Proceedings of the 9th Games and Natural Language Processing Workshop @LREC2022, pages 1–6
Marseille, 25 June 2022

An Analysis of Abusive Language Data Collected through a Game with a Purpose

Federico Bonetti[1,2], **Sara Tonelli**[1]
[1]Fondazione Bruno Kessler, Trento, Italy, [2]University of Trento, Italy
{fbonetti, satonelli}@fbk.eu

Abstract

In this work we present an analysis of abusive language annotations collected through a 3D video game. With this approach, we are able to involve in the annotation teenagers, i.e. typical targets of cyberbullying, whose data are usually not available for research purposes. Using the game in the framework of educational activities to empower teenagers against online abuse we are able to obtain insights into how teenagers communicate, and what kind of messages they consider more offensive. While players produced interesting annotations and the distributions of classes between players and experts are similar, we obtained a significant number of mismatching judgements between experts and players.

Keywords: game with a purpose, linguistic annotation, offensive language

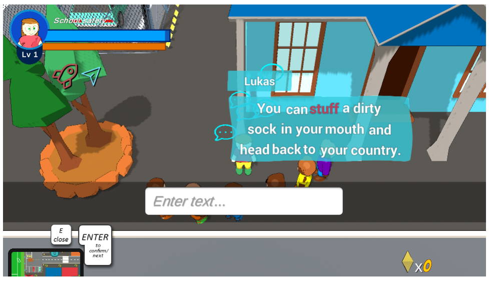
1. Introduction

Cyberbullying has been recognised as a major public health issue, which can lead to severe negative consequences for teenagers, from self-harm to suicide (Tokunaga, 2010; Kowalski et al., 2014). Nevertheless, cyberbullying attacks are frequent in private chats and channels, while only a small fraction of them is visible in public accounts. This makes it hard to study the behaviour of adolescents online, since data collection from major social media platforms is strictly limited. The few existing works dealing with NLP and cyberbullying resort to simulations (Sprugnoli et al., 2018; Menini et al., 2020), create datasets starting from school bulletin boards (Nitta et al., 2013) or extract posts from the few available online sources like ask.fm (Hee et al., 2015; Safi Samghabadi et al., 2020; Rathnayake et al., 2020), where however users are anonymous and it is not possible to identify teenagers among them.

Collecting reliable data, while respecting teenagers' privacy, is therefore of paramount importance to study cyberbullying phenomena. Novel ways to understand the behaviour of teenagers with respect to verbal abuse online are needed. Past works have proposed to use video games to empower teenagers in countering cyberbullying and increase their resilience (Calvo-Morata et al., 2019). In this work we employ *High School Superhero* (HSS) (Bonetti and Tonelli, 2021a) as a tool to involve teenagers, i.e. typical targets of cyberbullying, in a game where the main goal is to decrease the amount of offensive language used in a small town. The players have the possibility to critically evaluate potentially offensive sentences and make them not offensive. As a side effect, the game allows the collection of a large number of sentences judged by teenagers in the form of a gamified crowd-sourced task. Thus, playing with HSS can also lead to the creation of linguistic annotated datasets for abusive language detection. We focus this contribution on the analysis of the annotated data and the challenges of using HSS to collect abusive language annotations.

