2019

The Role of Previous Discourse in Identifying Public Textual Cyberbullying

Aurelia Power
Technological University Dublin

Anthony Keane
Technological University Dublin, anthony.keane@tudublin.ie

Brian Nolan
Technological University Dublin

See next page for additional authors

Recommended Citation

Power, A., Keane, A., Nolan, B., & Neill, B. O. (2019). The Role of Previous Discourse in Identifying Public Textual Cyberbullying.*Journal of Computer-Assisted Linguistic Research*, 3, 1–20. doi:https://10.4995/jclr.2019.11013

Power, Aurelia, Keane, Anthony, Nolan, Brian, and O'Neill, Brian (2019) The Role of Previous Discourse in Identifying Public Textual Cyberbullying. Journal of Computer-Assisted Linguistic Research 3: 1-20.

The Role of Previous Discourse in Identifying Public Textual Cyberbullying

Aurelia Power*, Anthony Keane, Brian Nolan, Brian O'Neill
Technological University Dublin, Ireland
*Corresponding author: aurelia.power@itb.ie

Received: 10 December 2018 / Accepted: 26 February 2019 / Published: 15 July 2019

Abstract
In this paper we investigate the contribution of previous discourse in identifying elements that are key to detecting public textual cyberbullying. Based on the analysis of our dataset, we first discuss the missing cyberbullying elements and the grammatical structures representative of discourse-dependent cyberbullying discourse. Then we identify four types of discourse dependent cyberbullying constructions: (1) fully inferable constructions, (2) personal marker and cyberbullying link inferable constructions, (3) dysphemistic element and cyberbullying link inferable constructions, and (4) dysphemistic element inferable constructions. Finally, we formalise a framework to resolve the missing cyberbullying elements that proposes several resolution algorithms. The resolution algorithms target the following discourse dependent message types: (1) polarity answers, (2) contradictory statements, (3) explicit ellipsis, (4) implicit affirmative answers, and (5) statements that use indefinite pronouns as placeholders for the dysphemistic element.

Keywords: cyberbullying, discourse, grammatical structures, natural language processing.

1. Introduction

Cyberbullying has been recognised as a predominant behavioural issue among young people using the Internet (Livingstone et al. 2011; Livingstone et al. 2014). Its negative emotional impact on the psychological wellbeing of the victim(s) can lead to serious pathological problems, such as depression, self-harm, suicide ideation, and suicide attempt (Sourander et al. 2010). Similar to face-to-face bullying, cyberbullying instances must satisfy three fundamental criteria: intention of harm, repetition, and power imbalance between the victim and the bully (Hinduja and Patchin 2009). In addition, cyberbullying instances must occur in the cyberspace which allows these acts to transcend the physical and temporal constraints that apply to face-to-face bullying. Because of the ability of cyberspace to persist over time, to be searched for repeatedly, to be replicated numerous times, and to multicast to potentially large invisible audiences (Boyd

DOI: https://doi.org/10.4995/jclr.2019.11013
PUBLISHER: UNIVERSITAT POLITÈCNICA DE VALÈNCIA.

2007), cyberbullying acts implicitly satisfy the criterion of repetition (Dooley and Cross 2010; Grigg 2010; Langos 2012; Power 2017; 2018).

While cyberbullying detection has recently received increased attention, with much of the research using techniques from text analytics and Natural Language Processing (NLP), the majority of related work relies on the explicit presence of features such as as profanities (Yin et al. 2009; Dinakar et al. 2012; Dadvar et al. 2013; Al-garadi et al. 2016), bad words (Reynolds et al. 2011; Huang et al. 2014), foul terms (Nahar et al. 2013), bullying terms (Kontostathis et al. 2013; Nandhini and Sheeba 2015), pejoratives and obscenities (Chen et al. 2012), emotemes and vulgarities (Ptaszynski et al. 2010; Ptaszynski et al. 2016), curses (Chatzakou et al. 2017) or negative words (Van Hee et al. 2015). Even those studies that target implicit forms of cyberbullying (Chen et al. 2012; Dinakar et al. 2012; Nitta et al. 2013; Ptaszynski et al. 2010; Ptaszynski et al. 2016) do not clearly define the boundaries of what is cyberbullying and the approaches described result in a considerable amount of false positives[1] that contain rude and violent language, despite the fact that the use of this type of language does not constitute cyberbullying on its own. On the other hand, although the focus of previous research in the field of cyberbullying detection has been to reduce the number of false negatives[2], no other study has investigated the linguistic role of prior discourse in identifying public textual cyberbullying, and the purpose of the present paper is to determine the contribution of antecedent messages in resolving the missing cyberbullying elements that we previously proposed (Power et al. 2017; 2018): the personal marker, the dysphemistic element, and the cyberbullying link between the personal marker and the dysphemistic element.

2. Public textual cyberbullying and discourse

We base the present investigation on the same view that we expressed previously (Power et al. 2017; 2018) that the presence of explicit terms/expressions does not suffice for a message or post to be classified as public textual cyberbullying; it must be linked to or it must target a specific person, or group of people. Thus, an instance constitutes public textual cyberbullying if it contains (either explicitly or implicitly) the personal marker/pointer - which identifies or points to the victim(s), the dysphemistic element - which is defined by Allan and Burridge (2006, 31) as the "word or phrase with connotations that are offensive either about the denotatum and/or to people addressed or overhearing the utterance", and the link between the previous two elements - which captures how the dysphemistic element targets the victim(s) identified or pointed to by the personal marker.

The three necessary and sufficient elements can be explicitly present in a given instance such as in the sentences *You are a cunt* and *You are not smart*. In these cases, the detection process identifies the personal markers (the personal pronoun *you*), the dysphemistic elements (the profane noun *cunt* and the adjectival phrase *not smart* that uses negation to invert the positive

[1] In the context of our research, false positives represent instances that have been incorrectly assigned the label of cyberbullying by a detection system.

