Power, Aurelia, Keane, Anthony, Nolan, Brian, and O'Neill, Brian (2018) Detecting Discourse-Independent Negated Forms of Public Textual Cyberbullying. Journal of Computer-Assisted Linguistic Research 2: 1-20.

Detecting Discourse-Independent Negated Forms of Public Textual Cyberbullying

Aurelia Power[1*], Anthony Keane[1], Brian Nolan[1], Brian O'Neill[2]
[1]Institute of Technology Blanchardstown, Dublin
[2]Dublin Institute of Technology
*Corresponding author: aurelia.power@itb.ie

Received: 19 December 2017 / Accepted: 21 February 2018 / Published: 12 July 2018

Abstract

Cyberbullying is a risk associated with the online safety of young people and, in this paper, we address one of its most common implicit forms – negation-based forms. We first describe the role of negation in public textual cyberbullying interaction and identify the cyberbullying constructions that characterise these forms. We then formulate the overall detection mechanism which captures the three necessary and sufficient elements of public textual cyberbullying – the personal marker, the dysphemistic element, and the link between them. Finally, we design rules to detect both overt and covert negation-based forms, and measure their effectiveness using a development dataset, as well as a novel test dataset, across several metrics: accuracy, precision, recall, and the F1-measure. The results indicate that the rules we designed closely resemble the performance of human annotators across all measures.

Keywords: cyberbullying detection, dependency parsing, negation, natural language processing.

1. Introduction

Cyberbullying has become increasingly more prevalent among the young people using the Internet (Livingstone et al. 2011; Livingstone et al. 2014), affecting the emotional and psychological wellbeing of the victim(s), which, in severe cases can lead to serious pathological issues, such as depression, self-harm, suicide ideation, and suicide attempt (Sourander et al. 2010). Cyberbullying, like face-to-face bullying, is typically defined in terms of three fundamental criteria: intention of harm, repetition, and power imbalance between the victim and the bully (Hinduja and Patchin 2009). However, while face-to-face bullying is restricted by temporal and geographical constraints, the specific environment in which cyberbullying occurs – the cyberspace – allows the act of cyberbullying to transcend such constraints. From this perspective, cyberspace is characterised by four key aspects: (1) the ability to persist over time,

DOI: https://doi.org/10.4995/jclr.2018.8917
PUBLISHER: UNIVERSITAT POLITÈCNICA DE VALÈNCIA.

(2) the ability to be searched for repeatedly, (3) the ability to be replicated numerous times, and (4) the ability to multicast to potentially large invisible audiences (Boyd 2007).

As a consequence, some authors have questioned whether repetition is a necessary element of cyberbullying (Dooley et al. 2009; Grigg 2010), while others have adopted a relaxed view of this criterion (Langos 2012; Power et al. 2017), that is, repetition can be achieved by an instance by simply being viewed multiple times by the victim, particularly in the case of cyberbullying that occurs in the public domain. For example, a single hurtful textual instance on social media can remain in cyberspace for an indefinite period of time, and, it can be viewed repeatedly not only by the victim, but also by a potentially large number of bystanders; in addition, the post may be re-posted or re-tweeted which can also lead to potential multiple-viewings by the victim; this, in turn, can lead to an intensified sense of powerlessness, because of the larger and more persistent audience.

Given the pervasive nature and long-lasting negative effects of public textual cyberbullying, it is essential that cyberbullying is addressed in all its forms, and, in the present paper, we focus on one common type of cyberbullying – discourse-independent public textual cyberbullying that uses negation. In this respect, we are the first to describe the role of negation in public textual cyberbullying interaction and design accordingly a set of rules to detect these forms. We measure the effectiveness of such rules on two datasets: a development dataset and a novel test set, across several metrics: accuracy, precision, recall, and the F1-measure. Our results indicate that the rules designed here closely resemble the human performance across all measures.

2. RELATED WORK

Research in the field of cyberbullying detection has recently gained momentum, with many approaches taking advantage of the advances in the fields of text analytics and Natural Language Processing (NLP). From this perspective, the task of cyberbullying detection was previously approached as a classification task (Yin et al. 2009) that involves data acquisition and pre-processing, feature extraction, and classification. These techniques were used mostly in targeting explicit textual cyberbullying language and rely on detecting features such 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).

Several other studies have also targeted the detection of implicit forms of cyberbullying (Chen et al. 2012; Dinakar et al. 2012; Nitta et al. 2013; Ptaszynski et al. 2010; Ptaszynski et al. 2016). However, they do not clearly define the forms that represent implicit cyberbullying. In fact, the majority of previous approaches do not provide clear boundaries of what constitutes cyberbullying in general and they target the detection of rude and violent language rather than cyberbullying instances. On the other hand, the view that we take here, as well as in previous work (Power et al. 2017), is 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. We repeat here the definition we advanced previously as follows:

> (1) **Definition:** A given public textual instance (whether expressed as a message, a post or a sentence) can be classified as cyberbullying if it contains the following three elements: (1) the personal marker/pointer, (2) the explicit or implicit dysphemistic element, and (3) the link between the personal marker/pointer and the dysphemistic element.

The personal marker/pointer refers to that element that is used to identify or point to the victim(s), while the dysphemistic element refers to what has been 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"; the link between the previous two elements capture the final element of our definition and identifies the means by which the dysphemistic element targets the victim(s) identified or pointed to by the personal marker/pointer. The characteristics of explicitness and implicitness are inherent characteristics that apply to the dysphemistic element only, and refer to whether cyberbullying instances contain explicit (profane offensive, or violent) terms or not. However, the three elements need not be explicitly present in a given instance, as long as they can be inferred from other contextual elements, such as the sentential structure or previous discourse. For example, in the sentence *You are a cunt*, the dysphemistic element is both explicitly present and realised by means of explicit profane language, but, in the sentence *You are not smart*, although the dysphemistic element is explicitly present, it is not realised by means of explicit language, but by means of negation. On the other hand, the instance *You clearly are*, although it contains no dysphemistic language, was labelled as public textual cyberbullying in our dataset, and it was only when we considered the previous post/message uttered by a different user - *I am not pathetic* - that we were able to identify the dysphemistic element in the form of the offensive adjective *pathetic* and we were able to resolve the sentence to its full form: *You clearly are pathetic*.

