotated corpus of WhatsApp chats has been released so far. Therefore, our corpus represents a novel resource, useful to study in particular cyberbullying in classmates' groups (Aizenkot and Kashy-Rosenbaum, 2018).

3 Project Description

The corpus presented in this paper is part of the CybeRbullying EffEcts Prevention activities (CREEP), a larger project based in Italy and supported by EIT Digital[5], for the monitoring of cyberbullying and the assistance of students and teachers. The project goals are *i)* to develop advanced technologies for the early detection of cyberbullying stances through the monitoring of social media and *ii)* to create a virtual coaching system for communicating preventive advice and personalised messages to adolescents at risk of cyberbullying aggressions. To this purpose, one of the means to reach the objectives is the realisation of a living lab, namely an in vivo research methodology aimed at co-creating innovation through the involvement of aware users in a real-life setting (Dell'Era and Landoni, 2014). This approach permits to tackle the cyberbullying phenomenon within the social media platform from a user-centric perspective, thereby favouring the co-creation of prevention strategies whilst avoiding top-down planning. When applied to the school context, the living-lab approach presents a threefold advantage. Researchers gain ecological validity by studying the phenomenon with target users through a role-play experiment. Teenagers can actively engage in the participatory design without being used as passive research subjects. Schools can count on a supplementary tool for raising awareness on cyberbullying and give pupils additional means to understand the phenomenon first-hand.

The creation of the corpus required the involvement of students in a role-playing simulation of cyberbullying and has so far involved three lower secondary schools' classes of teenagers aged 12-13 from 2 different schools based in Trento, in the North-East of Italy. The experimentation itself, described in the next section, was embedded in a larger process that required four to five meetings, one per week, involving every time two social scientists, two computational linguists and at least two teachers for the class. The content of the meetings is briefly described below:

1. **Lecturing on cyberbullying**. Researchers introduce the theme of cyberbullying and elicit personal experiences and students' opinions.

2. **Pilot annotation of online interactions**. Students in pairs annotate semantically six online threads gathered from Instagram and Twitter. They discuss the hate speech categories that will be then annotated also in the WhatsApp interactions.

3. **Introducing the experimentation**. Researchers present the experimentation, both the practicalities and the rules of the role-playing.

4. **Participatory analysis**. Researchers present the preliminary analysis of the experimentation and elicit students interpretations of the data gathered.

5. **Feedback to teachers and parents** (optional but recommended). Researchers present the whole project to parents of students involved and teachers, discussing results and future activities.

The relevance and implications of this approach are presented in Section 6.

4 WhatsApp Experimentation

The choice of WhatsApp is based on two main considerations: first, the application is one of the most preferred and used messaging applications (Fiadino et al., 2014). In 2015, almost 60% of Italian teens aged 12-17 used the app (Save the Children, 2014). In fact, in each class involved in the living lab, around 70% of the children already had an active WhatsApp account on their

[5] https://www.eitdigital.eu/

Scenario	Type of addressed problem
Your shy male classmate has a great passion for classical dance. Usually he does not talk much, but today he has decided to invite the class to watch him for his ballet show.	Gendered division of sport practices
Your classmate is very good at school, but does not have many friends, due to his/her haughty and 'teacher's pet' attitude. Few days ago, s/he realised that his/her classmates brought cigarettes to school and snitched on them with the teacher. Now they will be met with a three days suspension, and they risk to fail the year.	Interference in others' businesses
Your classmate is very good at school, and everyone think s/he is an overachiever. S/He studies a lot and s/he never goes out. S/He does not speak much with his/her classmates, that from time to time tease him/her for his/her unsocial life. Things have slightly changed recently: your classmates mum convinced teachers to increase the homework for all the students. A heedless teacher revealed the request to the class, and now some students are very angry at him/her.	Lack of independence, parental intromission.
Your shy classmate is good in all subjects but in gymnastics. For this reason, his/her classmates often tease on him/her when s/he exercises. Recently, the class has found out a video on the social network Musical.ly, where s/he dances gracelessly, on a 90s song that no one has never heard before.	Web virality

