

Guidelines for the Fine-Grained Analysis of Cyberbullying

Version 1.0

LT3 Technical Report – LT3 15-01

Cynthia Van Hee¹, Ben Verhoeven², Els Lefever¹, Guy De Pauw²,
Walter Daelemans² and Véronique Hoste¹

¹LT3, Language and Translation Technology Team
Faculty of Arts and Philosophy
Ghent University

²CLiPS - Computational Linguistics Group
Faculty of Arts
University of Antwerp

URL: <http://www.lt3.ugent.be/en/>¹

August 25, 2015

¹The reports of the LT3 Technical Report Series (ISSN 2032-9717) are available from <http://www.lt3.ugent.be/en/publications>. All rights reserved.

Contents

1	Introduction	1
2	Defining Harmfulness	3
3	Cyberbullying Roles	4
4	Textual Categories related to Cyberbullying	5
4.1	Threat or blackmail	5
4.2	Insult	6
4.3	Curse or exclusion	6
4.4	Defamation	6
4.5	Sexual Talk	7
4.6	Defense	7
4.7	Encouragement to the Harasser	7
4.8	Other	8
5	Annotation Procedure and Examples	9
6	Using the Brat Annotation Tool	14

Chapter 1

Introduction

Over the past few years, the problem of **cyberbullying** has received increased attention. A recent study of Tokunaga (2010) revealed that about 20% to 40% of all youngsters have experienced some form of cyberbullying at least once in their lives. These figures demonstrate that cyberbullying is not a rare problem. Moreover, it can have a serious impact on children's and teenager's well-being, with studies linking it to depression, low self-esteem and school problems (Price and Dalgleish, 2010; Šléglová and Černá, 2011; Vandebosch et al., 2006). To improve child safety online, it is of key importance to identify possibly threatening situations on the Web. However, the ever-growing amount of information online has made human monitoring an unfeasible task. There is thus an urgent need for **intelligent systems** that signal potentially threatening situations such as cyberbullying automatically (e.g., through dashboard applications for moderators on social networking platforms).

Definitions of traditional bullying are a common point of departure for defining cyberbullying. When conceptualizing cyberbullying, researchers often refer to Olweus's (1993) three criteria of bullying: **repetitiveness**, the **intention to cause harm** and an **imbalance of power**. Based on these three criteria, Smith et. al (2008, p. 376) provided the following definition of cyberbullying:

Cyberbullying is 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.

Some doubt exists, nevertheless, as to whether all three criteria are necessary conditions for cyberbullying. For instance Grigg (2010) stresses that online posts are persistent and may be viewed by various recipients. A single aggressive utterance can result in continued and widespread ridicule for the victim and act as a repeated humiliation. Furthermore, online communication is prone to misinterpretation, which makes it hard to decide upon intentionality (Kiesler, Sigel, and McGuire, 1984; Vandebosch et al., 2006). Moreover, a power imbalance is difficult to assess in online bullying as it may be related to anonymity, the level of technological skills of the bully and the victim or the inability of victims to escape the situation (Dooley, Pyzalski, and Cross, 2009; Smith et al., 2008). Finally, analyzing online content by means of formal criteria is often hindered by the lack of context that characterizes this type of data. Considering these limitations, we conceptualize cyberbullying as textual content that is published online by an individual and that is aggressive or hurtful against a victim.

This paper describes the guidelines for the fine-grained **annotation** of cyberbullying data. Annotators will add pieces of information to **social media** messages. These messages will be annotated in context, this means that they will be presented within their original content or event (e.g., a

chat conversation or a Facebook post with replies) when available. All messages presented within the conversation or event should be considered. The objective is to indicate in these conversations the messages that are (potentially) harmful. As a major part of our dataset was collected via social networking sites, these annotation guidelines are developed according to the structure of this particular type of user-generated content.

Essentially, there will be **two levels** of annotation. The first is the level of the message or the post itself. Considering the entire message, annotators will define as a first step whether it contains cyberbullying or not by indicating the harmfulness of the message. When the message is considered harmful and thus contains indications of cyberbullying, annotators should indicate the role of the author of the post. At this level, annotators also indicate posts that are written in another language than English. These do not need to be further annotated. At the second level, text spans with relevant information to the use case of cyberbullying will be identified and categorized. Please note that these levels do interact and that a decision for one level can have implications on the other level.

