Challenges in Designing Games with a Purpose for Abusive Language Annotation

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

Abstract

In this paper we discuss several challenges related to the development of a 3D game, whose goal is to raise awareness on cyberbullying while collecting linguistic annotation on offensive language. The game is meant to be used by teenagers, thus raising a number of issues that need to be tackled during development. For example, the game aesthetics should be appealing for players belonging to this age group, but at the same time all possible solutions should be implemented to meet privacy requirements. Also, the task of linguistic annotation should be possibly hidden, adopting so-called orthogonal game mechanics, without affecting the quality of collected data. While some of these challenges are being tackled in the game development, some others are discussed in this paper but still lack an ultimate solution.

1 Introduction

Cyberbullying has been recognised as a ubiquitous public health issue, with severe negative consequences for teenagers. Studies show that victims are more likely to suffer from psychosocial difficulties, affective disorders and lower school performance (Tokunaga, 2010; Kowalski et al., 2014), with consequences that are just as serious as for traditional bullying. Indeed, both bullying and cyberbullying are universal problems which affect all countries to a greater or lesser extent, regardless of the culture and country of origin of victims and aggressors (Zych et al., 2015).

Cyberbullying attacks are frequent in private chats and channels, while only a small fraction of them is visible in public accounts. This makes the analysis of cyberbullying phenomena difficult, since few data of this kind are accessible, making it hard to study the behaviour of adolescents online, especially if they are underage. 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 data from ask.fm (Hee et al., 2015; Safi Samghabadi et al., 2020), where however users are anonymous, making it difficult to focus only on adolescents.

Novel ways to understand the behaviour of teenagers with respect to verbal abuse online are therefore needed. Also, it is important to understand their perspective on different types of hate messages, so to develop systems that recognise cyberbullying by applying adolescents' point of view. Indeed, adolescents being part of virtual communities may share communication habits and slang that are different from those of adults, making it difficult for people outside these communities to judge the messages' offensiveness.

In order to deal with this lack of resources, we are developing a game called High School Superhero (Bonetti and Tonelli, 2020), whose goal is two-fold: on the one hand, it is meant to be used by young people to raise awareness on cyberbullying and on the use of offensive language, by playing the role of bystander in offensive interactions among peers. On the other hand, it is designed to collect annotations on hate speech texts, in particular whether players consider a message offensive or not, and which text span should be replaced or removed to make it not offensive. Accomodating these two tasks poses a series of challenges that we discuss in the remainder of this paper.

2 Related work

To date, there have been several attempts to gamify a wide range of linguistic annotation tasks. These include, among others, *Phrase Detectives* for anaphora resolution (Poesio et al., 2013), *The Knowledge Towers* (Vannella et al., 2014) and *Puzzle Racer* (Jurgens and Navigli, 2014) for concept-

Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing, pages 60–65
April 20, 2021.

image linking, *Infection* (Vannella et al., 2014), *OnToGalaxy* (Krause et al., 2010) and *JeuxDeMots* (Joubert et al., 2018) for semantic linking, *Argotario* (Habernal et al., 2017) for fallacious argumentation identification, *Zombilingo* (Fort et al., 2014) for dependecy syntax annotation, *Sentimentator* (Öhman and Kajava, 2018) for sentiment annotation, *WordClicker* (Madge et al., 2018) for part-of-speech tagging, *Wordrobe* (Venhuizen et al., 2013) and *Ka-Boom!* (Jurgens and Navigli, 2014) for sense annotation. Researchers stress the fact that games with a purpose (GWAPs) should be designed in such a way that they integrate the task without sacrificing their 'gamefulness', otherwise the tasks may be perceived as work (Vannella et al., 2014).

Concerning the use of gamification to raise awareness against cyberbullying, past works showed that increasing empathy is crucial to control cyberbullying (Barreda-Ángeles et al., 2021; Del Rey et al., 2016) and games can help in this sense as shown in 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.

3 High School Superhero

3.1 Setting and gameplay

High School Superhero is a role-playing game set in the school of a small fictional town. Both the school and the town are freely explorable, allowing a certain degree of free roaming. The player is a teenager who has been appointed by the school principal and a scientist to limit bullying.