Following from our definition, explicit terms are not the only means by which public textual cyberbullying can be realised. For instance, the messages/posts *You don't deserve a mum* or *You are not pretty* do not contain any explicit profane or offensive, or violent terms; they are instances of public textual cyberbullying that use negation to hurt or offend the victim(s). Such instances, however, have never being considered by the previous research, and, to address this limitation, we focus in this paper on those instances of public textual cyberbullying that contain the negated dysphemistic element in an explicit manner.

3. Negated forms of public textual cyberbullying

Instances of public textual cyberbullying that use negation constitute the second most common type of public textual cyberbullying in our dataset and are characterised by the fact that the dysphemistic element does not occur as explicit terms/expressions, instead it is realised by means of positive or neutral connotation terms/expressions used in conjunction with negation operators and triggers. Furthermore, negation-based instances of cyberbullying that we identified in our dataset can be overt, using explicit negation triggers, or covert, using negative connotation expressions.

Based on the dataset we have used, overt forms of negation-based public textual cyberbullying are more likely to occur than covert forms, and the grammatical structures representative of overt negation are typically realised by using the operators *not* and *no,* and the contracted forms *n't* added to the main verb, as well as by using indefinite pronouns such as *nobody* or *nothing* (Horn 1989; Lawler 2005). Some examples of overt negation-based instances found in the present development dataset are shown in (2).

(2)
a. Negating verbs or verb phrases is the most common; examples include *You don't deserve a mum* or *I don't like your face,* or *You can't spell.*
b. Negating the adjective or adjectival phrases, such as *You are not pretty,* or *You are not very smart.*
c. Using indefinite pronouns, such as *Nobody likes you* or *You deserve nothing and nobody.*

However, there are other common grammatical structures representative of negation (Horn 1989; Lawler 2005) that we did not encounter in our dataset, but for which anecdotal evidence can be found in everyday interaction; these structures include the following:

(3)
a. Negating the noun or noun phrases, such as *You have no taste in clothing* or *You are no beauty.*
b. Using negation prepositional phrases, such as *You are all nice, except Lena,* or *Students are all very clever, with your exception.*
c. Using negative frequency adverbs, such as *You were never smart.*
d. Using negative manner adverbs, such as *You are hardly worth talking to,* or *You can barely spell.*
e. Using negative probability adverbs, such as *It's unlikely that you will ever do well.*
f. Using negative verbs, such as *I doubt you are smart.*

In addition, we found in our dataset several instances that do not invoke any of the overt negation triggers discussed. These instances rely on the semantic content of verbs that express personal believes and opinions, such as *think,* or verbs that express attempting or making an effort, such as *try.* For instance, the sentence *You think you are pretty* implies that, in fact, the victim is not pretty, and it is only the victim's belief about oneself that she/he is pretty, while the sentence *You are trying to sound smart* implies that the victim is not smart, despite the obvious efforts.

4. EXPERIMENTAL SET-UP

4.1. Datasets and Data Labelling

We used two sources to acquire data: a dataset that Kavanagh (2014) used in her research and a dataset that Hosseinmardi et al. (2014a) and Hosseinmardi et al. (2014b) used to develop and test their framework from which we randomly selected a continuous portion. We then merged the two datasets, the entire dataset consisting of a total of 2038 instances; this larger dataset was subsequently divided into a development dataset of 1504 instances, and a test dataset of 534 instances. Both datasets originate from ASK.fm and they contain conversations corresponding to 16 users which are organised as pairs of questions and answers. These conversations are typically characterised by flaming or online fight, where insults and hurtful messages are exchanged.

To label the data, we used a rigorous process to minimise bias associated with data labelling. The datasets were presented to annotators in natural conversational order, to allow them to take advantage of the contextual information provided by the discourse. Two individuals were asked to label the instances in the development and test datasets using the labels of cyberbullying (CB), or 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 in the development dataset, 21.87% were labelled as cyberbullying, while in the test dataset, approximately 26.20% were labelled as cyberbullying. From those instances labelled as cyberbullying in the development dataset, approximately 10.06% were negation-based instances of public textual cyberbullying, while in the test dataset, approximately 8.21% represented instances of negation-based public textual cyberbullying.

In addition, we computed the annotator/inter-observer agreement score for 19.33% of all instances to be used as an upper bound against which accuracy, precision, recall, and F1-meassure scores can be compared; the intuition behind this is that a system's accuracy, recall, and precision are not expected to be higher than this score which represents the human performance. The percentage of annotator/inter-observer agreement score (IOA) was computed according to the following formula:

$$(4) \qquad IOA = \frac{\text{the number of instances for which both annotators provided the same label}}{\text{the total number of instances}} X\,100;$$

The results show that the two individuals agreed on 390 out of the 394 total messages/posts, yielding a score of 98.98% inter-observer agreement.

4.2. Pre-Processing[1]

To efficiently process the text files, we used an automatic procedure to remove all xml and html tags, and we retained only the usernames. In addition, we manually[2] removed those posts that were in languages other than English, and from the remaining posts, we have automatically removed all hyperlinks. In addition, we inversed the order of the question-answer pairs automatically, from the most recent pair to the least recent pair to reflect the conversational order. We then applied several techniques[3] that are typically used in text analytics and NLP, namely tokenization, case transformation, and lemmatisation (Navarro and Ziviani 2011).