All annotations are performed using **brat**, a Web-based tool for text annotation (Stenetorp et al., 2012).

Chapter 2

Defining Harmfulness

Social media posts or messages will be presented to the annotators in chronological order within their original conversation context when possible. Within every conversation, each post is preceded by a dummy token ¶. It is this dummy token that should be used for all post-level annotations (i.e., defining a harmfulness score and identifying the author's role (see Chapter 3)).

For each post, annotators define whether the post contains indications of cyberbullying and if so, whether these indications are severe. Based on this property, a harmfulness score on three-point scale (0-1-2) will be defined.

- **Harmfulness score = 0** → The post does not occur in a cyberbullying context or does not contain indications of cyberbullying.

(1) *Do you like Miley Cyrus?*
(2) *Hi bitchess, anyone in for a drink tonight?*

- **Harmfulness score = 1** → The post occurs in a cyberbullying context and contains indications of cyberbullying.

(3) *I don't like your photo, you've horse teeth.*

- **Harmfulness score = 2** → The post contains explicit and serious indications of cyberbullying (e.g., serious threats, incitements to commit suicide).

(4) *You're a peasant full of aids, just kill yourself*

Practical remark

For practical reasons, to speed up the annotation, the dummy token ¶ can be left unannotated when there is no sign of cyberbullying (i.e., when the post would receive a harmfulness score of 0). The harmfulness score should only be made explicit when it is 1 or 2. By convention, **serious threats** and explicit **instructions to perform sexual or harmful actions** or to **commit suicide** receive a harmfulness score of 2. To gain insight into the use of aggressive language in social media texts, annotators are asked to indicate cyberbullying-related text categories even if the post by which they are contained is not perceived harmful (see example (2) for instance, where *bitchess* should be marked as an insulting word).

Chapter 3

Cyberbullying Roles

Similar to traditional bullying, participants in a cyberbullying episode adopt well-defined **roles**. Based on the existing literature (Salmivalli, 2010; Salmivalli, Voeten, and Poskiparta, 2011; Xu et al., 2012), four cyberbullying roles are distinguished: **harasser**, **victim**, **bystander-defender** and **bystander-assistant**. When the harmfulness score of a post is 1 or 2, annotators should indicate the author's role within the cyberbullying event.

- **Harasser:** Person who initiates the harassment.
- **Victim:** Person who is harassed.
- **Bystander-defender:** Person who helps the victim and discourages the harasser from continuing his actions.
- **Bystander-assistant:** Person who does not initiate, but takes part in the actions of the harasser (e.g., by encouraging the harasser).

Some examples are listed below. The author's role is indicated between square brackets.

(5) *You're a fucking moron.* [harasser]

(6) *Why do you talk shit about me? Leave me alone.* [victim]

(7) *Stop making fun of people, I think she's a good person.* [bystander-defender]

(8) *LOL, you're right, he is a nobody.* [bystander-assistant]

Practical remark

Sometimes (e.g., when not enough context is present), it is difficult to distinguish between a victim and a bystander-defender, between a harasser and a bystander-assistant, or even between a victim and a harasser when the victim is verbally explicit in his response. In this case, annotators should select the role that is most appropriate with the information available. For example, when annotators are not entirely sure whether an offensive post is posted by a harasser or by a victim defending himself, they should assign the role of harasser.

Chapter 4

Textual Categories related to Cyberbullying

In the literature, different **forms of cyberbullying** are identified (O’Sullivan and Flanagin, 2003; Price and Dalgleish, 2010; Willard, 2007) and compared with forms of traditional bullying (Vandebosch and Van Cleemput, 2009). Based on these forms, this annotation scheme describes some specific **textual categories** that are often **inherent to a cyberbullying event** such as threats, insults, defensive statements from a victim, encouragements to the harasser, etc.