[2] Conversely, we consider false negatives those instances that have been incorrectly assigned the label of not-cyberbullying by a detection system.

connotations associated with the adjective *smart*), and the links between them (expressed in both cases as the conjugated form of the copular verb *be*). However, in other cases, all three or some cyberbullying elements need to be inferred from previous discourse[3], that is, from previous instances[4] that are part of the same online conversation/dialogue[5]. For example, the following instances from the dataset used for this research were labelled by annotators as cyberbullying, despite the fact that they appear at first to be harmless:

(1) a. Yes.
b. Not you.
c. Pig.

Even the third instance which contains explicitly the animal dysphemistic element cannot be labelled as cyberbullying unless the previous online discourse or message is considered, since it can be an answer to the question *What animal do you like best?*, or it could indeed be a metaphoric mapping from the animal domain to that of people intended to hurt someone. In all these instances, some of (or even all) the three necessary and sufficient cyberbullying elements are not explicitly present, and, only after we considered the online conversation and identified their respective previous messages/posts, we were able to understand why they qualify as cyberbullying, as shown in (2):

(2) a. U1[6]: I used to cut and harm. am I a freak???
U2: Yes.
b. U1: Top 10 gorgeous guys?
U2: Not you.
c. U1: you are a …
U2: What?
U1: Pig.

In the case of (2 a.), the *yes* answer is labelled as public textual cyberbullying because it stands for the complete sentence *you are a freak*, sentence that is inferable from the previous rhetorical question *am I a freak?*, as well as from the answer *yes* which affirms the proposition put forward in the question. Specifically, the previous instance contains the explicit dysphemistic element *freak*, as well as the personal marker and the cyberbullying link. Similarly, in the case of (2 b.), the answer can be labelled as public textual cyberbullying because it is a negation based instance which can be elaborated as follows: *you are not one of the top 10 gorgeous guys*. Such extension is an inference that can be drawn based on two elements: first, the previous message uttered by the first user is a question that contains the inferable information - the positive connotation term *gorgeous* and the personal marker *guys* which is a person-referring noun, and,

[3] According to Agne and Tracy (2009), discourse, in a broader sense, can refer to entire online conversations or dialogues, as well as single sentences. However, in this research, we limit previous discourse to refer only to distinct previous messages/posts.

[4] We grouped online messages and posts under the umbrella of the term *instances*.

[5] Dialogue and conversation are treated as synonyms in the context of the present research, since both represent free-flowing talk among two or more people (Agne and Tracy, 2009).

[6] For privacy reasons, we replaced the actual user names with indexed acronyms, such as U1 or U2.

secondly, the second user's answer *not you* which negates the positive characteristic of being gorgeous in relation to the first user denoted by the personal pronoun *you*. Unlike the previous example, the instance uttered by the first user does not contain explicitly the link verb; however, the link verb can be inferred from the sentential structure of the first user's utterance which can be rephrased as a full copular structure: *Who are the top 10 gorgeous guys?* Finally, in the case of (2 c.), U1's first utterance provides the old discourse inferable information that is needed to qualify the last instance as public textual cyberbullying. And, although the message *Pig* contains the potential dysphemistic element, it cannot be considered cyberbullying until it is related to a personal marker; thus, the first utterance provides the two missing cyberbullying elements: the personal marker embodied in the personal pronoun *you*, and the cyberbullying link denoted by the conjugated copular verb *are*. As a result, the second instance uttered by the first user can be rephrased as an animal metaphor-based cyberbullying instance: *you are a pig*.

3. Dataset and Pre-processing

To analyse discourse-dependent instances, we used the same dataset as in previous work to which the same cleaning and pre-processing techniques were applied (Power et al. 2018). The dataset was acquired from two sources: from Kavanagh' research (2014) on cyberbullying detection and from Hosseinmardi et al. (2014a) and Hosseinmardi et al. (2014b) dataset from which we randomly selected a continuous portion. The dataset consists of a total of 2038 instances that originate from the conversations (organised as pairs of questions and answers) of 16 users on ASK.fm. To label the dataset, two individuals were asked to annotate each instance using one of the following two categories: cyberbullying (CB) and not-cyberbullying (NCB). For those instances for which the annotators disagreed, a third individual was then asked to label them, and the label provided by the third individual constituted the final label. The results of the data labelling process show that approximately 23.30% of instances were labelled as cyberbullying, from which 3% constitute discourse dependent instances.

To efficiently process the dataset, we removed any xml and html tags, but kept the usernames, which subsequently were replaced by acronyms such as U1 or U2. We have also removed any non-English instances, and from those retained, we have removed all html links. In addition, we inversed the order of the question-answer pairs, from the most recent pair to the least recent pair to reflect the conversational order. Subsequently, we applied several commonly-used NLP techniques such as tokenization, case transformation, and lemmatisation (Navarro and Ziviani 2011). We applied further pre-processing techniques to obtain additional information that aids the detection process and to increase the accuracy level of the dependency parser. Such techniques include replacing acronyms[7], abbreviations, words where some characters were omitted, and words that contain digits to substitute groups of letters with their full form using an online dictionary (InternetSlang 2019). For example, *bj* was replaced by *blow job*, *anon* was replaced by *anonymous*, *fck* was replaced by *fuck*, and *h8* was replaced by *hate*. We also replaced all informal variations of the personal pronouns with their corresponding formal form; for instance, *u* and *ya* were replaced by *you*, while informal reflexive forms such as *yerself* or *meself*

[7] We also include here initialisms.

were replaced by *yourself* and *myself*, respectively. Other errors including repeated letters, meaningless symbols, transposition, missing, and wrong characters were addressed by using Norvig's spelling algorithm (2007). Finally, we have discarded any sequence of characters that did not have a sense or for which the acronym mapping procedure could not find a corresponding value, or the autocorrection tool could not resolve, as well as any icons, smileys, and emoticons; for example, sequences such as 'ahahh' or 'zsss' were removed on the basis that they did not appear in the acronyms/abbreviations dictionary, nor could they be autocorrected in a meaningful way.

4. Discourse-Old Cyberbullying Elements

To analyse discourse-dependent instances, we apply the information paradigm proposed by Prince (1981) who divides information into discourse-old and discourse-new. However, our focus is on discourse-old information, that is, how such information can be used to infer some or all three necessary and sufficient cyberbullying elements.