We also applied pre-processing techniques which served to not only increase the level of accuracy of the dependency parsing, but also to provide additional information that aids the detection process. For example, online textual instances make frequent use of abbreviations and acronyms,[4] such as *bj* which is an acronym for *blow job*, or *dc* which stands for *don't (doesn't) care.* In addition, public instances of online communication are often characterised by many deliberate spelling errors such as omitting certain characters, for instance, *fck (fuck)* or *dont (don't),* and substituting a letter or group of letters with digits, for example, *id1ot (idiot)* or *h8 (hate)*. These phenomena may be explained by the fact that the user tries to resemble the speed of face-to-face interaction or, in the case of cyberbullying, by the fact that the bully tries to avoid detection. Thus, we argue that acronyms, abbreviations, and deliberate errors may contain vital information and, to account for them, we used a dictionary of acronyms and abbreviations (Internet Slang 2017) to replace them with their corresponding full forms.[5] We also replaced all informal variations of the personal pronouns[6] commonly found in online interaction with their corresponding formal forms; for instance, *u, ya* were replaced by *you,* while informal reflexive forms such as *yerself* or *meself* were replaced by *yourself* and *myself*, respectively. Furthermore, to compensate for the lack of kinetics and proxemics found in face-to-face communication, textual forms of online interaction often contain icons intended to represent facial expressions, gestures, emotions, etc. However, in our development dataset we noted very few instances of such icons. In addition, their use were not directly related to cyberbullying. For these reasons, we removed

[1] All pre-processing techniques (apart from the removal of posts in other languages other than English) were carried out automatically and we applied them programmatically using the Java programming language (Oracle 2017).

[2] We removed these posts manually because we did not implement any software to identify other languages, given its complexity that goes beyond the scope of the present research.

[3] To implement these techniques, we took advantage of the rich collection of algorithms offered by the WEKA machine learning toolkit (Witten et al 2011), algorithms that are also implemented using the Java programming language.

[4] We also include here initialisms.

[5] The acronym/abbreviation replacement procedure was carried out automatically and we implemented such functionality in Java by using its Map datastructure to store the dictionary obtained from Iternetslang.com (2017) in the format of key-value pairs, where the acronyms/abbreviations constitute the keys and the respective full forms constitute the values; we then defined a search functionality and a replace functionality to search the datasets and everytime a key defined in the dictionary was found, it was replaced with the corresponding value.

[6] A similar replacement procedure to that used for acronyms/abbreviations was used in the case of informal pronouns (see footnote 4).

any icons, smileys, and emoticons. Other errors that we encountered in public textual cyberbullying, not necessarily deliberate, were repeated letters, meaningless symbols, transposition, missing, and wrong characters, and we addressed them by employing Norvig's spelling algorithm (2007). For example, words such as *killl* and *liek* were corrected to *kill* and *like*, respectively.

4.3. Discourse-Independent Negation-Based Cyberbullying Constructions

To capture cyberbullying instances, we first identify whether the three cyberbullying elements – the personal marker/pointer, the dysphemistic element, and the link between them – are explicitly found in an instance. Thus, negation-based forms of public textual cyberbullying that do not depend on previous messages or posts to infer its cyberbullying elements can be further divided into several types. However, given that all discourse-independent forms must contain the dysphemistic element in an explicit manner (in the present case either as overt negation, or covert negation), as well as the fact that negation requires the explicit presence of the personal marker/pointer, discourse-independent forms of negated public textual cyberbullying can only be characterised by four types of cyberbullying constructions: (1) full overt negation constructions where the personal marker, the overt negated dysphemistic element, and the verb link between them are all present in an explicit manner, (2) cyberbullying link-inferable overt negation constructions in which the personal marker and the overt negation dysphemistic element are explicitly present, but the link verb is inferable from the sentential structure of the instance, (3) full covert negation constructions in which the personal marker, the covert negated dysphemistic element, and the verb link between them are all present in an explicit manner, and (4) cyberbullying link-inferable covert negation constructions in which the personal marker and the covert negation dysphemistic element are explicitly present, but the link verb is inferable from the sentential structure of the instance.[7]

Examples corresponding to each of these types are shown in (5):

(5) a. *You don't deserve a mum.* (full overt negation construction)
b. *Not your brightest idea!* (cyberbullying link-inferable overt negation construction)
c. *You are trying to sound smart.* (full covert negation construction)
d. *You think you pretty!* [8](cyberbullying link-inferable covert negation construction)

4.4. The Detection Mechanism

The detection mechanism that we propose is comprised of detection rules[9] based on our definition of public textual cyberbullying, ensuring that, for any given instance, all three elements – the personal marker/pointer, the dysphemistic element, and the link between them – are captured. To identify each of these, we use several features of the lexical entry that are

[7] We found no evidence for this type of constructions in the datasets we have used.

[8] This example was modified here for demonstration purposes.

[9] The rules are also implemented using the Java programming language (Oracle 2017).

defined in the cyberbullying lexical database proposed by Power et al. (2017), both semantic features, such as the cyberbullying function and the cyberbullying referential domain, and grammatical, such as the syntactic category. In addition, we employ grammatical dependencies in the process of detection because they provide an overview of how each cyberbullying element is grammatically related to one another; for this purpose we use the bidirectional Stanford Dependency Parser (de Marneffe and Manning 2008a; 2008b) which represents each grammatical dependency as a binary relation between a governor (or a regent) and a dependent.

We implemented all detection rules using the following general format: given a set of dependencies, (1) check that the relevant dependency relations are present, (2) check whether overt negation is applicable, (3) check to see whether the relevant dependencies are related, (4) apply lemmatisation to the extracted dependency constituents, and, finally, (5) check to see whether these dependency constituents represent the necessary cyberbullying elements, that is, to see whether the dependencies constituents satisfy the conditions imposed by the methods[10] designed to identify various properties of the lexical entry stored in the database.

4.4.1. Overt Negation Rules

Overt negation cyberbullying instances contain the negation dysphemistic element in an explicit manner and to capture such instances we have developed rules that do not rely on the presence of explicit profane/obscene, offensive/insulting, or violent terms. Instead, they check for the presence of overt negation triggers that can be captured using the following grammatical relations: (1) the nominal subject which can be used to account for the presence of indefinite pronouns, such as *nobody,* or for the presence of negative verbs, such as *doubt*, (2) the negation modifier relation which can be used to check for the presence of overt operators such as *no, not* and *n't*, (3) the prepositional modifier which can be used to identify the presence of certain prepositions, such as *except*, and (4) the adverb modifier relation which can be used to account for the presence of adverbs such as *hardly* or *barely*. These rules are described as follows.

