Requirements of a Cyberbullying Management System

Group D

January 8, 2018

Contents

1 Introduction/Abstract

Cyberbullying is the use of electrical technology to harass, threaten, embarrass, humiliate or target another person. Cyberbullying is rapidly becoming a widespread phenomenon [1, 2], claiming more and more victims each day. It is imminent within any form of communication that some kind of bullying will follow. Digital communication is no exception. Managing large classes of grade students in real life with regards to bullying can be a daunting task on its own, but given that no teacher can be omnipresent even digitally, it might seem an insurmountable problem. The cyberbullying problem is growing.

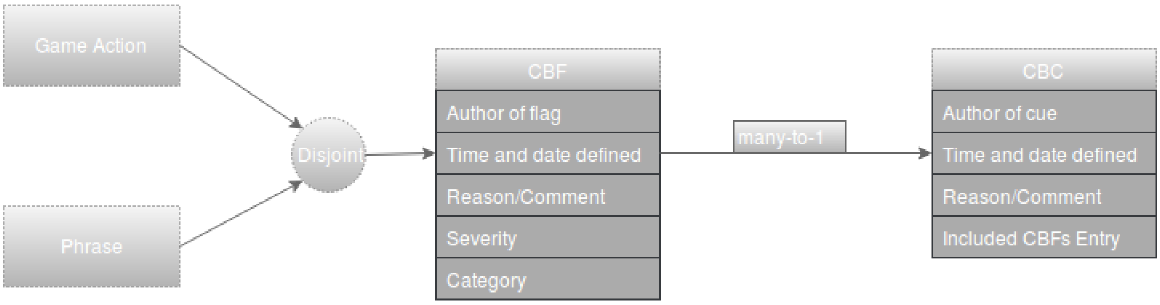
This document seeks to explore the requirements necessary of a system that could facilitate the management and detection of cyberbullying within an immersive game environment. Literature has been researched to narrow the possible categories of bullying. It has been concluded that bullying is a vast and complex subject, with context playing a huge role in any given situation, and therefore the actual detection has been given a minor role in this paper, as management of detected bullying has turned out to be a far more feasible problem to deal with in an abstract manner.

The requirements listed in this paper will produce and allow developers of immersive game environments to feed the system their own data and actions to decide what constitutes as bullying within their specialized game environment. The system handles recognition of bullying in many forms and provides relevant information to the stakeholder. The system will also produce several logs that can be of scientific value for further research.

Cyberbullying has become a major concern in today's technology society [2]. No matter what age, gender or nationality the user inherited from, this is a major problem all over the world. As an adult, cyberbullying is considered a federal crime which can lead to actions like fine or even jail. The interesting question is; how can we prevent this behavior and how can we detect it? [3]

2 Glossary

- **IGE** - Immersive Game Environment
- **CBMS** - Cyberbullying Management System
- **Player** - Any person who has the ability to participate in the game.
- **Student** - Any person who is a player but who does not have access to the CBMS
- **Teacher** - Any person who is a player and has access to the CBMS.
- **Admin** - A person with access to the CBMS but who is not a participant of the game.
- **User** - Player or Admin.
- **Game Action** - An action that is possible to perform in the game. This is specified by the developers and depends on the game.
- **Cyberbullying Flag** or **CBF** - A single Game Action or phrase that a Teacher or Admin identify as being worth noting as a Cyberbullying Flag Entry. A CBF has **occurred** when the action or phrase defining that CBF are detected.



A model of the CBF

- **Cyberbullying Flag Entry** - The entry in the logging system that the CBMS logs when any action defining a flag is detected in the game.
- **Cyberbullying Cue** or **CBC** - a collection of one or more CBFs defined by a teacher. A CBC has **occurred** when the CBMS detects that the collection of flags that defines the CBC has occurred within the game, that creates a CBC-entry
- **Cyberbullying Cue Entry** or **CBC Entry** - The logged aggregate data of all the flag entries defining the CBC.
- **Cyberbullying Event** or **CBE** - A CBC Entry that is assessed to require some corrective action by a teacher.
- **User log** - The logging subsystem containing all users. Each user in the log contain information regarding involvement in a CBC/CBE, created or modified CBFs, CBFs raised and which other users they have reviewed in the log.
- **Flag log** - The logging subsystem containing all the entries.
- **Log Entry** - A CBF entry, CBC entry or CBE entry.
- **Severity Index** - Each specific flag should be provided with a severity index. The severity index is a rank from 0-2 with purpose to rank how severe/urgent the cyberbullying matter is.
  - Rank 0: The flag does not contain any trace of direct cyberbullying.
  - Rank 1: The flag contains events of cyberbulling but it is not severe.
  - Rank 2: The flag includes serious indications of cyberbullying which is severe.