The personal marker is not required to be explicitly present in discourse dependent instances and the inference mechanism uses Prince's framework (1981) to represent the personal marker as a discourse-old entity or, in the case of a possessive modifier, to attach it to a discourse-old entity. Based on the development dataset, several forms of public textual cyberbullying in which the personal marker is a linguistic entity that must be inferred from the previous discourse were found. First, the personal marker must be inferred from the previous discourse in the case of yes/no answers where the dysphemistic element is also missing, as seen in the case of (1 a. and 2 a.). Secondly, the personal marker must be inferred from the previous discourse where the dysphemistic element is explicitly present as a metaphoric element as is the case with (1 c. and 2 c.). Additionally, the dysphemistic element can be explicitly present in a previous instance as a dysphemistic adjective that has a wide range of applicability, such as *ugly* or *awful*, which on its own cannot be labelled as cyberbullying. Take, for instance the following snippet of conversation:

(3) U1: Saw Ella[8]?
U2: Ugly!

The instance *Ugly!* can be labelled cyberbullying because the previous instance contains the personal marker (in the form of proper name *Ella*) to which the adjective *ugly* can be inferred to be applied, since the proper name is the only item that grammatically can be modified by *ugly* in that context. Ultimately, the resolution of the personal marker element is a reference issue that is typically associated with *who, whose,* or *to whom* questions, for example, *who is a freak?*, or *who is a pig?,* or *who is ugly?*

The dysphemistic element can be an inferable discourse-old entity too, as demonstrated in (1 a. and 2 a.) and (1 b. and 2 b.) Additional examples include the following instances:

[8] The name *Ella* was used here to replace the actual name to avoid any potential identity and privacy related issues.

(4) a. U1: Is she pretty?
U2: No.
b. U1: Who is stupid?
U2: you.

In the case of (4 a.), both elements, the personal marker and the dysphemistic element, must be inferred from the previous message. As such, the second user's answer can be replaced with the more elaborate answer *she is not pretty*. But, in the case of (4 b.), the second user's answer contains explicitly the personal marker (the personal pronoun *you*), but the dysphemistic element is a discourse-old entity - the adjective *stupid* - inferable from first user's question; in this case, U2's utterance can be resolved to the full copular instance: *you are stupid*.

Finally, while discourse dependent cyberbullying instances can contain the personal marker or the dysphemistic element on its own, they cannot contain the cyberbullying link on its own. Even when the cyberbullying link is explicitly present as the dysphemistic element itself (the case of reflexive links where the link is an explicit intransitive dysphemistic verb, such as *die*), the instance is not in fact a discourse dependent instance, since it explicitly contains the dysphemistic element and the cyberbullying link, while the personal marker can be inferred from its imperative structure, and not from previous messages. In addition, those instances that have copular and transitive sentential structures would not be considered acceptable utterances, if they only contained the conjugated copular and transitive verbs, respectively. Consider, for example, the following adapted instances[9]:

(5) a. U1: I am not pathetic.
U2: are*.
b. U1: My mom doesn't deserve death.
U2: Deserves*.

In (5), both statements that belong to the second users – *are* and *deserves*, respectively – are not acceptable utterances because they are grammatically incomplete sentences that do not contain enough semantic information. In the case of the copular construction (5 a.), the copular verb must link a subject to a complement (but both are missing), while in the case of the transitive construction (5 b.), the transitive verb should have two arguments, the subject and the object, but they are also both missing. To be acceptable utterances, both must contain at least the subject as follows:

(6) a. U1: I am not pathetic.
U2: You are.
b. U1: My mom doesn't deserve death.
U2: She does.

[9] We modified here the second users'statements for the purpose of demonstrating their inadequacy.

As a result, when analysing discourse dependent instances, the cyberbullying link resolution cannot be considered on its own, but always related to the resolution of the personal marker and/or the dysphemistic element.

5. Discourse Dependent Public Textual Cyberbullying Constructions

Once all the cyberbullying elements are inferred from previous discourse and the cyberbullying form identified, the same detection mechanism that we described in previous work (Power et al 2018) can be applied to any instance. However, to identify the missing elements we first need to consider the types of grammatical relations that represent them and because the cyberbullying link cannot appear on its own in discourse dependent instances, but it must be accompanied at least by its grammatical subject, there are only four discourse dependent cyberbullying constructions that we need to discuss: (1) fully inferable constructions – where all three cyberbullying elements, the personal marker, the dysphemistic element, and the link between them, are not explicitly present, but can be inferred from previous discourse, (2) personal marker and cyberbullying link inferable constructions – where the dysphemistic element is explicitly present, but the personal marker and the link must be inferred from previous discourse, (3) dysphemistic element and cyberbullying link inferable constructions – where the personal marker is explicitly present, but the dysphemistic element and the cyberbullying link are entities inferable from previous discourse, and (4) dysphemistic element inferable constructions – where the personal marker and the link are explicitly present, but the dysphemistic element must be inferred from prior discourse.

5.1. Fully Inferable Constructions

Fully inferable constructions are those that contain no cyberbullying element in an explicit manner, and, typically, they are represented by yes/no answers to polarity questions[10] (Krifka, 2011). In such instances, using Prince's paradigm (1981), all the cyberbullying elements are discourse-old information, and they are inferred from both, the question and the answer to the question, respectively. Some examples were shown earlier and are reconsidered here for further discussion:

(7) a. U1: I used to cut and harm. am I a freak???
U2: Yes.
b. U1: Is she pretty?
U2: No.

In the case of (7 a.), the second user's answer was labelled as cyberbullying because it can be rephrased as *you are a freak* which is an inference that draws from three sources: (1) the affirmative answer *yes* which precludes the existence of the negation relation and confirms the proposition put forward in the question, (2) the copular sentential structure *am I a freak* where *I* is the subject, *freak* is the subject complement, and *am* is the conjugated copular verb, and (3) the

[10] From a functional perspective, questions are functions, while their respective answers provide the arguments for such functions. As such, in the case of polarity questions, the answer confirms or negates the proposition put forward in the question – λp p or λp ¬p, respectively (Krifka, 2001).

change of conversational user from U1 to U2, which allows the subject *I* to be replaced with the personal pronoun *you*, and the conjugated copula *am* with *are*. The resulting inference *you are a freak* is an explicit form of cyberbullying that contains the personal marker (the pronoun *you*), the dysphemistic element (the offensive noun *freak*), and the cyberbullying link between them (the conjugated copula *are*).