The nominal subject relation and the direct object relation together capture cyberbullying instances when the sentence contains a negative indefinite pronoun such as *nobody* or *nothing*. In addition, the second component of the direct object relation must be a personal marker/pointer, such as a personal pronoun, a proper name or an indefinite pronoun (except first person pronouns). Also, the main verb must be a positive or neutral term, such as *want, like,* or *deserve.* For example, the instances *nobody likes you* and *you deserve nothing* are labelled as cyberbullying based on the *nsubj*(likes-2,[11] nobody-1) and *dobj*(likes-2, you-3) relations, and the *nsubj*(deserve-2, you-1) and *dobj*(deserve-2, nothing-3) relations, respectively. This rule applies exclusively to full

[10] These methods typically return a boolean value. For example, the method *isPersonalMarker* takes as argument a String represented by a lemma; this lemma is first checked to see whether it exists in the database and, if it does, it returns true if its cyberbullying function is that of personal marker (applicable to personal pronouns, proper names, and person-refering nouns), otherwise, false.

[11] The Bidirectional Stanford Dependency Parser (de Marneffe and Manning 2008a) represents each relation using the positions of the words in a given sentence; for instance, in the relation *nsubj*(likes-2, nobody-1), the numbers *2* and *1* indicate that *likes* is the second word in the sentence, while *nobody* is the first word in the sentence. Note that each sentence contains a *root* relation that has a fake node *ROOT* which is always represented at position 0.

overt negation constructions, since it specifies overt negation triggers as part of the dysphemistic element, as well as the explicit presence of the personal marker and the link verb.

The negation modifier and the nominal subject relations together capture cyberbullying instances when the first constituents of both relations are represented by the same positive or neutral noun, adjective, or a verb. Moreover, the second constituent of the nominal subject relation must be a personal marker/pointer (except first person pronoun). For example, the instances *you are not pretty* and *you were never smart* are labelled as cyberbullying based on the *nsubj*(pretty-4, you-1) and *neg*(pretty-4, not-3) relations, and the *nsubj*(smart-4, you-1) and *neg*(smart-4, never-3), respectively. This rule applies to both types of overt negation constructions.

The negation modifier and the nominal passive subject relations together capture cyberbullying instances in a similar manner to the previous rule, with the exception that the first constituents of both relations must be a participial form of a positive or neutral verb. For instance, the sentence *you won't be missed* is labelled as cyberbullying based on the *nsubjpass*(missed-5, you-1) and *neg*(missed-5, n't-3) relations. Likewise, this rule applies to both types of overt negation constructions.

The nominal subject, the negation modifier, and the clausal complement relations can also capture cyberbullying instances. In this case, the main verb must be a positive or neutral verb which appears as the first component in all three relations. In addition, the second component of the nominal subject must be a personal marker (except first person pronouns), while the second component of the complement relation must be a positive or neutral noun. For instance, *You don't know how to spell* can be labelled as cyberbullying based on the *nsubj*(know-4, You-1), *neg*(know-4, n't-3), and *ccomp*(know-4, spell-7) relations. This rule applies only to full overt negation constructions, since all cyberbullying elements must be explicitly present.

There are four rules in which the nominal subject, the negation modifier, and the direct object relations together capture cyberbullying instances. In the case of the first three rules, the first constituents of all three relations must be the same transitive positive or neutral verb. However, for the first rule, the second constituent of the nominal subject relation must be a personal marker/pointer (except first person pronoun), while the second constituent of the direct object relation must be a positive noun or an indefinite pronoun such as *anything* or *anyone.* For example, the sentence *you don't deserve a mom* is labelled as cyberbullying based on the *nsubj*(deserve-4, you-1), *neg*(deserve-4, n't-3) and *dobj*(deserve-4, mom-6) relations. In the case of the second rule, the second constituent of the nominal subject relation must be a personal marker/pointer (except second person pronoun), while the second constituent of the direct object relation must be also a personal marker/pointer (except first person pronoun). For instance, *we do not want you here* is labelled as cyberbullying based on the *nsubj*(want-4, we-1), *neg*(want-4, not-3) and *dobj*(want-4, you-5) relations. For the third rule, the second constituent of the nominal subject relation must be the same as the first constituent of the negation modifier relation in the form of the indefinite pronoun *one,* while the second constituent of the direct object relation must be also a personal marker/pointer (except first person pronoun). For example, the instance *no one wants you* is labelled as cyberbullying based on the *neg(*one-2, no-1), *nsubj*(wants-3, one-2) and *dobj*(wants-3, you-4) relations. Finally, the last rule applies to those instances that use the negation trigger *no* with the direct object of the verb. In such cases, the first constituents of the nominal subject and direct object relations must be the same transitive positive or neutral verb. In addition, the first constituent of the negation modifier relation and second component of the direct object must be a positive noun phrase. The personal marker/pointer (with the exception of first person pronouns) must be present in the nominal subject relation as the second component. For example, the instance *you deserve no husband* is labelled as cyberbullying based on the following relations: *nsubj*(deserve-2, you-1), *neg*(husband-4, no-3), and *dobj*(deserves-2, husband-4). Again, these four variations apply only to full overt negation constructions, since all cyberbullying elements are required to be explicitly present.

The negation modifier, the possession modifier, and the direct object relations together capture cyberbullying instances when the verb of the negation and direct object relations is the same positive verb, and the first constituent of the possession modifier relation is the same as the second constituent of the direct object relation. In addition, the first constituent of the possession modifier must be a second or third person possessive pronoun, or a proper name. For instance, the sentence *I don't like your face* is categorised as cyberbullying on the basis of *neg*(like-4, n't-3), *poss*(face-6, your-5) and *dobj*(like-4, face-6) relations. Like with all transitive constructions, this rule applies only to full overt negation constructions.

The root, the prepositional modifier, and the prepositional object relations capture explicit negation-based cyberbullying instances that are achieved by means of the adverb/preposition *except* or the prepositional phrase *with…exception.* For instance, the sentence *All in your year are nice except you* is categorised as cyberbullying based on the relations *root*(ROOT-0, nice-6), *prep*(nice-6, except-7) and *pobj*(except-7, you-8), where the preposition relations must modify the root's second constituent which is a positive term, and the object of the preposition must be a personal marker/pointer (except first person pronouns).This rule applies also only to full overt negation constructions.