- **Categories -** Each flag should be able to be classified into a category. We use categories provided by Van Hee in her report about cyberbullying [4]. We also added some extra categories to match the IGE. Below follows a direct excerpt from the report but with own examples but also the added categories.
  - *Threat*: expressions containing physical or psychological threats or indications of blackmail (e.g. When I see you after class I will punch you).
  - *Insult*: expressions containing abusive, degrading or offensive language that are meanth to insult the addressee. (e.g. I think you are the ugliest person I know)
  - *Curse/Exclusion*: expressions of a wish that some form of adversity or misfortune will befall the victim and expressions that exclude the victim from a conversation or a social group (e.g. Just kill yourself).
  - *Defamation*: expressions that reveal confident or defamatory information about the victim to a large public (e.g. She had the ugliest clothes when I saw her today.)
  - *Sexual talk*: expressions with a sexual meaning that are possibly harmful (e.g. Post a naked pic, now!!).
  - *Defense*: expressions in support of the victim, expressed by the victim himself or by a bystander (e.g. Shut up about my sister, she is not an idiot!)
  - *Encouragement to the harasser*: expressions in support of the harasser (e.g. Haha, you're so right, he's a fat donkey)
  - *Spamming*: When a player constantly during a long period of time talks with another player, but the target player does not respond.
  - *Stalking*: A player follows another player around for a long time and during several occations.
  - *Self-raised flag*: When a player in the IGE raise a flag by themselfs.

3 Scope Statement

3.1 Project Purpose and Justification

This document concerns a system for detecting bullying and notifying responsible figures such as teachers about it in an IGE. An IGE is a game where actions can be performed and communications can be conveyed either by actions in-game or communication via voice. By allowing automatic detection of cyberbullying, a teacher might get an easier time with their job.

3.2 Scope Description

The scope of this document is to design the requirements for a cyberbullying management system. The system is supposed to be usable with any immersive game environment (IGE) and is language agnostic. In other words it is meant to work as a general interface. It should flag possible cases of harassment either via text/voice, actions performed, or a combination of these by it's users. If the IGE provides the possibility of user generated content all content will be checked for possible harassment or hate speech.

3.3 Main Stakeholders

- Educators
- Children
- Parents - Minor Stakeholders

3.4 Boundaries

3.4.1 Out-of-scope

Implementation of the system in non-IGE-products. For example chat-systems. All voice communication will be translated into text and treated as it via a voice to text software. Tonality or any voice analysis will not be a part of the system to determine possible harassment.

3.4.2 In-Scope

1§ Flagging of possible harassment or hate speech conveyed via communications.

2§ Flagging of possible harassment conveyed via actions.

3§ Flagging of possible harassment conveyed via a combination of communications and actions.

4§ Flagging of user generated content that can be considered hate speech or harassment.

5§ Handling flags of different severity levels.

6§ Possibility to implement into any IGE.

7§ Possibility for developers to define what a specific actions are in their game and let them determine if it's a part of the core mechanic of the game that a user must do to actually play the game.

3.5 Strategy

Some actions in some games are classified as harassment. Say if you go around destroying peoples projects in Minecraft (also called "griefing") that could be considered harassment. However in some games that might be a core game mechanic that you need to do to actually play the game. That is why developers will get an interface to program against to define what features of their game is considered harassment or just a game mechanic.

If something is a core game mechanic a flag will not be raised when this is performed. This interface also makes it possible for programmers to define what a specific action is in their game data-wise. What in their game for example is considered hitting, running away from person b and so on. It might be when a 0 becomes a one in some variable, or it might be calling some function called "bool hitting(Person personToHit)" returning true or false depending on the result of the punch thrown.

3.6 Deliverables

- A requirement specification for a cyberbullying detection system
- A finished backlog

4 Specification

4.1 General System Specifications

1§ Classification and detection of textual/voice communication shall be language agnostic, or easily adapted to a new language.

2§ Detection, classification and reporting, shall, to an as large extent possible, happen automatically without the need for any teacher intervention.

4.1.1 CBF management

1§ Each message communicated shall, if considered a flag, be classified into one of the categories provided by Van Hee et al.[4]

2§ Each CBF and CBC shall be assigned a severity index depending on how serious the flag is. These should be customized by the teacher, according to the needs of the class.

3§ Teachers are able to decide which types of flags in the severity index that will immediately be reported and circumvent any methods that work to prevent the reporting of false positives.