Similarly, in the case of (7 b.), the second user's answer was labelled as cyberbullying because it can be rephrased as *she is not pretty* which is an inference that draws from two sources: (1) the negative answer *no* which uses the negation relation to invert the truth value of the proposition put forward in the question and (2) the question *is she pretty?* where *she* is the subject, *pretty* is the subject complement, and *is* is the conjugated copula. Again, once all the cyberbullying elements are present, the resulting inference *she is not pretty* can be treated as an implicit negation-based cyberbullying instance, where the personal marker is the pronoun *she*, the dysphemistic element is realised by negating the adjective *pretty*, and the link between them is the conjugated verb *is*.

5.2. Personal Marker and Cyberbullying Link Inferable Constructions

These discourse dependent cyberbullying forms do not contain the personal marker and the cyberbullying link in an explicit manner, and they must be inferred from previous messages. However, the dysphemistic element must be explicitly present. We reconsider below some previous examples:

(8)
a. U1: You are a …
U2: What?
U1: Pig.
b. U1: Saw Ella?
U2: Ugly.

In the case of (8 a.), the first user's answer *Pig* was labelled as cyberbullying, despite the fact that it contains no personal marker, or the cyberbullying link. However, they are both inferred from the first user's first utterance *You are a …*, which also contains the elliptical punctuation marks that signal that the sentence awaits completion. In such cases, the full sentence can be inferred by simply replacing the elliptical punctuation marks with the second utterance as follows: *You are a pig*. The resulting inference can then be treated as an animal metaphoric cyberbullying instance where the personal marker is represented by the personal pronoun *you*, the metaphoric dysphemistic element is the animal dysphemistic noun *pig*, and the cyberbullying link is the conjugated copula *are*.

In the case of (8 b.), the second user's answer was also labelled cyberbullying, although it contains no personal marker, or the cyberbullying link. However, it contains the explicit dysphemistic adjective *ugly* which is inferred to be applied to the personal marker *Ella* because it is contained in the first user's question which immediately precedes the answer *Ugly!* In addition, since the answer contains no negation trigger that targets the conjugated verb *saw*, it is a confirmation of the fact that the second user saw *Ella*. As a result, the inference drawn is: *I saw*

Ella and she is ugly. However, for the purpose of cyberbullying detection one needs to consider only the second clause where the personal pronoun *she* anaphorically refers to *Ella*, the direct object of *saw*. The final inference – *Ella is ugly* – can now be treated as an explicit form of cyberbullying which contains the personal marker *Ella*, the dysphemistic element *ugly*, and the cyberbullying link in the form of the conjugated copular verb *is*.

5.3. Dysphemistic Element and Cyberbullying Link Inferable Constructions

These constructions are the most common type of discourse dependent instances that we found in the present dataset and they do not contain the dysphemistic element or the cyberbullying link in an explicit manner, however, the personal marker is explicitly present. They typically constitute answers to *wh* questions[11] (Krifka, 2011), or contradictory statements. Some examples are shown below in (11.9):

(9)
a. U1: Who is stupid?
   U2: You.
b. U1: Ella looks beautiful.
   U2: No she doesn't.

In the case of (9 a.), the second user's answer you was labelled as a cyberbullying instance, despite the fact that it contains explicitly no dysphemistic element, although, it contains explicitly the personal marker *you* which can replace the pronoun *who*. Furthermore, because an answer to a *wh* question implies all the grammatical elements evoked in the question (Krifka, 2001), such as the subject (here replaced by *you*), the copular verb *be* (here in the conjugated form *are*), and the complement *stupid*, the second user's answer *you* can be rephrased as *you are stupid*. The resulting inference can then be treated as an explicit form of cyberbullying where the personal marker is realised by means of the personal pronoun *you*, the dysphemistic element is the offensive adjective *stupid*, and the cyberbullying link is represented by the conjugated copula *are*.

In (9 b.), the second user's statement constitutes a contradictory statement which inverts the propositional truth of the first user's statement by using negation. Like in the previous example, the dysphemistic element is not explicitly present, whereas the personal marker is represented by the third person pronoun *she*. The inference can be drawn based on several elements: (1) the use of the contraction *doesn't* which indicates that the dysphemistic element is realised by means of negation, (2) the explicit presence of the personal marker *she*, (3) the positive connotations term *beautiful* contained in the previous message, and (4) the verb *look* which is negated using the auxiliary conjugated contraction *doesn't* and which takes as complement the adjective *beautiful*. The resulting inference – *she doesn't look beautiful* – can now be viewed as a negation-based instance of cyberbullying.

[11] According to Krifka (2001), *wh* questions require answers that fill in a constituent for the *wh* word in the question.

5.4. Dysphemistic Element Inferable Constructions

These cyberbullying constructions do not contain the dysphemistic element in an explicit manner which must be inferred from previous discourse, but the personal marker and the cyberbullying link are both explicitly present. These constructions can take various forms: they can be statements that use indefinite pronouns as placeholders for the dysphemistic element or they can be contradictory statements intended to contradict the propositional truths of previous messages. Some examples are shown in (10):

(10)
a. U1: My favourite animals are monkeys.
U2: you look like one!
b. U1: I'm not pathetic.
U2: You clearly are.

In (10 a.), the dysphemistic element of the second user's statement can be inferred using the information from the previous message which contains the potential dysphemistic element in the form of the noun *monkeys* that can replace the indefinite pronoun *one*. As such, the resulting inference – *you look like monkeys* – can then be treated as an animal metaphoric instance where the personal marker is the second person pronoun *you*, the dysphemistic element is the noun *monkeys*, and the cyberbullying link is represented by the conjugated copular verb *look*.