Table 1: Scenarios adopted in our experimentation.

personal smartphones. Only a minority (around 5 or 6 teens) used their parents' smartphones to be able to participate. Second, the app provides all the functionalities of social networking services, and Whatsapp classmates' groups are identified by other studies as contexts of cyberbullying perpetration (Aizenkot and Kashy-Rosenbaum, 2018). Overall, a total number of 70 students participated in the experimentation.

After receiving the necessary authorisation from the school director, the school board, and teenagers' parents, the researchers presented the experimentation to the participants, conceived as a role-play. In each class, two WhatsApp groups with around 10 teens were created. Teachers were part of the groups and could assist to the conversation, but they never actively participated to the chat. In each group, one researcher played for the whole time the role of the victim. Students were instead given the following roles: cyberbully (2 students), cyberbully assistants (3-4 students), and victim assistants (3-4 students). Teachers divided the classes and assigned the roles to pupils, so to take into account previous class dynamics and childrens personalities. Each role-play lasted for 3 days, after which pupils changed roles within the same chat; students were allowed to participate in the chat only after the school hours, and could be excluded for a short time in case of misbehaviour. Teachers and researchers used dedicated mobile phones provided by the project, not their private ones. Each chat started with the following basic rules:

- Offenses must target only the designated victim (the researcher)
- Bad words are allowed, but do not exaggerate
- Do your best to play your role and try to interact in a realistic way
- Stick to the roles previously defined
- Do not hesitate to quit the chat if you feel offended
- Use the chat only after school hours

Moreover, in order to trigger the conversation, several scenarios, previously discussed and agreed

by the students, opened the role-play. The scenarios, reported in Table 1, aim to address different types of problematics that teenagers can encounter, from the gendered division of sport practices to the embarrassment caused by a viral video. At the end of the experimentation, a two-hour meeting with each class was organised to reflect with students and teachers on the content of the conversation, and to discuss some of the taken-for-granted aspects of cyberbullying and raise awareness on teenagers. During this occasion, students could reflect on the experience, highlight with researchers and teachers the most problematic interactions, and point out the benefits and drawbacks of the methodology.

Three middle-school classes were involved in this experimentation. Since WhatsApp groups are closed and not accessible from the outside, a preliminary agreement was signed involving students' parents, teachers and headmasters to allow the activity. The threads were then saved in anonymous form and manually annotated by two expert linguists. The original names were not completely removed, but they were replaced by fictitious names, so that it was still possible to track all the messages exchanged by the same person.

5 Corpus Description

The corpus of Whatsapp chats is made of 14,600 tokens divided in 10 chats. All the chats have been annotated by two annotators using the CAT web-based tool (Bartalesi Lenzi et al., 2012) following the same guidelines.

Our guidelines are an adaptation to Italian of the "Guidelines for the Fine-Grained Analysis of Cyberbullying" developed for English by the Language and Translation Technology Team of Ghent University (Van Hee et al., 2015c). Following these guidelines, the annotator should identify all the harmful expressions in a conversation and, for each of it, he/she should annotate: (i) the cyberbullying role of the message's author; (ii) the cyberbullying type of the expression; (iii) the presence of sarcasm in the expression; (iv) whether the expression containing insults is not really offensive but a joke. The guidelines identifies four cyberbullying roles: `Harasser` (person who initiates the harassment), `Victim` (person who is harassed), `Bystander-defender` (person who helps the victim and discourages the harasser), `Bystander-assistant` (person who takes part in the actions of the harasser). As for the type of cyberbullying expressions, we distinguish between different classes of insults, discrimination, sexual talk and aggressive statements: `Threat or blackmail`, `General Insult`, `Body Shame`, `Sexism`, `Racism`, `Curse or Exclusion`, `Insult Attacking Relatives`, `Harmless Sexual Talk`, `Defamation`, `Sexual Harassment`, `Defense`, `Encouragement to the Harassment`, and `Other`. Each message in the chat can contain more than one expression to be annotated with a different associated type thus making the annotation fine-grained. For example the message *fai schifo, ciccione!* / *you suck, fat guy* is made of two harmful expressions e.g. [*fai schifo,*]$_{\text{General\_Insult}}$ [*ciccione!*]$_{\text{Body Shame}}$.