2. Related work

In NLP, several games with a purpose (GWAPs) have been proposed in the past to address different linguistic tasks: *Phrase Detectives* (Poesio et al., 2013) for anaphora resolution; *OnToGalaxy* (Krause et al., 2010) for semantic linking; *The Knowledge Towers* and *Infection* for validating and extending ontologies (Vannella et al., 2014); *Puzzle Racer* and *KaBoom!* (Jurgens and Navigli, 2014) for sense-image mapping and word sense disambiguation; *WordClicker* (Madge et al., 2019) for Part-of-Speech tagging; *Zombilingo* (Fort et al., 2014) for dependency syntax annotation, and *Wordrobe* (Venhuizen et al., 2013) for word sense labeling. Concerning the use of gamification to raise awareness against cyberbullying, past works showed that increasing empathy is crucial to controlling cyberbullying (Barreda-Ángeles et al., 2021; Del Rey et al., 2016) and games can help in this sense as shown by (Calvo-Morata et al., 2019). They tested *Conectado*, a game where users take the perspective of bullied victims, with school teachers and students aged from 12 to 17. The authors showed that this change of perspective has a positive impact on awareness and empathy, since players can learn more about bullying and what consequences it can have. (DeSmet et al., 2018), on the other hand, stress the importance of promoting positive bystander behavior. In particular, they found that after playing their serious game, participants reported an increase in self-efficacy to end cyberbullying and intention to act as a positive bystander. Using High School Superhero in classes aims to pursue both goals: on the one hand, it should empower teenagers by making them more aware of the language used in online conversations and of their offensive potential. On the

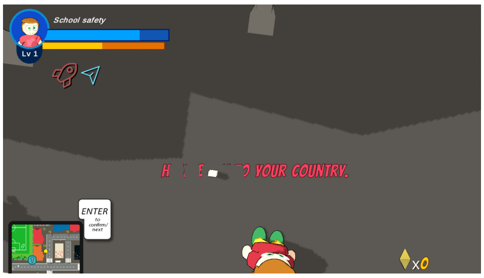


Lukas

You can stuff a dirty sock in your mouth and head back to your country.

Enter text...

Figure 1: Task mechanic 1 (overhearing dialogues)



H ... YOUR COUNTRY.

Figure 2: Task mechanic 2: Erasing graffiti

other hand, it allows the collection of sentences annotated as offensive or not. In this work, we focus in particular on the second aspect.

3. Design of High School Superhero

In this Section we summarise the main features of High School Superhero (HSS), the 3D game we have used to collect annotations about abusive language.
HSS is a 3D role-playing game set in a small town that allows players to change or erase parts of sentences displayed in different ways. After a character creation screen, players can explore a town to perform the task in dedicated spots. The theme and setting are relevant to the target domain of cyberbullying (Ahmad and Law, 2021). In fact, the very act of explaining to the players who they are within the fictional world (a student specifically chosen to fix the language spoken near and inside a school) and what their goal is in ethical terms (reducing the influence of bullies to save the students) may already foster on its own an appropriate sympathetic response (Belman and Flanagan, 2010; Ryan and Staines, 2016).

3.1. Task mechanics

The game contains 2 different types of activities, so-called *task mechanics* (Bonetti and Tonelli, 2021b). In Task Mechanic 1 (Figure 1), players can listen to conversations happening among non-player characters and see a preview of what they are going to say. In particular, when the player goes near a certain group of students, it is possible to overhear their conversation. Before every message, the player is able to read the speaker's mind: a cloud is shown where tokens are freely modifiable. Whenever a change is made, the student in the group says what the player has told them to say, then they act surprised and look puzzled. Both the modified sentence and the original sentence are kept in order to have examples of abusive sentences and possible fixes. The new sentence can be similar to the original one or rewritten from scratch, since the focus is on knowing if, not how, the sentences have been modified.
In Task Mechanic 2 (Figure 2), players erase graffiti off the ground or walls. Players are instructed to remove graffiti that contain abusive language. Players can erase tokens by using a sponge and a consumable called 'soap'. Words are considered erased when more than 80% of the word surface has been wiped. It is possible to go back and cancel the erasing if needed. This allows to make a new annotation, using additional soap, but it does not grant additional points, otherwise players would be able to spam annotations on the same graffiti to gain points.

3.2. Side mechanics

Side mechanics, in particular mechanics that do not contribute directly to the execution of annotation tasks, are also present. These include:

Collectible elements: Crystals are an in-game currency that can be spent to acquire both power-ups and task-related resources. Collectibles, such as coins, diamonds and the like, have been found to be quite effective in increasing the player's engagement and time spent in video games (Naglé et al., 2021).