The second user's statement in (10 b.) constitutes a contradictory statement which inverts the propositional truth of the first user's statement by using affirmation. Like in the previous example, the dysphemistic element is not explicitly present, whereas the personal marker is present as the second person pronoun you, as well as the verb link are. The inference process uses several elements: (1) the affirmation to indicate that the dysphemistic element is not realised by means of negation, (2) the explicit presence of the personal marker you, and (3) the previous message which contain the relevant offensive adjective pathetic. The resulting inference - you clearly are pathetic – can be treated as an explicit form of cyberbullying where the personal marker is the pronoun you, the dysphemistic element is the offensive adjective pathetic, and the verb link is the conjugated copula are.

6. Resolution Inferable Cyberbullying Elements

From a detection perspective, discourse dependent instances can be treated as explicit, negation-based, or animal metaphoric[12] constructions once all the cyberbullying elements have been successfully resolved. For this reason, no distinct detection rules need to be designed to target discourse dependent instances. However, we developed resolution algorithms to identify the personal marker, the dysphemistic element, and/or the cyberbullying link so that discourse dependent instances are replaced by complete inferences to which the same detection mechanism can be applied. These algorithms were implemented computationally using the Java programming language (Oracle 2019).

[12] Although there are other stylistic means of cyberbullying manifestation, in the present research we only investigated three forms: explicit, negation-based, and animal metaphors.

Based on the development dataset which is already pre-processed and parsed into dependency sets (using the bidirectional Stanford Dependency Parser, de Marneffe & Manning, 2015), these resolution algorithms target the following: (1) polarity answers, (2) contradictory statements, (3) explicit ellipsis, (4) implicit affirmative answers, and (5) statements that use indefinite pronouns as placeholders for the dysphemistic element. We describe each algorithm and the specific preconditions related to the dependency relations of an instance to undergo the resolution process. We also consider message recency and the number of previous messages because they impact the inference process, and factoring them into the design of the resolution algorithms greatly reduces the number of previous messages that need to be investigated when analysing discourse dependent instances, as well as being more economical from a computational perspective. This is the case because answers and statements that are too far removed from their corresponding questions or answers lose their strength and affect their semantic meaning; in the case of our dataset, this is facilitated by the fact that on Ask.fm they are always presented as pairs of questions and answers (or pairs of statements). Finally, we also factor in the users' screen-names to ensure that the inferences are drawn based on the messages uttered by the relevant user.

The first resolution algorithm applies to polarity answers which can represent all discourse dependent construction types, except personal marker and cyberbullying link inferable types. They target yes/no answers which may or may not contain the personal marker and/or the cyberbullying link in an explicit manner (for example, *yes, you* or *no, you aren't*). Thus, in order for this algorithm to be applied to a given instance, that instance must contain an affirmative polar answer (such as *yes, yup, yep, yeah*) or a negative polar answer (such as *no, nah, nope*), either as part of the root relation, or as an adverbial modifier. Once this pre-condition is satisfied, the previous message is examined to ensure that it explicitly contains a nominal subject relation, that is, to ensure that it contains the cyberbullying elements that need to be inferred. Subsequently, the inferred answer is built following the steps outlined in (11):

(11)
a. Initially, the answer is the same set of dependencies as the question.
b. If the dependency set contains a punctuation relation that has a question mark, that question mark is replaced by a full stop.
c. If the answer has an interrogative structure, then the nominal subject and the copular auxiliary verb relations are inverted, and their indices updated.
d. If the nominal subject is a first person pronoun (*I* or *we*) or a nominal phrase that contains a first person pronoun possessive modifier (*my* or *our*), the nominal subject relation is modified to contain a second person pronoun (*you*) or the possessive modifier relation is modified to contain a second person pronoun possessive modifier (*your*). Conversely, if the nominal subject is a second person pronoun (*you*) or a nominal phrase that contains a second person pronoun possessive modifier (*your*), the nominal subject relation is modified to contain a first person pronoun (*I* or *we*) or the possessive modifier relation is modified to contain a first

person pronoun possessive modifier (*my* or *our*). (Note that for the third person pronouns and proper names no modification is needed at this stage.)

e. In the case of copular constructions with the verb *be*, if the second component of the copula relation is a first person singular conjugated form (*am*, *'m*), during this step, it is replaced by the second person conjugated form (*are*). In the case of transitive constructions, if the second component of the direct object relation is a first person pronoun (me, us), it is replaced by a second person pronoun (*you*). Conversely, if the second component of the direct object relation is a second person pronoun (*you*), it is replaced by a singular first person pronoun (*me*). (Note that in the case of third person pronouns, proper names and other nouns, no changes are required.)

f. Finally, if the polar answer is a negation trigger such as *no*, a negation relation is inserted into the set of dependencies and the indices of the relevant relations are updated; this new relation must negate the root of the inferred answer.

An example of how the inference process occurs for polarity answers is shown in Table 1 where the relevant dependencies are underlined and bold font is applied to the relevant dependency constituent in each given step.

	U1: am I a freak? U2: yes.
a	[[cop(freak-4, am-1), nsubj(freak-4, I-2), det(freak-4, a-3), root(ROOT-0, freak-4), punct(freak-4, ?-5)]]
b	[[cop(freak-4, am-1), nsubj(freak-4, I-2), det(freak-4, a-3), root(ROOT-0, freak-4), punct(freak-4, **.-5**)]]
c	[[nsubj(freak-4, **I-1**), cop(freak-4, **am-2**), det(freak-4, a-3), root(ROOT-0, freak-4), punct(freak-4, .-5)]]
d	[[nsubj(freak-4, **you-1**), cop(freak-4, am-2), det(freak-4, a-3), root(ROOT-0, freak-4), punct(freak-4, .-5)]]
e	[[nsubj(freak-4, you-1), cop(freak-4, **are-2**), det(freak-4, a-3), root(ROOT-0, freak-4), punct(freak-4, .-5)]]
f	N/A

TABLE 1. AN EXAMPLE OF HOW RESOLUTION IS APPLIED TO POLAR ANSWERS.