With respect to the original guidelines by Van Hee et al. (2015c), we added a new type of insult called `Body Shame` to cover expressions that criticize someone based on the shape, size, or appearance of his/her body. We did this addition because body shaming has become an important societal issue that according to existing literature has a strong impact on the cybervictimization of teens and pre-treens (Frisén et al., 2014; Berne et al., 2014). We have also changed the original type `Encouragement to the Harasser` into `Encouragement to the Harassment`, so to include all the incitements between the bully and his/her assistants. We indeed noticed that the exhortations to continue the persecution and the expressions of approval for insults and acts of intimidation do not have a single direction (that is, from the assistants to the bully) but all the people taking part in the harassment encourage each other.

We calculated the inter annotator agreement between our annotators on one of the chats, made of 1,000 tokens and belonging to the scenario about the video posted on musical.ly. Results are shown in Table 2 in terms of Dice coefficient (Dice, 1945) for the extension of the annotated expression and in terms of accuracy for the attributes associated to each message. Since that the roles where pre-defined, we did not measure the agreement on the assignment of the cyberbullying roles. Results are satisfactory given that the agreement is equal or above 0.8 both for the extension and the attributes. These scores are

Extension		Attributes		
Dice coefficient		Accuracy		
exact match	partial match	type	sarcasm	non-offensive
0. 80	0.88	0.87	1	1

Table 2: Results of Inter Annotator Agreement

TYPES	
Defense	381 (31.7%)
General_Insult	313 (26.0%)
Curse_or_Exclusion	200 (16.6%)
Threat_or_Blackmail	81 (6.7%)
Encouragement_to_the_Harassment	63 (5.2%)
Body Shame	45 (3.7%)
Discrimination-Sexism	45 (3.7%)
Attacking_relatives	28 (2.3%)
Other	24 (2%)
Defamation	23 (1.9%)
TOTAL	**1203**

ROLES	
Bystander_assistant	358 (29.8%)
Harasser	343 (28.5%)
Bystander_defender	334 (27.7%)
Victim	168 (14.0%)

OFFENSIVE	
non-offensive	0

SARCASM	
sarcasm	27 (2.2%)

Table 3: Annotated data on WhatsApp data

higher than those reported for the annotation made using the guidelines to which we were inspired (Van Hee et al., 2015b). This difference could be explained by the fact that our annotators were directly involved in the creation of our guidelines.