Within every message, annotators indicate all text spans that correspond to one of the textual categories that are described below, even when used not in a cyberbullying context (i.e., when a message is not considered harmful). In example (2) for instance, annotators should mark *bitchess* as an insulting utterance, but the harmfulness score of the instance is 0.

Below are the categories to be annotated in text spans. Most of these forms were inspired by **social studies on cyberbullying** (Vandebosch et al., 2006; Vandebosch and Van Cleemput, 2009). Generally, a distinction is made between direct (e.g., *flaming*, threats) and indirect forms of cyberbullying (e.g., *masquerading*¹, *outing*², *popularity polls*). This annotation methodology focusses on a number of cyberbullying-related text categories that are considered relevant and applicable to our dataset. Most of these categories are related to **direct forms of cyberbullying** (as defined by Vandebosch et. al (2006)) and one that is related to *outing*, an indirect form of cyberbullying, namely *Defamation*. Additionally, a number of subcategories are defined to make the annotation scheme as concrete as possible (e.g., *Discrimination* as a subcategory of *Insult*).

The cyberbullying-related text categories are listed below. An example post is given for each category.

4.1 Threat or blackmail

This category contains expressions of physical or psychological threats towards the addressee and expressions indicating blackmail.

¹By *masquerading*, one understands all attempts of a person to steal the victim’s identity online (e.g., creating a fake profile with the victim’s name) (Vandebosch et al., 2006).

²*Outing* is considered a form of cyberbullying where the bully reveals private or embarrassing information about the victim to a large public (Vandebosch et al., 2006).

(9) *My fist is itching to punch you so hard in the face.*

(10) *Just do what I asked, or I'll post a naked picture of you.*

4.2 Insult

Expressions containing abusive, degrading or offensive language in order to insult the addressee. An insult can be classified into one of the following subcategories:

- **General Insult:** Expressions that insult or offend the victim.

(11) *You're a sad little fuck.*

- **Attacking Relatives and Friends:** Expressions that insult relatives or friends of the victim.

(12) *Your mother is so fat that she wouldn't even fit in the Grand Canyon!!!*

- **Discrimination:** Expressions of unjust or prejudicial treatment of the victim. Two types of discrimination are distinguished (i.e., sexism and racism). Other forms of discrimination should be categorized as general insults.

- **Sexism:** Expressions with a sexist nature, such as prejudice or discrimination based on the victim's sex, gender or sexual orientation.

(13) *You'd better shut up and return to ur kitchen.*

- **Racism:** Expressions of discrimination that are based on the victim's race, skin color, ethnicity, nationality, or religion.

(14) *Just shut up you're a fucking jew!*

4.3 Curse or 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.

(15) *Hopefully you'll burn in hell.*

(16) *Just kill yourself, nobody likes you.*

4.4 Defamation

Defamations are expressions that reveal confident, embarrassing or defamatory information about the victim to a large public. This category is related to the indirect form of cyberbullying that is known as *outing*.

(17) *She's a slut and she will influence you to be one*

(18) *I've heard his father lost his job bc he's an alcoholic.*

4.5 Sexual Talk

This category contains expressions that have a sexual meaning or connotation. A distinction is made between harmless and harmful sexual talk (i.e., sexual harassment).

- **Harmless Sexual Talk:** Expressions with a sexual meaning, such as sexting between equals.

(19) I wanna kiss you, and more than that.

- **Sexual Harassment:** Expressions with a sexual meaning that have a compelling character and that are considered undesirable (e.g., unwanted requests to talk about sex or to do something sexual).

(20) Post a naked pic, now!!

4.6 Defense

Expressions in support of the victim. These can be uttered by the victim himself or by a bystander.

- **Bystander Defends the Victim:** Expressions by which a bystander shows support for the victim or discourages the harasser from continuing his actions.

- **General Victim Defense:** General expressions in support of the victim.

(21) Shut up about my sister, she is not a slut!

- **Good Characteristics:** Expressions that specify positive characteristics of the victim.

(22) Just stop, she's such a nice girl!

- **Victim Defends Himself:** Assertive or powerless reactions from the victim.

- **Assertive Self-defense:** Expressions of disapproval and 'fighting back'.