Navigation power-ups: The Rocket Boots, a special pair of shoes, allow users to jump as high as some rooftops. An electric scooter allows users to move faster around the town. Lastly, the Glider allows players to jump off buildings and gently glide to the ground.

Quests, a hallmark of role-playing games, are also present. They have been implemented in the form of rather simple missions, where random characters ask the player to erase some graffiti in the area before the time is up.

4. Activity and Data Description

4.1. Activity Setup

The game was administered to selected students in the context of a project aimed at raising awareness on cyberbullying and online abuses targeting teenagers. We carried out in total 6 focus group sessions in 6 Italian middle and high schools. The procedure was approved by the Ethics Advisory Board of the project and of the authors' institution. Before the activities, the participants' parents signed a consent form. Also the participants gave their consent and, before using the game, were reminded that they could quit the activity in any moment.
The procedure was carried out in complete anonymity. Prior to playing, participants were briefed on the activity. They were briefly shown the game and told that it was about abusive language detection: sentences that

they deemed abusive should be corrected (annotated) by erasing words in the case of graffiti and by erasing *or* changing words in the case of dialogue lines. They were also told that they could decide to change or erase only a part of the sentence or no tokens at all, leaving the text unchanged if no offense was detected. This is important as participants may be eager to play and try out the mechanics regardless of the content of messages.

4.2. Data Selection

Since our goal is to analyse the quality of the offensive language dataset collected through the game, we carefully select the sentences to be displayed to the players. We rely on the dataset presented in (Sprugnoli et al., 2018), which contains simulations of cyberbullying interactions collected in classes through a Whatsapp chat. The sentences, in Italian, have been also manually labeled as abusive or not and associated with a category label such as *Body shaming*, *Threat or blackmail*, *Racism*, *Sexism*, *Curse or Exclusion*, *Generic offense*. Using this dataset allows us to compare the judgments collected through the game with the gold labels previously assigned by linguists during manual annotation.

Sentences were divided into different sets according to the target group. Indeed, in some classes we had to be careful not to administer certain types of sentences that could have some people re-experience distressful situations, therefore we followed teachers' suggestions on how to select the data. Sentences with explicit sexual content were always omitted. In general, students from the same class were shown the same sentences, and each class could potentially annotate up to 300 sentences.

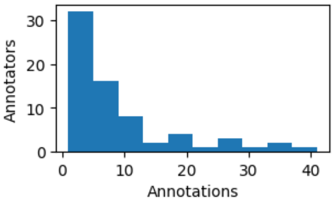


Figure 3: Distribution of annotations per annotator

5. Data Analysis

5.1. Annotation distribution

In total, 590 annotations were collected on 199 sentences from 70 players. The mean number of annotations per participant was 8.42 (SD=9.22); the median was 5; the mode was 2. 50% of participants contributed between 1 and 7 annotations while the top annotator provided as many as 41 annotations. See Figure 3 for a distribution of annotations.

We focus our analysis on the set of annotated sentences for which it is possible to obtain a majority vote, or that were annotated only once. These are overall 162 sentences.

		Players		Tot.
Expert		O	N	
	O	79	34	113
	N	26	23	49
	Tot.	105	57	162

Table 1: Expert judgements vs majority judgements (O=*Offensive*, N=*Not offensive*).

We report in Table 1) the distribution of the collected annotations. We compare them with the annotations from the original dataset, assigned by linguists. Overall, we observe a slight increase of offensive annotations in the dataset by experts. Furthermore, the two sets of annotations match only partially, in particular only 23 sentences were considered not offensive both by experts and players. Expert annotations in this study are shown mainly with the purpose of understanding the degree of mismatch and to observe patterns that differ among the offensive categories. Given the differences between interfaces (only teenagers used the gamified one) and the subjectivity of the task, agreement is not used as an annotation quality metric. For a detailed analysis of mismatches see the following Sections.

5.2. Experts vs. Players' Annotations

We display in Figure 4 the detailed distribution of the labels assigned by Experts (left), compared with the distribution of the labels in the dataset annotated by Players (right). The diagram refers to the 162 sentences analysed in Table 1.