Table 3 reports statistics about the annotated data in the WhatsApp chats. We identified a total of 1,203 cyberbullying expressions, corresponding to almost 6,000 tokens: this means that the phenomenon we are investigating covers 41.1% of the whole corpus. Roles are quite balanced in the chats with many interactions among all the participants. Expressions of type `Defense` written by the victim or bystander defenders are the most numerous (31.7%): they include expressions in support of the victim specifying his/her positive characteristics (e.g., *secondo me è bravo! / I think he's good!*) but also showing disapproval and indignation (e.g., *lasciatelo stare! / leave him alone!*). Undesirable sexual talk (type `Sexual_harassment`) and expressions of discrimination that are based on the victim's race, skin color, ethnicity, nationality, or religion are never found in the WhatsApp corpus. Besides, all rude remarks and bad words are offensive in the context in which they are used (i.e., the `non-offensive` tag is never used). The defenders of the victim often fight back responding to attacks with insults (e.g., *le sfigate siete voi / you are the losers*) and threats (*e.g. Se non la smetti la vedrai con tutti noi / If you do not stop you will see it with all of us*). These types of expressions correspond to the 37.1% of the annotations with the role `Bystander-defender`. Almost all the expressions in the category `Curse_or_Exclusion` (96%) are aimed at detaching the counterpart from social relations with expressions such as *chiudi il becco / shut up*, *nessuno ti vuole / nobody wants you*, *cambia classe / change class*. The strong majority of insults of type `Attacking_relatives`, corresponding to 82.1%, are addressed to the mother (e.g., *Tua madre fa schifo quanto te / Your mother sucks as much as you do*) whereas the others are attacks to sisters, brothers and friends in general. Different scenarios bring out different types of cyberbullying expressions and thus different types of insults. For example, the scenario about the ballet has a high presence of expressions with a sexist nature, starting from the idea that ballet is an activity only for girls, e.g. *Balli anche te così da gay? / Do you dance so gay too?*). The scenario related to the video on musical.ly has attracted many comments on the appearance of the victim and, as a consequence, we have high occurrence of insults of type `Body_Shame`. Typically, these insults contain references to animals stressing the heaviness of the victim, for example *Dimagrisci elefante / lose weight, you elephant*, *Sembra un bisonte quando corre! / He*

looks like a bison when he runs!.

6 Discussion

The use of a simulation to create the corpus described in the previous sections raises a number of issues. A detailed and thorough exploration of such issues is beyond the scope of the present paper and it would require a paper in its own right. Nonetheless, in this last section we discuss some of the issues faced with no pretense of exhaustiveness, distinguishing between operational, ethical and epistemic issues. As we shall see, though, there are considerable overlaps among these categories.

6.1 Operational Issues

First and foremost, the creation of a WhatsApp corpus has required several accompanying measures. The effort required to gather the corpus is only a fraction of the overall effort needed to set up the living lab. Participation of schools was ensured by making data gathering only a part of a larger set of activities aimed at paving the way for experimentation (lecture, annotation) and providing teachers and parents, our internal stakeholders, with results to be further used for educational purposes (feedback). All these activities required to invest extra-time. At the same time, though, the prolonged involvement and exchanges with students and teachers allowed us to build a trust meant to ensure a smooth participation to the role-playing and avoid that researchers are perceived as judgmental.

A second issue, partially related to the former, is the need of a consistent engagement of researchers in the creation of the corpus. The experimentation, far from having a predictable behaviour, needed to be monitored to ensure that participants would adhere to the rules (i.e. not insulting each other) or stop interacting. This required making several decisions (e.g. sending private messages to remind the rules to specific students) to avoid the failure of the role-playing. Borderline cases were frequent and each required choices to be made on the fly. As a rule, researchers adopted a flexible approach towards rule-breaking, deeming that an excessive intervention would have broken the 'suspension of disbelief' of the role-playing.

On a separate note, the unfolding of the research process required to make some decisions regarding conflicting needs. For instance, the students without a smartphone were excluded from the role-playing experimentation. We evaluated the possibility to provide a smartphone to these students to avoid their exclusion from the activities of the rest of the classmates. The option was ruled out preferring realism over participation.

6.2 Ethical Issues

The ethical issues were a main concern since the drafting of the study design. The study has been co-designed with the schools involved and parents' informed consent was gathered beforehand. The role-playing methodology was adopted, among other considerations, as it did not require gathering sensitive information regarding minors. Students were assigned a role and they were asked to play it considering a fictional (even if realistic) scenario, so no information regarding lived experiences was collected.