(23) do you realize how much words hurt you dumb bitch leave me the fuck alone.

- **Powerless Self-defense:** Expressions showing the victim's indignation and helplessness.

(24) And why do u hate me? What have I done to you?

4.7 Encouragement to the Harasser

Expressions in support of the harasser.

(25) Haha, you're soo right, hes a nobody.

4.8 Other

This category is provided for cases that contain any other harmful utterance than the ones described above, or when annotators are not sure which category to assign a certain post to.

Note: Whenever sarcasm is contained by one of the aforementioned categories, annotators can mark this. An example of this is sentence (26).

(26) *Wow, that's a nice pic of you.. horse teeth!!!*

Practical remarks

- **Other language:** Posts that are written in another language than English should be marked through an annotation at the post level (i.e., on the dummy token ¶). Note that posts written in another language than English do not need to be annotated further.
- **Text span:** A *text span* is that part of the text that relates to the action under investigation. One text span should correspond with one utterance or proposition. This can be a word (e.g., a curse word like *bitch!*), a phrase (e.g., an insulting noun phrase like *you sad little fuck*) or an entire sentence (e.g., a threat like *If you don't shut up right now, you'll be dead tomorrow*). As a convention, the **subject and verb** of a sentence part to be categorized (when available) should be included in the annotated text span. Text spans do not normally cross sentence boundaries. **Conjunctions** that connect words, sentences or clauses (e.g., *and, or, because*) should be annotated as well, they are included in the text span they introduce. **Punctuation marks** at the end of the text span should be included in the annotated text span. **Newline characters** within a message have been replaced by pipes or vertical bars and need not be annotated.
- **Double annotations:** Double annotations are possible. Whenever two cyberbullying related categories can be identified in a text span (e.g., a defensive statement that includes an insult), they should both be annotated (see for instance example (28)).

Chapter 5

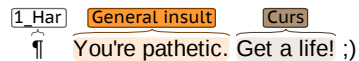
Annotation Procedure and Examples

In this chapter, we describe the different steps in the annotation procedure. Essentially, the annotation scheme describes two levels of annotation. Firstly, at **post level** annotators indicate the **harmfulness** of a post on a three-point scale (0-1-2). If the post is considered harmful (i.e., its harmfulness score is 1 or 2), annotators indicate the **author's role** in the cyberbullying event (i.e., harasser, victim, bystander-defender, or bystander-assistant). In a second step, at the (sub)sentence level annotators mark phrases where different **textual categories related to cyberbullying** (e.g., threats, insults, defenses, etc.) are expressed, even when the post itself is not considered harmful (see for example sentence (2)). Remember that neutral text is to be left unannotated.

There are no restrictions as to what form these annotations can take. They can be adjectives, noun phrases, verb phrases, and so on. The only condition is that the annotation **cannot span more than one sentence and less than one word**. Moreover, **double annotations are possible** (see example (28) below). However, annotators should try to annotate each phrase separately. Conjunctions like *and*, *or* are included in the text span they correspond to, as well as punctuation marks. Emoticons need not be annotated. Below are given some annotation examples in brat.

Note: Posts that are written in **another language** than the corpus language (i.e., Dutch or English) should be marked through an annotation at the post level (i.e., on the dummy token ¶), but no further annotations are required for these posts.

(27)


¶ You're pathetic. Get a life! ;)

- Harmfulness score = 1, the message contains indications of cyberbullying, but they are not severe.
- Author's role = harasser
- *You're pathetic* = general insult
- *Get a life!* = curse/exclusion

(28)

1 Vic Assertive self-Defense General Insult
Assertive self-Defense
↑ Look at yourselves you'r 1000000000000 times uglier than me

- Harmfulness score = 1
- Author's role = victim
- *Look at yourselves* and *you'r 1000000000000 times uglier than me* = expressions of the victim 'fighting back' (i.e., assertive self-defense).
- *you'r 1000000000000 times uglier than me* is also a general insult.

(29)

1 Bystander_defender General victim defense General victim defense
↑ Omg you racist or however it is written just let her be happy?