The second resolution algorithm was designed for contradictory statements and applies to dysphemistic element and cyberbullying link inferable constructions and to dysphemistic element constructions. This algorithm targets statements that use negation or affirmation to contradict the truth of the proposition put forward in the previous message and which contain explicitly the personal marker and/or the cyberbullying link. The statements may also contain the polar *yes* or *no* term, but it is not a required element. Thus, for this algorithm to be applied to

a given instance, the second component of its root relation must be a conjugated copular *be* verb (such as *are, is*) or a conjugated auxiliary verb (such as *does, do*). Once this pre-condition is satisfied, the previous message is also examined to ensure that it explicitly contains a nominal subject relation, that is, to ensure that it contains the cyberbullying elements that must be inferred. Subsequently, the inferred answer is built following the steps outlined in (12):

(12)
a. Initially, the contradictory statement is replaced by the same set of dependencies as the previous message.
b. If the nominal subject is a first person pronoun (*I* or *we*) or a nominal phrase that contains a first person pronoun possessive modifier (*my* or *our*), the nominal subject relation is modified to contain a second person pronoun (*you*) or the possessive modifier relation is modified to contain a second person pronoun possessive modifier (*your*). Conversely, if the nominal subject is a second person pronoun (*you*) or a nominal phrase that contains a second person pronoun possessive modifier (*your*), the nominal subject relation is modified to contain a first person pronoun (*I* or *we*) or the possessive modifier relation is modified to contain a first person pronoun possessive modifier (*my* or *our*). (Note that for the third person pronouns and proper names no modification is needed at this stage.)
c. In the case of copular constructions with the verb *be*, if the second component of the copula relation is a first person singular conjugated form (*am*, *'m*), during this step, it is replaced by the second person conjugated form (*are*). In the case of transitive constructions, if the second component of the direct object relation is a first person pronoun (me, us), it is replaced by a second person pronoun (*you*). Conversely, if the second component of the direct object relation is a second person pronoun (*you*), it is replaced by a singular first person pronoun (*me*). (Note that in the case of third person pronouns, proper names and other nouns, no changes are required.)
d. Finally, if the contradictory statement contains a negation relation, then a negation relation is inserted into the set of dependencies to negate the root of the inferred statement[13]; otherwise, the negation relation that is contained in the initial set of relations is removed. The indices of the relevant relations are also updated.

An example of how these steps are implemented for contradictory statements is shown in Table 2 where the relevant dependencies are underlined and bold font applied to the relevant dependency constituent in each given step.

[13] Note that in this step changing the form of a conjugated verb from singular to plural or vice versa is not of concern, since, during the detection process, the relevant terms are subjected to a lemmatisation process anyway.

	U1: everyone deserves a mom. U2: you don't.
a	[[nsubj(deserves-2, everyone-1), root(ROOT-0, deserves-2), det(mum-4, a-3), dobj(deserves-2, mum-4)]]
b	[[nsubj(deserves-2, **you-1**), root(ROOT-0, deserves-2), det(mum-4, a-3), dobj(deserves-2, mum-4)]]
c	N/A
d	[[nsubj(deserves-4, you-1), aux(deserves-4, do-2), neg(deserves-4, n't-3), root(ROOT-0, deserves-4), det(mom-6, a-5), dobj(deserves-4, mom-6)]]

TABLE 2. AN EXAMPLE OF HOW RESOLUTION APPLIES TO CONTRADICTORY STATEMENTS.

The third resolution algorithm applies to elliptical statements which are representative of personal marker and cyberbullying link inferable constructions. These statements are those utterances that are preceded by the at least two messages, the first one containing the elliptical punctuation marks (…), the first and last messages having to be uttered by the same user (having the same user screen name). In addition, in the last message, the second component of the root dependency must be a noun or an adjective. Once these pre-conditions are satisfied, the resolution steps outlined in (13) can be implemented:

(13)
a. Initially, the inferred statement is the first statement of the three of utterances.
b. Then the elliptical punctuation marks are deleted.
c. Next, the last statement is appended.
d. Finally, the dependency parser is applied to the inferred statement.

An example of how these steps can be implemented is shown in Table 3.

	U1: you are a … U2: what? U1: dog.
a	you are a
b	you are a
c	you are a dog
d	[[nsubj(dog-4, you-1), cop(dog-4, are-2), det(dog-4, a-3), root(ROOT-0, dog-4)]]

TABLE 3. AN EXAMPLE OF HOW RESOLUTION IS APPLIED TO ELLIPSIS.

The fourth resolution algorithm targets implicit affirmative answers that do not contain explicitly the *yes* answer, but a dysphemistic phrase to which no negation relation is applied, and, because of this, it applies only to personal marker and cyberbullying link inferable constructions. Moreover, based on the present dataset, there are several pre-conditions that need to be satisfied before this algorithm is applied: first, the previous message must be a

yes/no question that has an underlying transitive structure (thus its corresponding set of dependencies must contain the direct object relation) and, secondly, the answer must explicitly contain an explicit dysphemistic adjective or noun. The inferred answer is then built based on the steps described in (14):

(14)
a. The second component of the direct object relation is first extracted from the first statement.
b. Then the inference is built using a copular structure in which the result of the previous step is linked to the answer; in other words, the result of previous step becomes the nominal subject of the sentence and the answer becomes the complement of the copula *be*.
c. Finally, the dependency parser is applied to the resulting inference.

An example of how these steps are applied to implicit affirmative answers is shown in Table 4:

	U1: saw Ella? U2: ugly!
a	Ella
b	Ella is ugly.
c	[[nsubj(ugly-3, Ella-1), cop(ugly-3, is-2), root(ROOT-0, ugly-3)]]

TABLE 4. AN EXAMPLE OF HOW RESOLUTION RULES IS APPLIED TO IMPLICIT AFFIRMATIVE ANSWERS.

Finally, we designed a resolution algorithm to resolve statements that use indefinite pronouns as placeholders for the dysphemistic element, and in order to do so, again, only the previous message needs to be considered. This algorithm applies to dysphemistic element constructions only, since the only inferable element is the dysphemistic element. However, several pre-conditions must be satisfied: first, the previous message must have a copular structure with the verb *be* or a transitive structure (intransitive constructions are excluded because in that case the dysphemistic element is the same linguistic item as the intransitive verb, thus, constituting, in fact, discourse independent instances) and, secondly, the second message must contain the indefinite pronoun. Subsequently, the resolution steps outlined in (15) can be applied:

(15)
a. First, if the previous statement has a copular structure, the second component of the root relation is extracted from its dependency set; if it has a transitive structure, the second component of the direct object relation is extracted from its dependency set.
b. Then the indefinite pronoun is replaced with the result of step a.
c. Finally, the dependency parser is applied to the resulting inference.