Beside these formal considerations, ethics has been a central concern of the research team throughout the process and it was addressed by-design as far as possible. As briefly mentioned in the Project description (see Section 3), the role-playing activity was part of a larger set up. The purpose was to ensure a framing of the experimentation and its possible outcomes in a broader perspective, allowing students to play different roles. Therefore, the introductory lecture in class addressed the issue of cyberbullying adopting a distal perspective (cyberbullying as a topic). The annotation phase required students to be exposed to the raw material of cyberbullying in all its unpleasantness but filtering their perception by adopting the perspective of researchers (cyberbullying as an object of study). The role-playing was performed after each student had already been familiarized with cyberbullying. The role-playing allowed a protected space to experiment cyberbullying, avoiding students to impersonate the bullied (victims were always impersonated by researchers) and experiencing different roles (cyberbullying as a lived experience from multiple perspectives: bully, support to the bully, support to the victim). The participatory analysis allowed students to retrospectively frame and provide meaning to the lived experience in a protected environment (cyberbullying as a prop for reflectivity). The whole process was designed to allow the honest and realistic outburst during the role-playing but framing it with accompanying measures aimed at

avoiding/minimizing the negative effects of harsh interactions during the role-playing. The presence and vigilance of two researchers and a teacher in each phase of the process ensured a failsafe mechanism to prevent the derangement of the experiment and its containment.

6.3 Epistemic Issues

At last, we shall address the issue of validity of the corpus created with the experimentation. Can the corpus be considered a substitute for a set of actual interactions or should it be considered just as the result of a playful experience with little (if any) resemblance to reality? The lack of a WhatsApp corpus of non-simulated cyberbullying activities does not allow for a comparison to assess the plausibility of the cyberbullying interactions gathered. Evaluating the realisticness of such interactions is both the main challenge and the key to replicate and extend the methodology adopted to other domains. While a clear-cut methodology to perform such evaluation was not developed for the case at hand, we tentatively assessed the plausibility of the corpus indirectly.
Since we could not assess the verisimilitude of the outcome (i.e. corpus) we collected information to evaluate the credibility of the process. We focused on three dimensions:

- observer effect, to estimate the self-censorship implied in being watched by adults;
- evolution of the interaction, to estimate the resemblance to actual online harassing;
- engagement in the interaction, to understand the constant awareness of being involved in a simulated activity.

The classes did not seem too worried about being under observation. Both classes had already experienced cyberbullying issues in the past years and teachers were informed by the students about it. In those occasions, to the surprise of the teachers, students voluntarily showed them the text to request their help, despite the presence of vulgar content. Moreover, during the experimentation, researchers had to remind several students to behave themselves and remember the basic rules of roleplaying. As for the evolution of the interaction, students declared that the experimentation mirrored a 'normal' heated conversation on WhatsApp, with violent but short-lived outbursts regarding a single issue. About the engagement in the interaction, the participatory analysis revealed that some participants were forgetful of the text they sent, as if they got carried away by the role playing. While we are aware that a only a thorough methodology could respond to the question of the validity of the corpus gathered through the role playing, these preliminary findings suggest that the spirit of the game may lead participants to act with in a way that resembles real life.

7 Conclusions

In this work, we present and release a WhatsApp dataset in Italian created through a role-play by three classes of students aged 12-13. The data, which are freely available, have been anonymized and annotated according to user role and type of insult. Given the difficulty to retrieve WhatsApp data, since their chats are only accessible to the group members, we believe that datasets of this kind can give a better insight into the language used by pre-teens and teens in closed communities and into the dynamics of cyberbullying. The work has also highlighted the importance of creating a living lab with a setting suitable for experimentation and educational activities. Nevertheless, several open issues must be taken into account, from ethical issues related to exposing pre-teens to offensive language, to the problem of data realisticness when using a role-play environment.

Acknowledgments

Part of this work was funded by the CREEP project (http://creep-project.eu/), a Digital Wellbeing Activity supported by EIT Digital in 2018. This research was also supported by the HATEMETER project (http://hatemeter.eu/) within the EU Rights, Equality and Citizenship Programme 2014-2020. In addition, the authors want to thank all the students and teachers who participated in the experimentation.