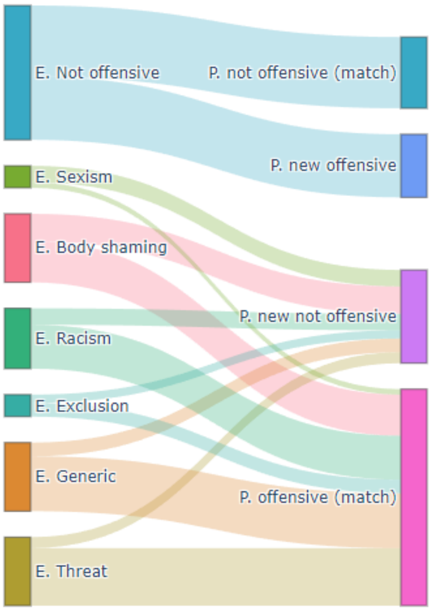


Figure 4: Distribution of the categories in the annotations by Experts (E) (left) and Players (P) (right). Players' annotations marked with *(match)* have the same offensive/not offensive label as in the expert dataset

The figure shows that most of the sentences referring to specific offensive categories labeled by experts have been recognised as offensive also by players. An interesting exception is the *Sexist* category, whose sentences have been mostly considered not offensive in the game, highlighting the need to raise awareness on misogynistic and sexist language among teenagers. A similar trend exists for the *Body shaming* category, and interestingly during the focus groups students often referred to this category as one of the most important insults to tackle. Also sentences in the *Exclusion* category, which encompasses cases of direct attacks aimed at detaching the counterpart from social relations such as "shut up" and "go away", have been considered not offensive in some cases.

In general, players tended to tag sentences as *not offensive* more frequently than linguists. The mismatching sentences could be due to differences in perception between the two groups of annotators or to considering sentences in a dialogue context (see Section 5.3). It is likely that linguists, who originally annotated the dialogue turns, focused more on the single utterances without considering much the thread context. This would confirm also the findings in (Menini et al., 2021), showing that sentences are less likely to be labeled as offensive when annotators consider the discourse context. However, it should be noted that *P. new offensive* sentences could be caused also by a certain eagerness to try out the game and its mechanics, while *P. new not offensive* sentences could be caused by not paying attention or accidental skipping. To partially solve this problem, however, we let players skip dialogue sentences only after 1 second from the onset.

5.3. Qualitative Analysis

One legitimate concern is that, when changing sentences, players could write something even more offensive just to have fun. However, students seem to have gone by the guidelines. For example, *You butterball you're really fat, yesterday I saw you on the guinness world records as the fattest person alive* ('Palla di lardo sei proprio un ciccione ieri ti ho visto sul guinness world records per il più grasso al mondo') was changed to *You **fork**ball you're really **beautiful** yesterday I saw you on the guinness world records as the **most beautiful** person alive*. Although 'fork' does not really make sense, it still does not make the sentence offensive, and therefore it does not go against the purpose of keeping pairs of negative and positive/neutral examples. Regarding similar examples with other sentences, in sentence *You don't know? Haha, what a loser*, 'loser' was changed to '**good person**'.

Concerning the graffiti, sentence *Indeed, you horrid nerd* ('Appunto, secchiona orribile') was changed to *Indeed you ~~horrid~~ nerd* by one participant, to *Indeed you ~~horrid nerd~~* by another and lastly it was erased completely by another still. It looks like participants preferred to erase the whole sentence rather than offensive words or random words. For example, *It was always your fault!* ('È sempre stata colpa tua!') was erased completely by 3 users. One changed it to *It was always ~~your~~ fault*. The reference to the victim was erased, which is acceptable in the context of neutralizing offenses.