4§ A set of flags predefined by game developers or other 3rd parties shall be present.

5§ If CBF is not assigned to be a part of any CBC that flag will be considered a CBC on its own unless explicitly marked to be ignored.

6§ The set of messages and flags shall, if needed, be augmented a teacher/supervisor.

4.1.2 CBC Management

1§ A CBC is a CBF alone or an combination of CBFs.

2§ A CBC containing a high severity index should trigger notifications.

3§ A teacher shall be able to choose if certain types of CBCs should trigger notifications.

4§ A teacher shall be able to reclassify a CBC-entry to a CBE.

5§ The end of a CBC-entry occurs when an amount of time has passed that is defined by a teacher. This called the timeout of the CBC-entry.

6§ Depending if the CBC-entry matches a CBC when it times out it will be logged as a CBC entry, if it does not match any it shall be stored as a Non-CBC-entry.

4.1.3 CBE management

1§ A CBE is a CBC-entry that manually have been notice and highlighted by the teacher as a CBE.

2§ When a CBC-entry is marked as an CBE then it means that it is up for evaluation.

3§ When a teacher classifies a CBC-entry as a CBE, that teacher is logged to the CBE.

4.2 UI subsystem

4.2.1 Teacher

1§ A teacher shall be able to create a CBF by specifying what action will constitute that CBF, along with its severity index.

2§ A teacher shall be able to create and combine several CBFs into CBCs.

3§ A teacher shall be able to view and edit their own CBF, or CBC at any time.

4§ All visible logged data is equally accessible by all teachers. Teachers shall be able to see CBFs that have occurred that are defined by other teachers.

5§ A teacher shall be able to delete their own (and only their own) definitions of CBF or CBC. This shall have no effect on the logging system (i.e. deleting the definition a CBF does not remove a previously occurred CBF from the log).

6§ The system shall be able to notify the teacher even outside of working hours.

7§ There shall exist an option for teachers to monitor a specific students activity more closely than other students.

8§ The UI system shall be able to communicate with the logging system. It shall allow the teacher to filter through the log data. Filtering may include the following options:

- Student name(s)
- Time of CBF/CBC/CBE
- Severity
- Only sorting by CBC, CBE, CBF
- Number of students involved
- Flagger (i.e. the teacher that defined the flag)

9§ It should also be able to present only the events from a given time interval.

10§ The UI system shall allow teachers to tweak notification settings, according to, for example, category of event or severity.

11§ The system shall not limit teachers to only be notified about CBCs that they defined.

12§ The UI shall provide a way to get quick overview over the most recent CBE/CBC's that have occurred.

13§ For any CBE or CBC-entry presented to the teacher, there shall be an option to view the the in-game actions that triggered the CBF within the context of the game.

14§ The UI system shall not allow the teacher to alter the log in any way.

15§ The teacher shall not be able to alter the severity indexes of other teachers' flags.

4.2.2 Player

1§ The player shall be able to flag an event that was interpreted as hostile. This shall be defined by the system as pre-defined CBF.

2§ A player shall also be able to report that another student is being bullied. This shall be a pre-defined CBF.

3§ The CBFs for a player reporting themselves or another player shall not be the same pre-defined flag.

4§ The player will otherwise not have any access to any part of the system.

4.3 Logging subsystem

4.3.1 Flag log

1§ All actions and communication, whether flagged or not shall be logged and permanently stored.

2§ All CBF-entries, CBC-entries and CBE's shall be stored permanently as log-entries.

3§ Log-entries shall be archived after an amount of time that the teacher decides.

4§ All log-entries can be archived. This removes it from the list of current log-entries however it shall not permanently erase them.