An example of how these steps are implemented is shown in Table 5.

	U1: my favourite animals are monkeys U2: you look like one
a	Monkeys
b	You look like monkeys.
c	[[nsubj(look-2, you-1), root(ROOT-0, look-2), prep(look-2, like-3), pobj(like-3, monkeys-4)]]

TABLE 5. AN EXAMPLE OF HOW RESOLUTION IS APPLIED TO IMPLICIT AFFIRMATIVE ANSWERS.

7. Conclusion

The approach we propose in this paper represents an effective way of reducing the number of false negatives (missed cyberbullying instances) that a system fails to detect due to the fact that such instances lack the key cyberbullying elements. Although the number of instances that draw from prior messages is relatively small in the present dataset, we successfully formalise a framework that describes how information found in prior discourse (antecedent messages or posts) can be used to identify the missing cyberbullying elements so that discourse dependent online instances can be subjected to the same detection process that applies to instances that contain explicitly all these elements (Power et al 2018).

First, we introduce the necessary and sufficient elements that our definition of public textual cyberbullying posits (Power et al 2017; 2018) and how they can be inferred from prior discourse using Prince's paradigm of information (1981), that is, how the missing cyberbullying elements can be viewed as discourse-old information.

We then investigate examples of discourse dependent instances that were present in our dataset and, subsequently, we identify and characterise four types of cyberbullying constructions that characterise discourse dependent instances: 1) fully inferable constructions – where all three cyberbullying elements, the personal marker, the dysphemistic element, and the link between them, are missing and they can be inferred from previous messages, (2) personal marker and cyberbullying link inferable constructions – where the dysphemistic element is explicitly present, but the personal marker and the link must be inferred from previous discourse, (3) dysphemistic element and cyberbullying link inferable constructions – where the personal marker is explicitly present, but the dysphemistic element and the cyberbullying link are entities inferable from previous discourse, and (4) dysphemistic element inferable constructions – where the personal marker and the link are explicitly present, but the dysphemistic element must be inferred from prior discourse.

Finally, we describe resolution algorithms designed to resolve the missing cyberbullying elements by building inferences that represent complete discourse independent instances to which the same detection mechanism described in previous work (Power et al 2018) can be applied. The resolution algorithms developed presently target the following types of instances: (1) polarity answers, (2) contradictory statements, (3) explicit ellipsis, (4) implicit affirmative answers, and (5) statements that use indefinite pronouns as placeholders for the dysphemistic

element. Additionally, to increase computational efficiency, the resolution algorithms apply only to instances that satisfy certain pre-conditions which we identified in the underlying grammatical structures.

References

Agne, R.R., and Tracy, K. 2009. "Conversation, Dialogue, and Discourse". In *21st Century Communication: A Reference Handbook,* edited by W.F. Eadie, 177 – 185. Sage Publications Inc.

Al-garadi, M.A., Varathan, K.D. and Ravana S.D. 2016. "Cybercrime Detection in Online Communications: The Experimental Case of Cyberbullying Detection in the Twitter Network." *Computers in Human Behaviour,* 63: 433 – 443.

Allan, K. and Burridge, K. 2006. *Forbidden Words: Taboo and Censoring of Language.* Cambridge: Cambridge University Press.

Boyd, D. 2007. "Why Youth (Heart) Social Network Sites: The Role of Networked Publics in Teenage Social Life." In *MacArthur Foundation Series on Digital Learning, Youth, Identity, and Digital Media,* edited by David Buckingham, 1 – 26. Cambridge, MA: MIT Press.

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., and Vakali, A. 2017. "Mean Birds: Detecting Aggression and Bullying on Twitter." Cornell University Library: Last modified March 5th. https://arxiv.org/abs/1702.06877.

Chen, Y., Zhou, Y., Zhu, S. and Xu, H. 2012. "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety." Paper presented at the ASE/IEEE International Conference on Social Computing, 71 - 80. Washington, DC, September 3-5.

Dadvar, M., Trieschnigg, D., R. Ordelman, R., and de Jong, F. 2013. "Improving cyberbullying detection with user context." Paper presented at the 35th European conference on Advances in Information Retrieval, 693 – 696. Moscow, March 24-27.

de Marneffe, M.C., and Manning, C.D. 2008a. "The Stanford typed dependencies representation." Paper presented at the COLING 2008 Workshop on Cross-framework and Cross-domain Parser Evaluation. Manchester, UK August 23 - 23.

de Marneffe, M.C., and Manning, C. 2008b. "Stanford typed dependencies manual." Last modified March 5th. http://nlp.stanford.edu/software/stanford-dependencies.shtml.

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., and Picard, R. 2012. "Common sense reasoning for detection, prevention, and mitigation of cyberbullying." *ACM Transactions on Interactive Intelligent Systems*, 2: 18:1-18:30. doi: 10.1145/2362394.2362400.

Dooley, J.J., Pyzalski, J., and Cross, D. 2009. "Cyberbullying versus face-to-face bullying – A theoretical and conceptual review." *Journal of Psychology*, 217: 182–188. doi: 10.1027/0044-3409.217.4.182.

Goncalves, M. 2011. "Text Classification". In *Modern Information Retrieval, the concepts and technology behind search*, edited by Ricardo Baeza-Yates and Berthier Ribeiro-Neto, 281 – 336. Pearson Education Limited.

Grigg, D.W. 2010. "Cyber-Aggression: Definition and Concept of Cyberbullying." *Australian Journal of Guidance and Counselling*, 12: 143–156.

Hinduja, S., and Patchin, J.W. 2009. *Bullying beyond the schoolyard: preventing and responding to cyber-bullying*. Thousand Oaks, CA: Corwin Press.