Regarding sentences originally labeled as *Not offensive* that were annotated as *Offensive* by players, consider this sentence, which is in the *P. new offensive* category: *At least I have intelligence* ('Almeno ho l'intelligenza'). This may not be offensive in the sense that it is not overtly offensive per se. It could imply two different things: that the person who utters the sentence has many flaws except stupidity; or that the counterpart is not intelligent. It is possible that players interpreted it according to its most hateful meaning, also because of the context of the dialogue and the focus group activity, where they could have acted like they were being tested on their readiness to spot offensive language. Distributing the game 'in the wild', with a written tutorial modified according to the first feedback described in this paper, may yield different results.

Interestingly, in *It's true he did not cause the team to lose, he caused it to be disqualified* ('È vero non ha fatto perdere la squadra, la ha fatta squalificare'), 'be disqualified' was changed to '**qualify**' by one participant and to '**win**' by another. These annotations are particularly worth examining, since the sentence is not overtly offensive, as it does not contain any specific insult; however, it may imply that whoever caused the team to be disqualified deserves hate or contempt. Through HSS it seems possible to retrieve judgements that come from reasoning about the background of a given utterance, given that a certain number of sentences in a sequence refer to the same topic or situation.

6. Conclusion

In this work we have presented an analysis of the annotations collected through the 3D game with a purpose "High School Superhero" on a cyberbullying dataset. The game was deployed in the context of focus groups held with 6 Italian classes of students. We gathered in total 590 annotations from 70 participants.

We observed considerable mismatch between annotations by linguists and those collected through the game. This might be due to differences in the perception of the offenses by the two different groups of annotators or to behaviour caused by the game, such as accidental skipping (which however regarded the dialogues alone, and which was curbed by a quality control step) or eagerness to change the sentences. We plan to counter this last problem in the future by making annotation of non-offensive sentences more rewarding. The findings of this paper need however to be confirmed by further research, one limitation being the small sample size. Another aspect worth exploring in the future is a qualitative analysis of players' behaviour based on post-hoc questionnaires. This would shed more light on why annotators annotated as they did.

7. Acknowledgements

We thank all participants that played with the game to annotate the dataset. Part of this work has been funded by the KID_ACTIONS REC-AG project (n. 101005518) on "Kick-off preventIng and responDing to children and AdolesCenT cyberbullyIng through innovative mOnitoring and educatioNal technologieS".

8. Bibliographical References

Ahmad, A. and Law, E. L.-C. (2021). Educators as Gamemasters: Creating Serious Role Playing Game with. *Proceedings of the ACM on Human-Computer Interaction*, 5(CHI PLAY):1–29, October.

Barreda-Ángeles, M., Serra-Blasco, M., Trepat, E., Pereda-Baños, A., Pàmias, M., Palao, D., Goldberg, X., and Cardoner, N. (2021). Development and experimental validation of a dataset of 360°-videos for facilitating school-based bullying prevention programs. *Computers & Education*, 161:104065.

Belman, J. and Flanagan, M. (2010). Designing Games to Foster Empathy. 14(2):11.

Bonetti, F. and Tonelli, S. (2021a). Challenges in designing games with a purpose for abusive language annotation. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 60–65, Online, April. Association for Computational Linguistics.

Bonetti, F. and Tonelli, S. (2021b). Measuring orthogonal mechanics in linguistic annotation games. *Proc. ACM Hum. Comput. Interact.*, 5(CHI):1–16.

Calvo-Morata, A., Freire-Moran, M., Martinez-Ortiz, I., and Fernandez-Manjon, B. (2019). Applicability of a Cyberbullying Videogame as a Teacher Tool: Comparing Teachers and Educational Sciences Students. *IEEE Access*, 7:55841–55850.

Del Rey, R., Lazuras, L., Casas, J. A., Barkoukis, V., Ortega-Ruiz, R., and Tsorbatzoudis, H. (2016). Does empathy predict (cyber) bullying perpetration, and how do age, gender and nationality affect this relationship? *Learning and Individual Differences*, 45:275–281, January.