The nominal subject, the adverbial modifier, and the root relations capture cyberbullying instances that use adverbs such as *hardly* or *barely.* The nominal subject of the sentence must be a personal marker/pointer (except first person pronouns), while the verb phrase described by the root relation must contain a positive term. In addition, the verb phrase must be modified by one of the negative adverbs. For instance, the sentence *You are hardly worth anything* is classified as cyberbullying on the basis of *nsubj*(worth-4, You-1), *advmod*(worth-4, hardly-3), and *root*(ROOT-0, worth-4) relations. This rule can be applied to both overt negation constructions.

Two nominal subject relations together with the conjunct and the negation modifier relations capture cyberbullying instances when the first constituent of the first nominal subject relation is a positive adjective, noun, or verb that is also present as the first constituent of the conjunct relation. In addition, the negation modifier must modify the verb of the second clause, and the subject of the second clause must be a personal marker/pointer (except first person pronouns).

For instance, sentences such as *She rocks, but you don't* and *She is nice and you are not* are labelled as cyberbullying on the basis of the following relation sets: *nsubj*(rocks-2, she-1), *nsubj*(do-5, you-4), *conj*(rocks-2, do-5), and *neg*(do-5, n't-6), and *nsub*j(nice-3, she-1), *nsubj*(are-6, you-5), *conj*(nice-3, are-6), and *neg*(are-6, not-7), respectively. A variation of the previous rule captures similar sentences, such as *she is a nice person, and you are not*, but it requires an additional relation, the adjectival modifier relation, in which case its second component (the adjective) must represent the positive constituent. For example, the relations *nsubj*(person-5, she-1), *amod*(person-5, nice-4), *nsubj*(are-8, you-7), *conj*(person-5, are-8)and *neg*(are-8, not-9), can qualify instances as cyberbullying. This rule and its variation also can be applied to both types of overt negation constructions.

A rule that captures similar instances as the previous one, except that there is no explicit conjunction, combines two nominal subject relations with the parataxis and the negation modifier relations. The first nominal subject relation must contain as its first component a positive modifier which must also be found in the parataxis relation, and a personal marker as its second component. The second nominal subject relation must also contain a personal marker/pointer, but different from the first one. Finally, the negation modifier must be applied to the verb of the second clause. For instance, the sentence *She is beautiful, you are not* is classified as cyberbullying based on the *nsubj*(beautiful-3, she-1), *nsubj*(are-6, you-5), *parataxis*(beautiful-3, are-6), *neg*(are-6, not-7) relations. A variation of this rule involves an additional relation, the adjectival modifier relation. For example, the sentence *she is a nice person, you are not* is labelled as cyberbullying based on *nsubj*(person-5, she-1), *amod*(person-5, nice-4), *nsubj*(are-8, you-7), *parataxis*(person-5, are-8), and *neg*(are-8, not-9) relations. Again, this rule and its variation can be applied to both types of overt negation constructions.

Two nominal subject relations and the clausal complement relation capture instances that contain negative verbs, such as *doubt*. The two different nominal subject relations must contain two different subjects which must belong to the people cyberbullying referential domain, but only the second one must be a personal marker. On the other hand, the first component of the complement relation must be the same verb as the first component of the first nominal subject relation. In addition, the second component of the complement relation must be a positive adjective or noun. For example, the instance *I doubt you are smart* is labelled as cyberbullying based on the following relations: *nsubj*(doubt-2, I-1), *nsubj*(smart-5, you-3), and *ccomp*(doubt-2, smart-5). This rule applies to both overt negation constructions.

The negation modifier and the root relations capture explicit negated cyberbullying instances when the first constituent of the negation modifier relation is the same as the second constituent of the root relation, which must be a positive person referring noun that can act as a personal marker. For instance, the sentence *Not a genius!* is labelled as cyberbullying based on the *neg*(genius-3, Not-1), and *root*(ROOT-0, genius-3) relations. This rule also applies to both overt negation constructions.

The negation modifier, the adjectival modifier, and the root relations capture similar instances as above. For example, the sentence *Not a nice girl! i*s labelled as cyberbullying based on the *neg*(girl-4, Not-1), *amod*(girl-4, nice-3), and *root*(ROOT-0, girl-4) relations. As can be seen, the first constituents of the negation and adjectival modifier relations, and the second constituent of the root relation must be the same positive or neutral person referring noun that can act as a personal marker, while the second constituent of the adjectival modifier must be a positive adjective. Similarly, this rule applies to both overt negation constructions.

The negation modifier, the possession modifier, the adjectival modifier, and the root relations can capture both types of overt negation cyberbullying instances when the first constituents of the negation, the possession and the adjectival modifiers are the same as the second constituent of the root relation, being represented by a neutral or positive noun. Moreover, the second constituent of the possession modifier relation must be a personal marker in the form of a possessive (except first person possessive) pronoun, and the second constituent of the adjectival modifier relation must be a positive adjective (superlative). For instance, the sentence *Not your brightest idea!* can be classified as cyberbullying based on the following relations: *neg*(idea-4, Not-1), *poss*(idea-4, your-2), *amod*(idea-4, brightest-3), and *root*(ROOT-0, idea-4).

The nominal subject, the negation modifier, the prepositional modifier, and the prepositional object relations capture instances of overt negation-based public textual cyberbullying when the first constituents of the nominal subject, the negation modifier, and the prepositional modifier are represented by the same positive or neutral verb. In addition, the second constituent of the prepositional modifier relation and the first constituent of the prepositional object relation must constitute the same preposition. To account for the presence of the personal marker/pointer, the prepositional object relation' second constituent must be a personal pronoun (except first person pronouns), or a proper name. For example, the instance *He doesn't care about you* is labelled as cyberbullying based on the following relations: *nsubj*(care-4, he-1), *neg*(care-4, n't-3), *prep*(care-4, about-5), and *pobj*(about-5, you-6). This rule applies to full overt negation constructions only, since all elements must be explicitly present.

4.4.2. Covert Negation Rules

We have also developed three rules to capture those cyberbullying instances that use covert negation for the dysphemistic element. Similar to the rules designed to capture instances of overt negation cyberbullying, these rules also ignore any explicit cyberbullying terms/expressions, such as profane, insulting or violent. In addition, these rules state that instances must not contain any negation relation that targets any of the cyberbullying elements, to avoid labelling as cyberbullying instances such as *you think you are not pretty* or *you don't think you are pretty*. The three rules apply to both types of covert negation constructions.