References

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 & Society*, page 1461444818782702.

Valentina Bartalesi Lenzi, Giovanni Moretti, and Rachele Sprugnoli. 2012. Cat: the celct annotation tool. In *In Proceedings of LREC 2012*, pages 333–338.

Jennifer Bayzick, April Kontostathis, and Lynne Edwards. 2011. Detecting the presence of cyberbullying using computer software. In *3rd Annual ACM Web Science Conference (WebSci 11)*, pages 1–2.

Sofia Berne, Ann Frisén, and Johanna Kling. 2014. Appearance-related cyberbullying: A qualitative investigation of characteristics, content, reasons, and effects. *Body image*, 11(4):527–533.

Claudio Dell'Era and Paolo Landoni. 2014. Living lab: A methodology between user-centred design and participatory design. *Creativity and Innovation Management*, 23(2):137–154.

Lee R Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Alejandro Dorantes, Gerardo Sierra, Tlauhlia Yamín Donohue Pérez, Gemma Bel-Enguix, and Mónica Jasso Rosales. 2018. Sociolinguistic corpus of whatsapp chats in spanish among college students. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 1–6.

Pierdomenico Fiadino, Mirko Schiavone, and Pedro Casas. 2014. Vivisecting Whatsapp Through Large-scale Measurements in Mobile Networks. *SIGCOMM Comput. Commun. Rev.*, 44(4):133–134.

Ann Frisén, Sofia Berne, and Carolina Lunde. 2014. Cybervictimization and body esteem: Experiences of swedish children and adolescents. *European Journal of Developmental Psychology*, 11(3):331–343.

Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishr. 2015. Prediction of cyberbullying incidents on the Instagram social network. *arXiv preprint arXiv:1508.06257*.

ISTAT. 2014. Il Bullismo in Italia: Comportamenti offensivi tra i giovanissimi. https://www.istat.it/it/files/2015/12/Bullismo.pdf. [Online; accessed 26-July-2018].

Kelly Reynolds, April Kontostathis, and Lynne Edwards. 2011. Using machine learning to detect cyberbullying. In *Machine learning and applications and workshops (ICMLA), 2011 10th International Conference on*, volume 2, pages 241–244. IEEE.

Save the Children. 2014. I nativi digitali. Conoscono davvero il loro ambiente? . https://www.savethechildren.it/blog-notizie/i-minori-e-internet-italia-infografica. [Online; accessed 25-July-2018].

Peter K. Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. 2008. Cyberbullying: its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry*, 49(4):376–385.

Junming Sui. 2015. *Understanding and fighting bullying with machine learning*. Ph.D. thesis, Ph. D. dissertation, The Univ. of Wisconsin-Madison, WI, USA.

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

Simone Ueberwasser and Elisabeth Stark. 2017. Whats up, switzerland? a corpus-based research project in a multilingual country. *Linguistik online*, 84(5).

Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2018. Automatic detection of cyberbullying in social media text. *arXiv preprint arXiv:1801.05617*.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2015a. Automatic detection and prevention of cyberbullying. In *International Conference on Human and Social Analytics (HUSO 2015)*, pages 13–18. IARIA.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2015b. Detection and fine-grained classification of cyberbullying events. In *International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 672–680.

Cynthia Van Hee, Ben Verhoeven, Els Lefever, Guy De Pauw, Véronique Hoste, and Walter Daelemans. 2015c. Guidelines for the fine-grained analysis of cyberbullying. Technical report, Language and Translation Technology Team, Ghent University.

AJP Verheijen and WPMS Spooren. 2017. The impact of whatsapp on dutch youths school writing. In *Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities (cmccorpora17)*. Bolzano: Eurac Research.

Lieke Verheijen and Wessel Stoop. 2016. Collecting facebook posts and whatsapp chats. In *International Conference on Text, Speech, and Dialogue*, pages 249–258. Springer.