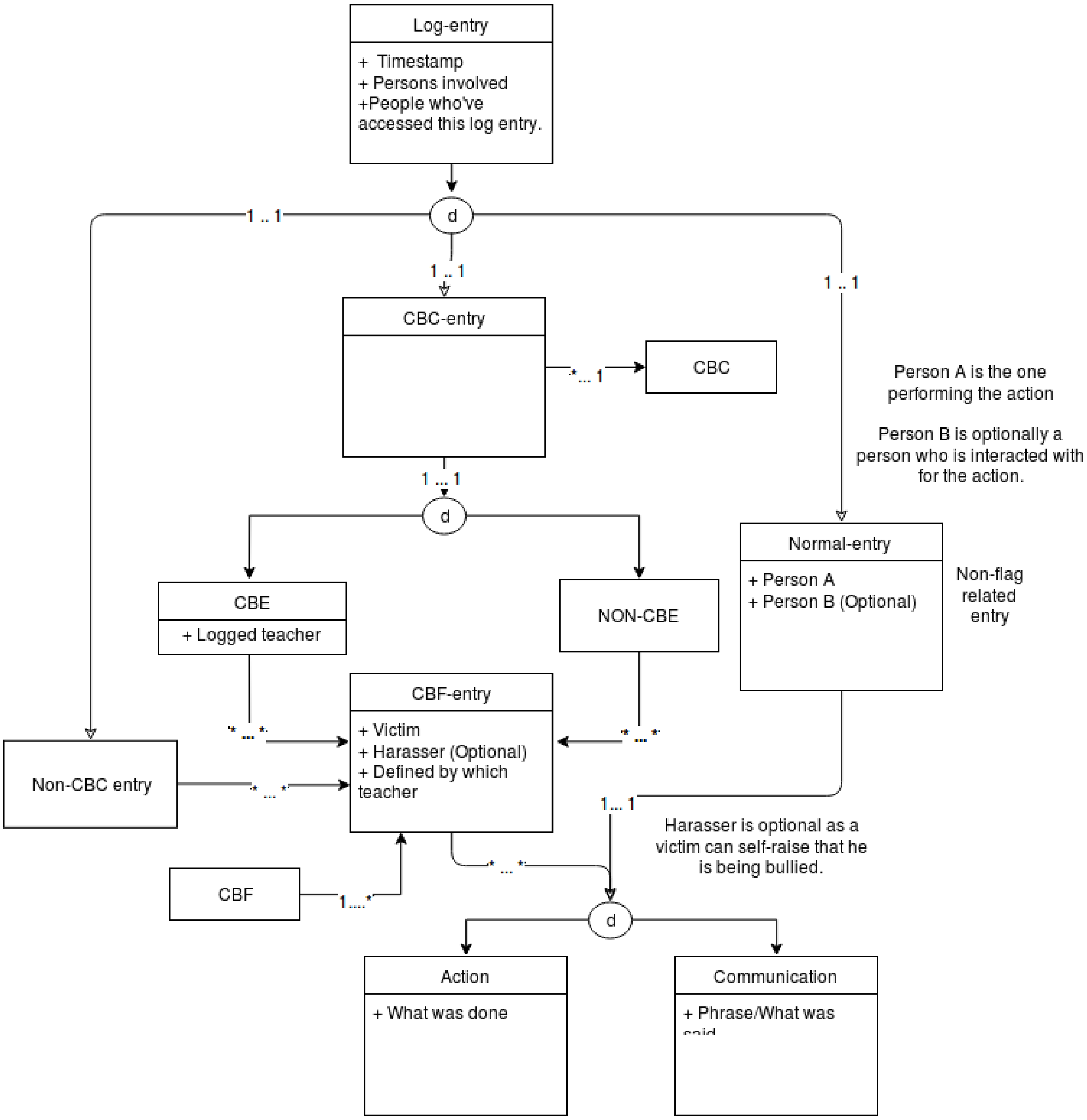
5§ It is not possible to permanently delete a log-entry, log-entries can only be archived.

6§ Archived log-entries shall be accessible if needed for forensic work on any case of bullying.

7§ Logged CBE's cannot be archived by a teacher who was a part of the CBC-entry that led to the CBE.

8§ The system shall maintain a log of what action teachers took after viewing a specific CBC-entry.

The logging system has the following relation model



4.3.2 User log

1§ The system shall store a history log of every user. This shall mention CBF, CBC, CBE entries they have been involved in.

2§ The system shall allow both teachers and admins to review each others user log histories. Both users shall also be able to review any student's log history.

3§ Reviewing another users history log shall be mentioned in the reviewers history log.

4§ It is not possible to delete or modify any users log history.

5§ All user histories in the user log is stored permanently.

4.4 Security and Integrity

4.4.1 User Privileges

1§ A teacher shall be able to monitor a players behavior in a detailed fashion if that is deemed necessary.

2§ Students have no access to the management system, apart from self reporting abusive behavior.

3§ A teacher shall not be able to view or review any CBC that they themselves are a art of.

4§ In the situation where a teacher is part of CBC another teacher or admin must review that CBC.

5§ An admin user has superior privileges to a teacher.

6§ An admin user must not be a participant of the game.

5 Usage Example

Alice and Bob are two students playing the game. Alice repeatedly steals items from Bob and Alice keeps attacking Bob. Each of these actions have been marked by the Teacher as being CBFs. Log entries for the CBF events are generated. If the teacher also adds these events to a CBC, The CBMS detects that these are not isolated events and therefore creates a CBC entry composed of these flags.