DeSmet, A., Bastiaensens, S., Cleemput, K. V., Poels, K., Vandebosch, H., Deboutte, G., Herrewijn, L., Malliet, S., Pabian, S., Broeckhoven, F. V., Troyer, O. D., Deglorie, G., Hoecke, S. V., Samyn, K., and Bourdeaudhuij, I. D. (2018). The efficacy of the Friendly Attac serious digital game to promote prosocial bystander behavior in cyberbullying among young adolescents: A cluster-randomized controlled trial. *Computers in Human Behavior*, 78:336 – 347.

Fort, K., Guillaume, B., and Chastant, H. (2014). Creating *Zombilingo* , a game with a purpose for dependency syntax annotation. In *Proceedings of the First International Workshop on Gamification for Information Retrieval - GamifIR '14*, pages 2–6, Amsterdam, The Netherlands. ACM Press.

Hee, C. V., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., Pauw, G. D., Daelemans, W., and Hoste, V. (2015). Detection and fine-grained classification of cyberbullying events. In Galia Angelova, et al., editors, *Recent Advances in Natural Language Processing, RANLP 2015, 7-9 September, 2015, Hissar, Bulgaria*, pages 672–680. RANLP 2015 Organising Committee / ACL.

Jurgens, D. and Navigli, R. (2014). It's All Fun and Games until Someone Annotates: Video Games with a Purpose for Linguistic Annotation. *Transactions of the Association for Computational Linguistics*, 2:449–464, December.

Kowalski, R. M., Giumetti, G. W., Schroeder, A., and Lattanner, M. R. (2014). Bullying in the digital age: a critical review and meta-analysis of cyberbullying research among youth. *Psychological bulletin*, 140 4:1073–137.

Krause, M., Takhtamysheva, A., Wittstock, M., and Malaka, R. (2010). Frontiers of a paradigm: exploring human computation with digital games. In *Proceedings of the ACM SIGKDD Workshop on Human Computation - HCOMP '10*, pages 22–25, Washington DC. ACM Press.

Madge, C., Bartle, R., Chamberlain, J., Kruschwitz, U., and Poesio, M. (2019). Incremental Game Mechanics Applied to Text Annotation. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, pages 545–558, Barcelona Spain, October. ACM.

Menini, S., Aprosio, A. P., and Tonelli, S. (2020). A multimodal dataset of images and text to study abusive language. In Johanna Monti, et al., editors, *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021*, volume 2769 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Menini, S., Aprosio, A. P., and Tonelli, S. (2021). Abuse is contextual, what about nlp? the role of context in abusive language annotation and detection. *arXiv preprint arXiv:2103.14916*.

Naglé, T., Bateman, S., and Birk, M. V. (2021). Pathfinder: The Behavioural and Motivational Effects of Collectibles in Gamified Software Training. *Proceedings of the ACM on Human-Computer Interaction*, 5(CHI PLAY):1–23, October.

Nitta, T., Masui, F., Ptaszynski, M., Kimura, Y., Rzepka, R., and Araki, K. (2013). Detecting cyberbullying entries on informal school websites based on category relevance maximization. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 579–586, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., and Ducceschi, L. (2013). Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on*

Interactive Intelligent Systems, 3(1):1–44, April.

Rathnayake, G., Atapattu, T., Herath, M., Zhang, G., and Falkner, K. (2020). Enhancing the identification of cyberbullying through participant roles. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 89–94, Online, November. Association for Computational Linguistics.

Ryan, M. and Staines, D. (2016). Four Lenses for Designing Morally Engaging Games. page 16.

Safi Samghabadi, N., López Monroy, A. P., and Solorio, T. (2020). Detecting early signs of cyberbullying in social media. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 144–149, Marseille, France, May. European Language Resources Association (ELRA).

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

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.

Vannella, D., Jurgens, D., Scarfini, D., Toscani, D., and Navigli, R. (2014). Validating and Extending Semantic Knowledge Bases using Video Games with a Purpose. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1294–1304, Baltimore, Maryland. Association for Computational Linguistics.

Venhuizen, N. J., Evang, K., Basile, V., and Bos, J. (2013). Gamification for Word Sense Labeling. In *Proceedings of the International Conference on Computational Semantics (IWCS)*, pages 397–403.