The main activity consists of engaging players in two different tasks: in Task 1 players are instructed to change other people's messages if they consider them offensive, so that they become inoffensive. To do this they can participate in conversations among non-playing characters and activate a special device that allows them to read their minds. The words displayed in the thought bubble can be changed freely, but the game allows for trap gold sentences to be inserted for quality control. Once the thought is changed, the message is displayed with a surprised reaction from the author. In this way we know which tokens are considered offensive and at the same time we can get alternative sentences. Both the original sentence and the changed one are stored and can be exported at the end of the game, so to be used as a sort of minimal pair containing an offensive and a non-offensive sentence. Changing a token consumes one battery bar, which has to be replenished by exchanging collectibles from specific terminals. Task 2 is similar to Task 1

Figure 1: Task 1: Change other people's messages

except messages are displayed on walls and floors in the form of graffiti or writings on blackboards, and players can erase the words they deem abusive. Erasing has a cost in terms of consumable sponges, which can be replenished like the batteries used in Task 1. In this way, we encourage players to delete only the offensive part of a message (an adjective, a slur), while keeping the rest unchanged, easing the identification of the exact offensive span for future linguistic analyses. The collectibles used to replenish the resources employed in the two tasks are scattered around the town, which means that exploration is deeply interconnected with the tasks.

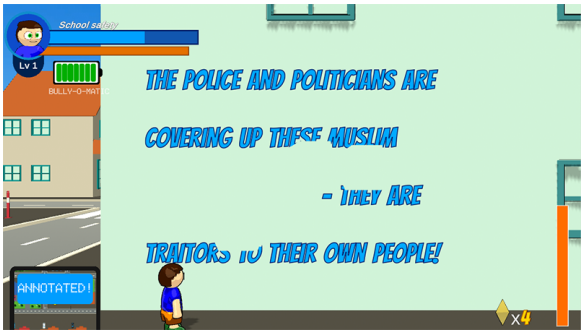

Figure 2: Task 2: Erase abusive words

3.2 Platforms and data management

The game is meant to be played either offline or online. It can be installed on Windows, Machintosh and Linux platforms, but it can also be played via browser thanks to Unity Web Player. At the moment, we do not have plans to bring the game to mobile platforms, such as Android and iOS, especially because the game offers a type of interaction that requires a fair degree of accuracy (namely, clicking or tapping words and erasing them) that can be compromised by the different sizes of smartphone and tablet screens. Furthermore, in this particular type of gameplay, where the player is moved via traditional input, i.e. directional keys or a control stick, and where the third person camera is rotated freely, desktop environments, with a keyboard and a mouse or a controller, represent the optimal solution. The data to be loaded into the game can be put in a folder inside the game directory, as long as the format is .txt, .xml, .csv or .json and the sentences are properly tagged or follow a specific pattern. The aim is to make it easy to feed data to the game, which, provided that there are the appropriate tags, automatically sorts the data and assigns it to each task (namely conversations and graffiti) and adapts it accordingly.

4 Challenge 1: Appealing aesthetics

The first challenge we identified during the development of High School Superhero is the game aesthetics: the game should not have a poor graphics, but rather be comparable with similar games that teenagers may know. It should also foster identification and involvement.

4.1 Graphics style

Since the game is primarily targeted at teens and young teens, the graphics style is cartoonish, with bright colors, and simple, rounded shapes. It was inspired by some of the latest successful commercial games like Fortnite, Animal Crossing: New Horizons, The Legend of Zelda: Breath of the Wild, Immortals: Fenyx Rising. This style allows one to keep the graphics simple when it comes to 3D modeling and shading, and therefore makes it possible to set up a virtual world without employing professional artists.

The game is being developed with Unity3D[1] and the resources, such as meshes and shaders, are both created from scratch using Blender[2] (an extremely powerful open-source 3D modeling and rendering program) and downloaded for free from the Unity Asset Store. The choice of using 3D graphics was also inspired by the fact that, although a vast portion of successful commercial video games is three-dimesional, to our knowledge a 3D game with a purpose for linguistic annotation had not been developed yet.

4.2 Avatar customization

We implemented avatar customization for three reasons: first, a custom avatar is the hallmark of any role-playing game and, more generally, this feature is starting to be ubiquitous, with aesthetic modifications and cosmetic collectibles appearing even in games where the avatar has only minimal importance (see Microsoft's Forza Horizon 4, a racing game where the driver can be customized, as an example). Second, character customization proved to increase the perceived agency and foster identification and empathy with one's avatar (Turkay and Kinzer, 2014), which is important for the specific research domain (abusive language). Third, to a higher sense of agency may follow higher intrinsic motivation according to Self-Determination Theory (Ryan and Deci, 2000; Turkay and Adinolf, 2015). Note that for avatars no pre-defined gender categories are given, and the player is free to customize it selecting the preferred physical traits, clothes and colors in any combination. With the purpose of increasing identification and empathy we are also implementing a minimal story line.