Hosseinmardi, H., Han, R., Lv, Q., Mishra, S., and Ghasemianlangroodi, A. 2014a. "Towards Understanding Cyberbullying Behavior in a Semi-Anonymous Social Network." Paper presented at the International Conference on Advances in Social Networks Analysis and Mining. Beijing, August 17-20.

Hosseinmardi, H., Rafiq, R. I., Li, S., Yang, Z., Han, R., Lv, Q., and Mishra, S. 2014b. "A Comparison of Common Users across Instagram and Ask.fm to Better Understand Cyberbullying." Paper presented at the 7th International Conference on Social Computing and Networking. Sydney, December 3-5.

Huang, Q., Singh, V.K., and Atrey, P.K. 2014. "Cyber Bullying Detection using Social and Textual Analysis." Paper presented at the 3rd International Workshop on Socially-Aware Multimedia, 3 – 6. Orlando, Florida, November 7.

InternetSlang 2019. "Internet Slang – Internet Dictionary." Last modified March 5th. http://www.internetslang.com/.

Kavanagh, P. 2014. "Investigation of Cyberbullying Language & Methods." MSc diss., ITB, Ireland.

Kontostathis, A., Reynolds, K., Garron, A. and Edwards, L. 2013. Detecting Cyberbullying: Query Terms and Techniques. Paper presented at the 5th Annual ACM Web Science Conference. Paris, May 2-4.

Krifka, M. 2001. "For a structured meaning account of questions and answers (revised version)". In *Audiatur Vox Sapientia. A Festschrift for Arnim von Stechow*, edited by C. Fery& W. Sternefeld 287-319. Berlin: Akademie Verlag (= studiagrammatica 52).

Krifka, M. 2011. "Questions". In *Semantics. An international handbook of Natural Language Meaning*, edited by K. von Heusinger, C. Maienborn & P. Portner ,Vol. 2, 1742-1785. Berlin: Mouton de Gruyter.

Langos, C. 2012. "Cyberbullying: The Challenge to Define." *Cyberpsychology, Behavior, and Social Networks*, 15(6): 285-289. doi: 10.1089/cyber.2011.0588.

Livingstone, S.,Haddon, L., Görzig, A., and Ólafsson, K. 2011. "EU Kids Online: final report 2011." Last modified March 5th. http://eprints.lse.ac.uk/45490/.

Livingstone, S., Mascheroni, G., Ólafsson, K., and Haddon, L. with the networks of EU Kids Online and Net Children Go Mobile 2014. "Children's online risks and opportunities: Comparative findings from EU Kids Online and Net Children Go Mobile". Last modified March 5th. http://eprints.lse.ac.uk/60513/.

Nahar, V., Li, X. and Pang, C. 2013. "An Effective Approach for Cyberbullying Detection." *Communications in Information Science and Management Engineering*, 3:238 – 247.

Nandhini, B.S., and Sheeba, J.I. 2015. "Online Social Network Bullying Detection Using Intelligence Techniques." *Procedia Computer Science*, 45: 485 – 492.

Navarro, G. and Ziviani, N. 2011. "Documents: Languages & Properties". In *Modern Information Retrieval, the concepts and technology behind search*, edited by Ricardo Baeza-Yates and Berthier Ribeiro-Neto, 203 – 254. Pearson Education Limited.

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." Paper presented at the 6th International Joint Conference on Natural Language Processing. Nagoya, October 14-19.

Norvig, P. 2007. "How to Write a Spelling Corrector." Last modified March 5th. http://norvig.com/spell-correct.html.

Oracle 2019. Java™ Platform, Standard Edition 9 API Specification. Last modified March 5th. Available online at https://docs.oracle.com/javase/9/docs/api/overview-summary.html.

Power, A., Keane, A., Nolan, B., and O'Neill, B. 2017. "A Lexical Database for Public Textual Cyberbullying Detection". Special issue of *Revista de lenguas para fines específicos*, entitled *New Insights into Meaning Construction and Knowledge Representation.*

Power, A., Keane, A., Nolan, B., and O'Neill, B. 2018. "Detecting Discourse-Independent Negated Forms of Cyberbullying". *Journal of Computer-Assisted Linguistic Research*, 2: 1 – 20.

Prince, E.F. 1981. "Toward a taxonomy of given/new information". In *Radical Pragmatics*, edited by P. Cole, 223 - 254. Academic Press.

Ptaszynski, M., Dybala, P., Matsuba, T., Rzepka, R. and Araki, K. 2010. "Machine Learning and Affect Analysis Against Cyber-Bullying." Paper presented at the 36th AISB Annual Convention. March 29- April 1.

Ptaszynski, M., Masui, F., Nitta, T., Hatekeyama, S., Kimura, Y., Rzepka, R., and Araki, K. 2016. "Sustainable Cyberbullying Detection with Category-Maximised Relevance of Harmful Phrases

and Double-Filtered Automatic Optimisation." *International Journal of Child-Computer Interaction*, 8: 15 – 30.

Reynolds, K., Kontostathis, A. and Edwards, L. 2011. "Using Machine Learning to Detect Cyberbullying." Paper presented at the 10th International Conference on Machine Learning and Applications Workshops. Hawaii, December 18-21.

Sourander, A., Brunstein-Klomek, A., Ikonen, M., Lindroos, J., Luntamo, T., Koskelainen, M., Ristkari, T., Hans Helenius, H. 2010. "Psychosocial risk factors associated with cyberbullying among adolescents: A population-based study." *Arch Gen Psychiatry*, 67: 720-728.

Unicode 2019. "Emoticons." Last modified March 5th. http://www.unicode.org/.

Van Hee, C., Lefever, E.,Verhoeven, B.,Mennes, J.,Desmet, B., DePauw, G., Daelemans, W., and Hoste, V. 2015. "Detection and Fine-GrainedClassificationofCyberbullyingEvents." Paper presented at the annual conference on RANLP. Hissar, September 5-11.

Yin, D., Xue, Z., Hong, L., Davison, B.D., Kontostathis, A., and Edwards, L. 2009. "Detection of harassment on web 2.0." Paper presented at the 1st conference on CAW. Madrid, April 20-24.