The nominal subject and the clausal complement relations together capture implicit negations of positive attributes of a person or group of people. For instance, the sentence *You think you are pretty* implies that in fact you are not pretty, it is only your belief/thought. This rule states that there should be two nominal subject relations with the same second constituent that is a personal marker/pointer in the form of a second and third personal pronoun. In addition, the first constituents of the fist nominal subject and clausal complement relations must be the same verb, and the second constituent of the clausal complement relation must be an adjective with positive connotations. Thus, the sentence *you think you are pretty* is labelled as cyberbullying based on the *nsubj*(think-2, you-1) and *ccomp*(think-2, pretty-5) relations.

The nominal subject, open clausal complement and the direct object relations are also intended to capture implied negations of positive attributes of a person or group of people. For instance, the sentence *You are trying to sound smart* implies that you are not smart, despite the obvious efforts. This is captured by the fact that the main verb implies making an effort (for instance, *try, seek, attempt, essay, assay*), the nominal subject is a personal marker or pointer (excluding first person pronouns), and the direct object of the open clausal complement of the main verb is a noun or adjective with positive connotations, as shown by the *nsubj*(trying-3, you-1), *xcomp*(trying-3, sound-5), and *dobj*(sound-5, smart-6) relations.

The nominal subject, the direct object, and the adverb modifier relations capture implicit negation-based cyberbullying that contain verbs expressing beliefs/opinions about oneself, such as *think.* These verbs must be present as the first constituents of the nominal subject and direct object relations. The second constituents of the both nominal and direct object relations must be the same personal marker/pointer (second or third personal pronoun) which must be further modified by an adverb with positive connotations. For instance, the sentence *you think you pretty* is labelled as cyberbullying based on the following relations: *nsubj*(think-2, you-1), *dobj*(think-2, you-3), and *advmod*(you-3, pretty-4).

5. Results and discussion

To experimentally evaluate the performance of our approach to detecting negation-based instances of public textual cyberbullying, we used the standard metrics of precision, recall and F1-measure, as well as accuracy (Goncalves 2011). These metrics are relative to the given label of cyberbullying and were represented as percentage. Accuracy is described in terms of the total number of correctly assigned labels relative to the total number of instances or labels, using the following formula:

$$(6) \qquad Accuracy = \frac{total\ number\ of\ correctly\ assigned\ labels}{total\ number\ of\ instances}\ X\ 100;$$

Precision is represented by the fraction of all instances that were correctly assigned the label of cyberbullying by the detection system (true positives) out of the total number of instances that the system assigned the label of cyberbullying, both correctly and incorrectly (true positives + false positives), while recall is represented by the fraction of instances labelled correctly as cyberbullying by the system (true positives) out of all the instances manually labelled by the annotators as cyberbullying (true positives + false negatives). The two equations below describe precision and recall, respectively, in terms of percentage:

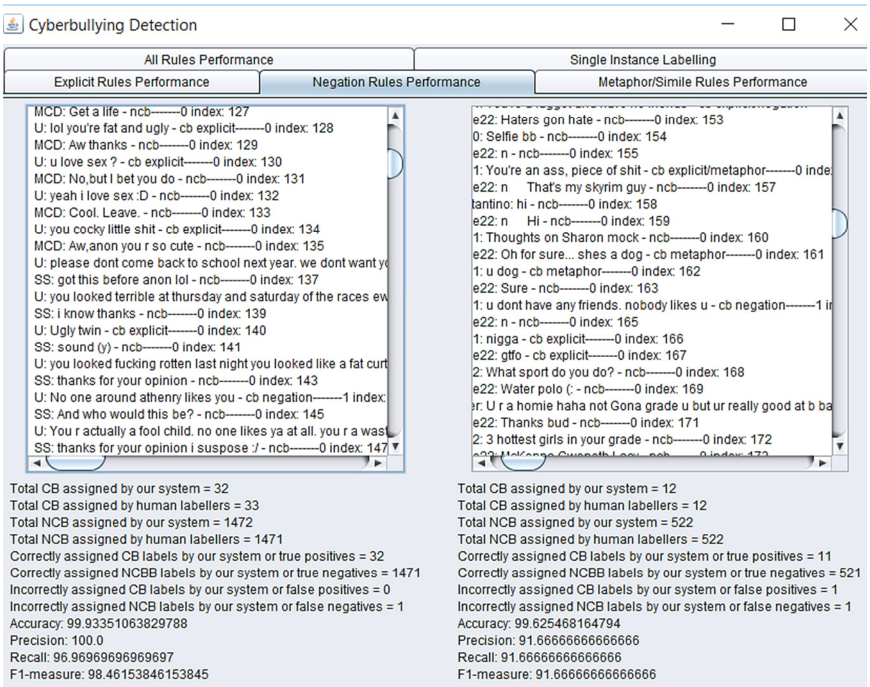
$$(7) \qquad Precision = \frac{number\ of\ correctly\ labelled\ instances\ as\ cyberbullying}{total\ number\ of instancecs\ labelled\ as\ cyberbullying}\ X\ 100;$$

(8) $$Recall = \frac{number\ of\ correctly\ labelled\ instances\ as\ cyberbullying}{number\ of\ instances\ manually\ labelled\ \ as\ cyberbullying}\ X\ 100;$$

Precision and recall can be combined into a single measure, the F-measure, allowing for different weights to be assigned to either precision or recall (Goncalves, 2011). Here we used the F1-measure (or the harmonic mean) by which both metrics, precision and recall, are given the same weight, since we believe that precision and recall are equally important in determining whether the detection approach is successful. To compute the harmonic mean we have used the following formula:

(9) $$F1 - measure = \frac{2\ X\ (Precision\ X\ Recall)}{Precision+Recall}\ X\ 100;$$

We applied the covert and overt negation rules to the development and test datasets using the application[12] that we have designed especially for applying and testing detection rules, and the results that we obtained in this experiment are depicted in Figure 1 which shows a screenshot of the application running when the tab for negation rules is selected.