Figure 3: Avatar customization interface

5 Challenge 2: Add Orthogonal Mechanics

According to Tuite (2014), orthogonal game mechanics are those mechanics that "do not serve the

[1] https://unity.com/

[2] https://www.blender.org/

purpose". This type of mechanics is functional to a design strategy that Krause et al. (2010) call disjoint design, where "the design and the actual task do not correspond to each other directly". For example, a game where one must aim in order to select a label or an item to be labeled is said to employ orthogonal game mechanics, because aiming requires additional skills that do not strictly pertain to the annotation task. A game with a purpose for linguistic annotation that implements this type of mechanics is *OnToGalaxy* (Krause et al., 2010), where players control a spaceship and, given a concept and a type of semantic relationship, they have to shoot the enemy spaceships accordingly. The goal is to select the spaceships whose label is in relation with the given concept (by ignoring them) and to discard the spaceships whose label is not in relation with the given concept (by shooting them). An example of game with non-orthogonal game mechanics, instead, is *WordClicker* (Madge et al., 2019), where the mechanics directly overlap with those of a clicker game, which is a widespread and quite addictive type of game.

While orthogonal game mechanics may not always the best option, because they require players to have additional skills beside being able to perform the annotation task, they can be useful sometimes to engage players and make the interaction more attractive and interesting, allowing for increased resemblance with mainstream games.

One of the challenges with the development of our game is to introduce a certain degree of orthogonality, thus hiding the task mechanics, while at the same time preserving the overall annotation quality. Task 1 of High-School Superhero does not make use of orthogonal game mechanics, because abusive words are selected by clicking them, as in a non-gamified interface. Task 2, on the other hand, is orthogonal since words are selected by being rubbed with an eraser or a sponge, and one must pay attention not to exceed the boundaries of the abusive words. Although some quality control techniques are used, such as assessing the proportion of the word that has been erased, Task 2 requires the additional ability of erasing with a certain degree of accuracy. Specifically, in the present implementation of orthogonal mechanics, the rate of erased words can be affected by the depletion of resources, while the annotation accuracy can be affected by the sponge size and shape. Comparing the two strategies will be useful to understand whether orthogonal mechanics (potentially more engaging) represent a good alternative to non-orthogonal mechanics in terms of annotation rate and reliability.

6 Challenge 3: Respect Users' Privacy and other Ethical Considerations

Since this game aims to collect annotations that can be potentially updated across multiple sessions, and since we want annotations to be linked to specific demographic information, a login system and demographic questionnaires are employed. Indeed, the game will be used in the framework of an international collaboration with schools, and the link to play High School Superhero will be shared only with the schools involved in cyberbullying awareness initiatives. We did not want complex authentication systems to clutter the game interface, so a Google Login system is used where only the email address is read. The address is then converted into a SHA256 string and used as a user ID, while the address is deleted. In this way, we can control how many players have interacted with the game and check basic demographics, without making use or storing any sensitive information. This is particularly relevant in our context, since we involve underage subjects. Providing a completely free access without any kind of login would allow for a fully anonymous use of the game, but we would run the risk to have players producing low-quality data, or losing the possibility to target a specific age group.

Other issues related to having teenagers play our game concern their exposure to offensive language. In order to be aware of cyberbullying and be ready to respond to this phenomenon appropriately it is necessary that players interact with offensive texts. However, this may affect their well-being and have an impact on particularly vulnerable subjects. In order to have access to students, we will have to ask for the approval from the Ethical Committee in our University. Furthermore, we introduced the possibility, when using a local installation of the game, to manually upload the list of texts that will be displayed during the game, so that educators can pre-screen the texts removing extremely hateful messages. This approach is not ideal, since manual filtering is time-consuming and may not work well on a large scale. Alternative approaches will be explored to pre-select texts using NLP or remove messages based on lists of taboo words.

7 Conclusions and Future Work

In this paper we have discussed a number of challenges we are addressing in the development of High School Superhero, an online game for teenagers whose goal is two-fold: on the one hand, it can be used to raise awareness on cyberbullying and the use of offensive language among peers. On the other hand, it can serve for collecting instances of offensive and non-offensive language. In the future, we plan to evaluate the game with students in classes along different dimensions, focusing especially on motivation by employing questionnaires based on Self-Determination Theory. We are particularly interested in assessing differences in motivation because it is a key aspect for GWAP dissemination and data collection and can vary a lot across different tasks and different applications.

Acknowledgments

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". We would like to thank Alessia Smeraglia for the help provided in designing the virtual world and the game play of High School Superhero during her internship.

References

Miguel Barreda-Ángeles, Maria Serra-Blasco, Esther Trepat, Alexandre Pereda-Baños, Montserrat Pàmias, Diego Palao, Ximena Goldberg, and Narcís Cardoner. 2021. Development and experimental validation of a dataset of 360°-videos for facilitating school-based bullying prevention programs. *Computers & Education*, 161:104065.