- Harmfulness score = 1
- Author's role = bystander-defender
- *Omg you racist or however it is written* and *just let her be happy?* = general statements in defense of the victim.

(30)

2 Har Threat or Blackmail
↑ I'm gonna punch you so hard in the face tomorrow.

- Harmfulness score = 2, the message contains serious indications of cyberbullying.
- Author's role = harasser
- *I'm gonna punch so hard in the face tomorrow* = threat

(31)

1 Har Attacking relatives
↑ Lets be honest, your mom is pretty much the BIGGEST skank ever!

- Harmfulness score = 1
- Author's role = harasser
- *your mom is pretty much the BIGGEST skank ever!* = insult of a family member of the victim

(32)

2_Har Curs General insult General insult
¶ Kill yourself. No one loves you, you're a used piece of shit.

- Harmfulness score = 2
- *Kill yourself* = curse/exclusion
- *No one loves you* and *you're a used piece of shit* = general insult

(33)

1_Har Defamation
¶ I heard I heard you get touched by your dad.

- Harmfulness score = 1
- Author's role = harasser
- *I heard I heard you get touched by your dad* = expression revealing defamatory information about the victim.

(34)

1_Har Sarcasm General insult General insult
¶ Wow, you look really hot. Big freakin horseteeth.

- Harmfulness score = 1
- Author's role = harasser
- *Wow, you look really hot* = general insult
- *Big freakin horse teeth* = general insult

(35)

1_Vic Powerless self-Defense Powerless self-Defense
¶ Why are you guys so mean? What did I do wrong?

- Harmfulness score = 1
- Author's role = victim
- *Why are you guys so mean?* and *What did I do wrong?* = powerless self-defense.

(36)

1_Bystander_defender GenDef Good characteristics General victim defense
↑ no shes not ugly, she is my beautiful sister so leave her alone.

- Harmfulness score = 1
- Author's role = bystander-defender
- *no she's not ugly* = general victim defense.
- *she is my beautiful sister* = specifying good characteristics of the victim.
- *so leave her alone* = general victim defense.

(37)

Harmless sexual talk
↑ omg i wish i could kiss you.

- Harmfulness score = 0, no author role should be indicated
- *omg i wish i could kiss you* = harmless sexual talk

(38)

2_Har Sexual harassment
↑ Send me a naked pic, now!!

- Harmfulness score = 2
- Author's role = harasser
- *Send me a naked pic, now!!* = sexual harassment

(39)

1_Vic AssDef General insult Assertive self-Defense
↑ Leave me alone. you're obviously some jealous ass bitch

- Harmfulness score = 1
- Author's role = victim
- *Leave me alone* and *you're obviously some jealous ass bitch* = assertive self-defense
- *you're obviously some jealous ass bitch* = also general insult

(40)

GenIn
¶ Hey bitches, feel like seeing a movie tonight?

- Harmfulness score = 0, no author role should be indicated
- *bitches* = insult

(41)

1_Har Curse or Exclusion Racism
¶ Go back to where you came from, you stupid muslim.

- Harmfulness score = 1
- Author's role = harasser
- *Go back to where you came from* = curse/exclusion
- *you stupid muslim* = racist insult

(42)

1_Bystander_assistant Encouraging harasser General insult
¶ Haha, soo right! she is an ugly kid

- Harmfulness score = 1
- Author's role = bystander-assistant
- *Haha, soo right!* = encouragement to the harasser
- *she is an ugly kid* = general insult

Chapter 6

Using the Brat Annotation Tool

Using brat for text annotation is intuitive. As mentioned before, this scheme describes two annotation levels: 1) the **post level** and 2) the **(sub)sentence level**. Annotations at the post level should be made on the **pilcrow sign ¶** preceding each post, which functions as a dummy-token. Annotations at the (sub)sentence level are made on specific text spans within the post.

In brat, selecting or double-clicking on the dummy token ¶ will open a window where the annotators can 1) define a **harmfulness score** for the post and 2) specify the **author's role**. Example (43) shows how this can be annotated.