Cyberbullying Detection
All Rules Performance
Single Instance Labelling
Explicit Rules Performance
Negation Rules Performance
Metaphor/Simile Rules Performance

```
MCD: Get a life - ncb-------0 index: 127
U: lol you're fat and ugly - cb explicit-------0 index: 128
MCD: Aw thanks - ncb-------0 index: 129
U: u love sex ? - cb explicit-------0 index: 130
MCD: No,but I bet you do - ncb-------0 index: 131
U: yeah i love sex :D - ncb-------0 index: 132
MCD: Cool. Leave. - ncb-------0 index: 133
U: you cocky little shit - cb explicit-------0 index: 134
MCD: Aw,anon you r so cute - ncb-------0 index: 135
U: please dont come back to school next year. we dont want y
SS: got this before anon lol - ncb-------0 index: 137
U: you looked terrible at thursday and saturday of the races ew
SS: i know thanks - ncb-------0 index: 139
U: Ugly twin - cb explicit-------0 index: 140
SS: sound (y) - ncb-------0 index: 141
U: you looked fucking rotten last night you looked like a fat curt
SS: thanks for your opinion - ncb-------0 index: 143
U: No one around athenry likes you - cb negation-------1 index:
SS: And who would this be? - ncb-------0 index: 145
U: You r actually a fool child. no one likes ya at all. you r a wast
SS: thanks for your opinion i suspose :/ - ncb-------0 index: 147
```

```
Total CB assigned by our system = 32
Total CB assigned by human labellers = 33
Total NCB assigned by our system = 1472
Total NCB assigned by human labellers = 1471
Correctly assigned CB labels by our system or true positives = 32
Correctly assigned NCBB labels by our system or true negatives = 1471
Incorrectly assigned CB labels by our system or false positives = 0
Incorrectly assigned NCB labels by our system or false negatives = 1
Accuracy: 99.93351063829788
Precision: 100.0
Recall: 96.96969696969697
F1-measure: 98.46153846153845
```

```
e22: Haters gon hate - ncb-------0 index: 153
0: Selfie bb - ncb-------0 index: 154
e22: n - ncb-------0 index: 155
1: You're an ass, piece of shit - cb explicit/metaphor-------0 inde
e22: n     That's my skyrim guy - ncb-------0 index: 157
tantino: hi - ncb-------0 index: 158
e22: n     Hi - ncb-------0 index: 159
1: Thoughts on Sharon mock - ncb-------0 index: 160
e22: Oh for sure... shes a dog - cb metaphor-------0 index: 161
1: u dog - cb metaphor-------0 index: 162
e22: Sure - ncb-------0 index: 163
1: u dont have any friends. nobody likes u - cb negation-------1 i
e22: n - ncb-------0 index: 165
1: nigga - cb explicit-------0 index: 166
e22: gtfo - cb explicit-------0 index: 167
2: What sport do you do? - ncb-------0 index: 168
e22: Water polo (: - ncb-------0 index: 169
r: U r a homie haha not Gona grade u but ur really good at b ba
e22: Thanks bud - ncb-------0 index: 171
2: 3 hottest girls in your grade - ncb-------0 index: 172
```

```
Total CB assigned by our system = 12
Total CB assigned by human labellers = 12
Total NCB assigned by our system = 522
Total NCB assigned by human labellers = 522
Correctly assigned CB labels by our system or true positives = 11
Correctly assigned NCBB labels by our system or true negatives = 521
Incorrectly assigned CB labels by our system or false positives = 1
Incorrectly assigned NCB labels by our system or false negatives = 1
Accuracy: 99.625468164794
Precision: 91.66666666666666
Recall: 91.66666666666666
F1-measure: 91.66666666666666
```

FIGURE 1. OUR APPLICATION RUNNING, SHOWING THE PERFORMANCE OF THE NEGATION RULES WHEN APPLIED TO THE DEVELOPMENT DATASET AND THE TEST DATASET.

[12] The application has additional functionalities, that are not discussed in the present paper. They include providing performance levels for other sets of rules, such as rules designed to capture explicit instances or metaphoric instances, as well as performance levels for the entire detection system. The application also provides a facility to label individual instances.

The results of applying the covert and overt negation rules to the development and test datasets are formally presented in Table 1 as precision, recall, F1-measure, and accuracy scores.

	%Accuracy	**%Precision**	**%Recall**	**%F1-measure**
Development Dataset	99.93	100.0	96.96	98.46
Test Dataset	99.62	91.66	91.66	91.66

TABLE 1. PERFORMANCE OF OUR APPROACH IN THE TASK OF DETECTING NEGATION-BASED CYBERBULLYING INSTANCES.

These scores indicate that the performance of our approach closely resembles the performance of the human annotators on the development dataset, while on the test dataset, despite the expected lower performance, it also remains close to the human performance. Moreover, the differences in scores between the development dataset and the test dataset are relatively low, suggesting that the rules developed here for detecting negation-based instances have a high degree of generalisation: in terms of accuracy, there is a small difference of 0.31, while the differences in scores for precision and recall are 8.4 and 5.3, respectively. Overall, the difference in the F1-measure scores between the development dataset and the test dataset is relatively low - 6.8. To account for the higher difference in precision and recall scores, we further investigated those instances for which the present rules yielded false positives and false negatives and we found that most errors are due to parsing errors, as well as to the relatively small set of proper names that were included in the lexical database (Power et al. 2017); thus, instances containing Spanish proper names were missed since the database includes only English proper names. In addition, in the case of the precision scores, the higher difference is explained by the fact that, on the development dataset, the present rules achieved the maximum score, as well as by the fact that the test dataset contained fewer negation-based instances.

From a practical, but also, ethical standpoint, designing a system that correctly identifies cyberbullying, but also not-cyberbullying instances, that is, a system that demonstrates not only a high level of true positives, but also a high level of true negatives, may positively impact the victimisation level, as well as the user-retention levels; in turn, these levels may influence whether a social platform is more likely to adopt such system. For these reasons, we also looked at the proportion of cyberbullying and not-cyberbullying labels assigned by our approach compared to those assigned by human annotators. The results are shown in Figure 2.