Federico Bonetti and Sara Tonelli. 2020. A 3D role-playing game for abusive language annotation. In *Workshop on Games and Natural Language Processing*, pages 39–43, Marseille, France. European Language Resources Association.

Antonio Calvo-Morata, Manuel Freire-Moran, Ivan Martinez-Ortiz, and Baltasar Fernandez-Manjon. 2019. Applicability of a Cyberbullying Videogame as a Teacher Tool: Comparing Teachers and Educational Sciences Students. *IEEE Access*, 7:55841–55850.

Rosario Del Rey, Lambros Lazuras, José A. Casas, Vassilis Barkoukis, Rosario Ortega-Ruiz, and Haralambos Tsorbatzoudis. 2016. Does empathy predict (cyber) bullying perpetration, and how do age, gender and nationality affect this relationship? *Learning and Individual Differences*, 45:275–281.

Ann DeSmet, Sara Bastiaensens, Katrien Van Cleemput, Karolien Poels, Heidi Vandebosch, Gie Deboutte, Laura Herrewijn, Steven Malliet, Sara Pabian, Frederik Van Broeckhoven, Olga De Troyer, Gaetan Deglorie, Sofie Van Hoecke, Koen Samyn, and Ilse De Bourdeaudhuij. 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.

Karën Fort, Bruno Guillaume, and Hadrien Chastant. 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.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational Argumentation Meets Serious Games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 7–12, Copenhagen, Denmark. Association for Computational Linguistics.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Recent Advances in Natural Language Processing, RANLP 2015, 7-9 September, 2015, Hissar, Bulgaria*, pages 672–680. RANLP 2015 Organising Committee / ACL.

Alain Joubert, Mathieu Lafourcade, and Nathalie Le Brun. 2018. The JeuxDeMots Project is 10 Years Old: What We have Learned. In *Proceedings of the 2018 LREC Workshop "Games and Gamification for Natural Language Processing (Games4NLP)"*, pages 22–26, Miyazaki, Japan.

David Jurgens and Roberto Navigli. 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.

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

Markus Krause, Aneta Takhtamysheva, Marion Wittstock, and Rainer Malaka. 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.

Chris Madge, Richard Bartle, Jon Chamberlain, Udo Kruschwitz, and Massimo Poesio. 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. ACM.

Chris Madge, Massimo Poesio, Udo Kruschwitz, and Jon Chamberlain. 2018. Testing TileAttack with Three Key Audiences. In *Proceedings of the 2018 LREC Workshop "Games and Gamification for Natural Language Processing (Games4NLP)"*, pages 6–11, Miyazaki, Japan.

Stefano Menini, Alessio Palmero Aprosio, and Sara Tonelli. 2020. A multimodal dataset of images and text to study abusive language. In *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.

Taisei Nitta, Fumito Masui, Michal Ptaszynski, Yasutomo Kimura, Rafal Rzepka, and Kenji Araki. 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. Asian Federation of Natural Language Processing.

Emily Öhman and Kaisla Kajava. 2018. Sentimentator: Gamifying Fine-grained Sentiment Annotation. In *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference (DHN 2019)*, volume 2084, pages 98–110, Helsinki, Finland.

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. 2013. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems*, 3(1):1–44.

Richard M Ryan and Edward L Deci. 2000. Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being. *American Psychologist*, page 11.

Niloofar Safi Samghabadi, Adrián Pastor López Monroy, and Thamar Solorio. 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. European Language Resources Association (ELRA).

Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. 2018. Creating a WhatsApp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59, Brussels, Belgium. Association for Computational Linguistics.

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.

Kathleen Tuite. 2014. GWAPs: Games with a Problem. In *Proceedings of the 9th International Conference on the Foundations of Digital Games*.

Selen Turkay and Sonam Adinolf. 2015. The effects of customization on motivation in an extended study with a massively multiplayer online roleplaying game. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 9(3).

Selen Turkay and Charles K. Kinzer. 2014. The Effects of Avatar-Based Customization on Player Identification:. *International Journal of Gaming and Computer-Mediated Simulations*, 6(1):1–25.

Daniele Vannella, David Jurgens, Daniele Scarfini, Domenico Toscani, and Roberto Navigli. 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.

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

Izabela Zych, Rosario Ortega-Ruiz, and Rosario Del Rey. 2015. Systematic review of theoretical studies on bullying and cyberbullying: Facts, knowledge, prevention, and intervention. *Aggression and Violent Behavior*, 23:1–21. Bullying, Cyberbullying, and Youth Violence: Facts, Prevention, and Intervention.