The creation of the CBC notifies their teacher. The teacher then manually inspects this CBC to see if the actions taken by Alice do indeed constitute bullying. Their teacher decides that this do indeed constitute bullying and marks this CBC as a CBE and takes appropriate action. The teacher could also decide to look at the logs of previous CBCs and CBEs to try to get a better picture of Alice' behavior.

5.0.1 Use Cases

Teacher may manage and oversee flags in the cyberbullying management system.

- Teacher may define new CBFs and CBCs.
- Teacher may edit and delete CBFs and CBCs.
- Teacher may put specific student under special attention.
- Teacher may change settings such as notification frequency.

Teacher may review logged data.

- Teachers may review the user log.
- Teachers may review the flag log.
- Teacher may filter log entries by CBE, CBCe or CBFe.
- Teacher may jump to certain time and place in the game specified by any log entry.

- Teacher may upgrade CBC events to CBE.
- Teacher may review actions taken for CBEs.
- Teacher may review other users log by specifying real or player name.

Admin provides support.

- Admin may edit teacher privileges.
- Admin may review all the log entries.
- Admin may perform system updates.
- Admin may review the user log.

6 Risk Analysis

We have identified some risks with our CBMS. These risks range from risks related to both the implementation of the CBMS and general risks concerning the subject of data gathering and processing of personal information.

6.1 System Risks

These are the risks that concern the actual functional aspects of the CBMS. The most prudent course of action to limit these risk is to employ a development process that yields ample resources for testing and verification.

6.1.1 Information Leakage

One of the most serious risks is that of information leakage. Since the system will log all of the players activities including their conversation a compromise of the systems integrity could have severe consequences for the individuals whose communications were disclosed. Even though there is no financial or material damage that could be caused by this the students could be put in an extremely compromising situation if conversations regarding, for example domestic abuse, that where held in confidence where to be made available.

The best way of safeguarding against this is to during the entire life cycle of the system pay great attention to security risks and to frequently audit the system for potential exploits.

6.1.2 Wrong User Credentials

Another serious risk is that of inappropriate privileges being granted to people who are not qualified. An inappropriate privilege escalation could not only lead to the integrity of the stored information being compromised but it could also enable students to get away with bullying without being detected.

This risk is also best mitigated through being extra careful when granting privileges to people.

6.1.3 Inappropriate Admin authority

The admin users of the systems do, unfortunately constitute a single point of failure. Since the admins have near total authority over the system it would have grave consequences if their powers were abused. Unfortunately there is no way to, from the design perspective, prevent the damage that would be caused by a corrupt or malicious user with admin privileges.

6.1.4 Teacher incompetence

The system is predicated on the fact that the teachers do indeed investigate CBCs and act accordingly. If a teacher where to be incompetent or negligent and not pay attention this could lead to players getting away with bullying. This risk is somewhat mitigated trough there being many teachers available in the system and hopefully some of them will be competent and pay enough attention.

6.2 Other Risks

These are risks that lie outside of the development and maintenance aspect of the CBMS but should still be taken into consideration when deploying the CBMS and when the CBMS is in use.

6.2.1 Legal Issues

Since the CBMS is designed to be implemented in different languages and in different cultures it is not possible to a priori know the legal status of the CBMS. Some countries may have very restrictive laws regarding the storage and processing of personal information. An additional complications happens in the situation where criminal conduct have been observed trough the CBMS. Some countries may have strict regulations regarding the handling of evidence in criminal proceedings.

The only way to mitigate this problem is to make sure that there is some entity with knowledge about the concerned countries legal system present at the deployment of the CBMS. This to ensure that any restrictions or user agreements necessary for the CBMS, in that legal setting, can be applied prior to usage.

6.2.2 Ethical Issues

There are some inherent ethical issues concerning the CBMS. The CBMS does indeed log the behavior and communication of children outside of school hours. This raises concerns regarding whether or not it is appropriate for one or many teachers to take part of students communications indiscriminately.

This issue can not be resolved in any decisive way and it is up to the customer of the CBMS to deal with any ethical considerations of the system.

References

[1] Susan Keith and Michelle E Martin. Cyber-bullying: Creating a culture of respect in a cyber world. *Reclaiming children and youth*, 13(4):224, 2005.

[2] Robin M Kowalski, Susan P Limber, Sue Limber, and Patricia W Agatston. *Cyberbullying: Bullying in the digital age.* John Wiley & Sons, 2012.

[3] Luca Caviglione, Mauro Coccoli, and Alessio Merlo. *Social Network Engineering for Secure Web Data and Services.* IGI Global, 2013.

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