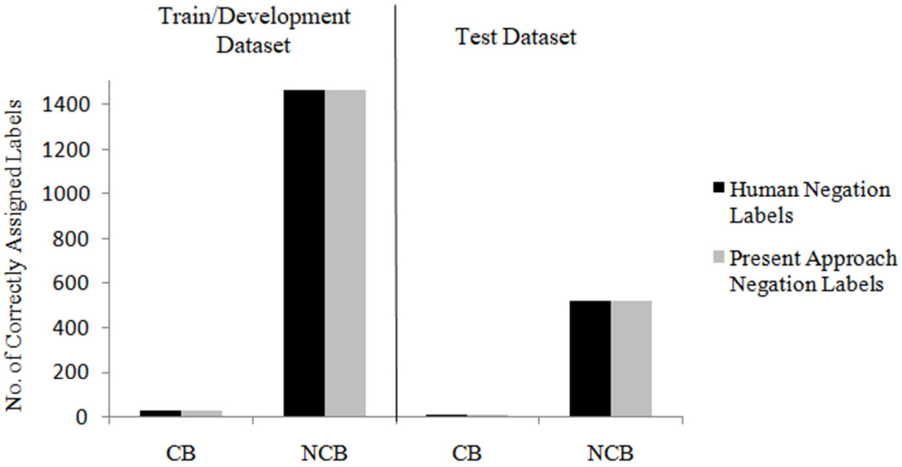


FIGURE 2. NUMBER OF CORRECT NEGATION CB AND NCB LABELS YIELDED BY OUR APPROACH, RELATIVE TO THE NUMBER OF NEGATION CB AND NCB LABELS ASSIGNED BY HUMAN ANNOTATORS, FOR BOTH DATASETS.

In terms of the number of correct negation CB labels (true positives) and the number of correct NCB labels (true negatives), our approach closely resembled the human performance in the task of detecting discourse independent negated cyberbullying instances. As shown in Figure 2, our approach correctly labelled as cyberbullying (CB) 32 instances out of 33 (96.96%) on the development dataset and 11 instances out of 12 (91.66%) on the test dataset. Similar results were obtained for not-cyberbullying instances: our approach correctly labelled as not-cyberbullying (NCB) all 1471 instances (100%) on the development dataset, and 521 instances out of 522 (99.80%) on the test dataset. Thus, the proportion of false negatives is relatively low: 3.04% (1 negation cyberbullying instance was missed) on the development dataset and 8.34% (1 negation cyberbullying instance was missed) on the test dataset.

6. CONCLUSIONS

In this paper, we presented an effective functional approach to detecting public textual cyberbullying. Specifically, we addressed negation-based instances. We first considered the underlying grammatical structures which characterise negation, then the corresponding cyberbullying constructions which were identified using the linguistically motivated definition of cyberbullying which we have advanced in previous work (Power et al. 2017). The definition posits three necessary and sufficient elements to qualify an instance as cyberbullying, namely, the personal marker/pointer, the dysphemistic element, and the link between them. Subsequently, we described the overall mechanism for detecting cyberbullying instances which uses the grammatical and cyberbullying information encapsulated in the cyberbullying lexical database (Power et al. 2017), as well as the grammatical dependencies among sentential components (de Marneffe and Manning 2008a; 2008b). Finally, we described detection rules for both overt and covert negated forms of public textual cyberbullying, and applied them to two datasets – a development dataset and a test dataset; the results indicate that our approach closely approximates human performance on both datasets, across accuracy, precision, recall, and the harmonic mean (F1-measure).

Despite the high level performance achieved by our approach, there are several areas which future research might consider. First, from a generalisation standpoint, only online interactions in English were targeted. Future research might consider expanding the capabilities of the detection system to other languages, by either incorporating translation, or by designing dependency parsers, lexical databases, and detection rules specific to a given language. In addition, only English proper names were considered presently, and given that online interaction is not constrained by geographical boundaries, future research might include in the cyberbullying lexical database proper names that originate in other languages. Secondly, we discarded any smileys and emoticons, since they rarely occurred. However, we recognise that, in other contexts and other datasets, the presence of such emoticons, such as frown, or angry faces, may be indicative of cyberbullying, as shown by Ptaszynski et al. (2010) and Ptaszynski et al. (2016). In addition, although not encountered in the present datasets, Unicode (2017) allows not only emoticons to be inserted in text, but also other symbols for gestures or animals that may constitute cyberbullying, such as fist-making or monkey face. Nevertheless, the detection model that we proposed here can easily be extended to account for such instances, by associating each combination of characters with a lexical entry that can be already found in the cyberbullying lexical database, or by creating a new lexical entry that encodes the required linguistic and cyberbullying information. For instance, the Unicode combination of *U+1F64A* represents a monkey (Unicode 2017), and like with the acronym and abbreviation resolution procedure, one can replace such character sequence with the word *monkey*, and, subsequently, apply the same pre-processing techniques and detection rules described in the present paper.

References

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: 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." https://nlp.stanford.edu/software/dependencies_manual.pdf.

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: Corw2017.

Horn, L. R. 1989. *A Natural History of Negation*. Chicago: University of Chicago 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. 2017. "Internet Slang – Internet Dictionary." Accessed October 19. 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.

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

Lawler, J. 2005. "Negation and NPIs." http://www.umich.edu/~jlawler/NPIs.pdf

Livingstone, S.,Haddon, L., Görzig, A., and Ólafsson, K. 2011. "EU Kids Online: final report 2011."
http://eprints.lse.ac.uk/45490/1/EU%20Kids%20Online%20final%20report%202011%28lsero%29.pdf.

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". http://eprints.lse.ac.uk/60513/1/__lse.ac.uk_storage_LIBRARY_Secondary_libfile_shared_repository_Content_EU%20Kids%20Online_EU%20Kids%20Online-Children%27s%20online%20risks_2014.pdf.

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." Accessed October 19. http://norvig.com/spell-correct.html.

Oracle. 2017. Java™ Platform, Standard Edition 9 API Specification. Accessed October 19. https://docs.oracle.com/javase/9/docs/api/index.html?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.*

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. 2017. "Emoticons." Accessed October 19. 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.

Witten, I.H., Frank, E., and Hall, M.A. 2011. Data Mining: Practical Machine Learning Tools and Techniques (3rd edition). Elsevier Inc., USA.

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.