43)

Harmfulness_Role

1

- 1_Harasser
- 1_Victim
- 1_Bystander_defender
- 1_Bystander_assistant

2

- 2_Harasser
- 2_Victim
- 2_Bystander_defender
- 2_Bystander_assistant

As is shown above, when a post is not considered harmful (i.e., its score is 0), no author role should be indicated.

Similarly to the annotation at the post level, selecting a piece of text at the (sub)sentence level will open a window where the annotators can select the corresponding category for that text span from a list. Parent categories are implicitly marked when a child category of them is selected, as is shown in example (44).

44)

The screenshot shows a window titled "Entity type" with a list of categories. The categories are: "Threat or Blackmail", "Insult", "Discrimination", "Curse or Exclusion", "Defamation", and "Sexual talk". Each category has a radio button next to it. "Insult" and "Discrimination" are expanded, showing sub-categories: "General insult", "Attacking relatives", "Sexism", and "Racism". "Racism" is selected, indicated by a blue dot in the radio button. The text "Racism" is highlighted in orange. Other categories and sub-categories are highlighted in various colors: "Threat or Blackmail" (red), "General insult" (orange), "Attacking relatives" (orange), "Sexism" (orange), "Curse or Exclusion" (grey), "Defamation" (yellow), "Harmless sexual talk" (purple), and "Sexual harassment" (purple).

All annotations that are added to a document are automatically saved. The annotators can proceed with the next post or return to a previously annotated post by using the right and left arrow buttons in the upper left corner of the screen.

References

- Dooley, Julian J., Jacek Pyzalski, and Donna S. Cross. 2009. Cyberbullying Versus Face-to-Face Bullying: A Theoretical and Conceptual Review. *Zeitschrift Fur Psychologie-journal of Psychology*, 217(4):182–188.
- Grigg, Dorothy Wunmi. 2010. Cyber-Aggression: Definition and Concept of Cyberbullying. *Australian Journal of Guidance and Counselling*, 20(2):143–156.
- Kiesler, Sara, Jane Sigel, and W.Timothy McGuire. 1984. Social psychological aspects of computer-mediated communication. *American Psychologist*, 39(10):1123–1134.
- Olweus, Dan. 1993. *Bullying at school: What we know and what we can do*. Malden, MA: Blackwell Publishing.
- O’Sullivan, Patrick B. and Andrew J. Flanagin. 2003. Reconceptualizing ‘flaming’ and other problematic messages. *New Media & Society*, 5(1):69–94.
- Price, M. and J. Dalglish. 2010. Cyberbullying: Experiences, Impacts and Coping Strategies as Described by Australian Young People. *Youth Studies Australia*, 29(2):51–59.
- Salmivalli, Christina. 2010. Bullying and the peer group: A review. *Aggression and Violent Behavior*, 15(2):112–120.
- Salmivalli, Christina, Marinus Voeten, and Elisa Poskiparta. 2011. Bystanders Matter: Associations Between Reinforcing, Defending, and the Frequency of Bullying Behavior in Classrooms. *Journal of Clinical Child & Adolescent Psychology*, 40(5):668–676.
- Smith, Peter K., 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.
- Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun’ichi Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, pages 102–107, Avignon, France.
- Tokunaga, Robert S. 2010. Review: Following You Home from School: A Critical Review and Synthesis of Research on Cyberbullying Victimization. *Computers in Human Behavior*, 26(3):277–287.
- Vandebosch, Heidi and Katrien Van Cleemput. 2009. Cyberbullying among youngsters: profiles of bullies and victims. *New Media & Society*, 11(8):1349–1371.
- Vandebosch, Heidi, Katrien Van Cleemput, Dimitri Mortelmans, and Michel Walrave. 2006. Cyberpesten bij jongeren in Vlaanderen: Een studie in opdracht van het viWTA. On-line.
- Šléglová, Veronika and Alena Černá. 2011. Cyberbullying in Adolescent Victims: Perception and Coping. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 5(1):2.
- Willard, Nancy E. 2007. *Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress*. Research Press.
- Xu, Jun-Ming, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from Bullying Traces in Social Media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT ’12*, pages 656–666, Stroudsburg, PA, USA. Association for Computational Linguistics.