Detection and Fine-Grained Classification of Cyberbullying Events

Cynthia Van Hee[1], Els Lefever[1], Ben Verhoeven[2], Julie Mennes[1], Bart Desmet [1]
Guy De Pauw[2], Walter Daelemans[2] and Véronique Hoste[1]

[1]LT3 - Language and Translation Technology Team, Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
[2]CLiPS - Computational Linguistics Group, University of Antwerp
Prinsstraat 13, 2000 Antwerp, Belgium

Abstract

In the current era of online interactions, both positive and negative experiences are abundant on the Web. As in real life, negative experiences can have a serious impact on youngsters. Recent studies have reported cybervictimization rates among teenagers that vary between 20% and 40%. In this paper, we focus on cyberbullying as a particular form of cybervictimization and explore its automatic detection and fine-grained classification. Data containing cyberbullying was collected from the social networking site Ask.fm. We developed and applied a new scheme for cyberbullying annotation, which describes the presence and severity of cyberbullying, a post author's role (harasser, victim or bystander) and a number of fine-grained categories related to cyberbullying, such as insults and threats. We present experimental results on the automatic detection of cyberbullying and explore the feasibility of detecting the more fine-grained cyberbullying categories in online posts. For the first task, an F-score of 55.39% is obtained. We observe that the detection of the fine-grained categories (e.g. threats) is more challenging, presumably due to data sparsity, and because they are often expressed in a subtle and implicit way.

1 Introduction

Young people are gaining more frequent and rapid access to online, mobile and networked media. Although most of the time, children's Internet use is harmless, there are some risks associated with the online activity, such as the use of social networking sites (e.g. Facebook). The anonymity and freedom provided by social networks makes children vulnerable to threatening situations on the Web, such as grooming by paedophiles or cyberbullying.

According to Smith et al. (2008), cyberbullying is defined as *an aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time against a victim who cannot easily defend him or herself.* Their definition is based on three criteria (repetitiveness, intentionality, and an imbalance of power between the harasser and the victim) that are recognized as inherent characteristics of bullying by Olweus (1996). Some doubt exists, nevertheless, as to whether all three criteria are necessary conditions for cyberbullying. For example Dooley and Cross (2010) and Grigg (2010) stress that online posts are persistent, a single aggressive act can result in continued and widespread ridicule for the victim. Furthermore, it is hard to decide upon intentionality since online communication is prone to misinterpretation (Kiesler et al., 1984; Vandebosch et al., 2006). Finally, the assessment of a power imbalance is complicated in online bullying as it may be related to ICT proficiency, anonymity or the inability of victims to get away (Dooley and Cross, 2010). In general, when working with social media data, the available context is often limited. This makes it hard to decide upon the repetitive character of a cyberbullying incident, to determine whether the victim of an aggressive act is able to defend himself or to decide whether the bully is acting intentionally. Considering these limitations, we restrict the scope of our research to the detection of textual content that is published online by an individual and that is aggressive or hurtful against a victim.

Tokunaga (2010) analyzed a body of quantitative research on cyberbullying and found that cybervictimization rates vary between 20% and 40% on average (Dehue et al., 2006; Hinduja and Patchin, 2006; Li, 2007; Smith et al., 2008; Ybarra and Mitchell, 2008). The rate varies among different studies depending on location, interval and the conceptualisations researchers use in describing cyberbullying. Indeed, according to The EU Kids Online Report (2014)[1], 17% of 9 to 16 year olds had been bothered or upset by something online in the past year, whereas Juvonen et al. (2008) found that no less than 72% of 12 to 17 year olds had encountered cyberbullying at least once within the year preceding the questionnaire. According to a recent study by Van Cleemput et al. (2013), 11% of 2,000 Flemish secondary school students had been bullied online at least once in the six months preceding the survey. These figures demonstrate that cyberbullying is not a rare phenomenon. Evidently, it can have a serious impact on children's and youngsters' well-being. This is shown by a number of studies that link cyberbullying to depression, school problems, low self-esteem and even self-harm (Price and Dalgleish, 2010; Šléglová and Černá, 2011; Vandebosch et al., 2006). It is therefore of key importance to identify possibly threatening situations on the Web before they can cause harm.

As it is unfeasible for humans to keep track of all conversations produced online, researchers have started to explore automatic procedures for signalling harmful content. This would allow for large-scale social media monitoring and early detection of harmful situations including cyberbullying. Research has also focussed on the desirability of such automatic systems. Van Royen et al. (2015), for example, found that a major part of their respondents favoured automatic monitoring, provided that effective follow-up strategies are included and that privacy and autonomy are guaranteed. Reynolds et al. (2011), Dinakar et al. (2012), and Dadvar et al. (2014) describe some of the first forays into the automatic detection of cyberbullying. However, to the best of our knowledge, we present the first study on recognizing cyberbullying events in social media content by means of a fine-grained textual annotation of the corpus, rather than implementing a binary distinction (i.e. cyberbullying versus non-cyberbullying).

For the annotation of the data, we consider fine-grained categories related to cyberbullying such as insults and threats. Implementing this fine-grained distinction allows for insight into various types of cyberbullying and the degree to which they are alarming (e.g. threats are considered more alarming than a single insult). Additionally, the annotation scheme allows to identify, for each cyberbullying post, the role of the author (i.e. bully, victim, bystander) and the harmfulness. The idea is that this information allows a more detailed reconstruction of cyberbullying events, which can be used to enhance the detection process.

We present experiments on the identification of cyberbullying events and the classification of online posts in fine-grained categories related to cyberbullying. The focus of our experiments is on a Dutch dataset, but the technique is language-independent, provided there is annotated data available in the target language.

2 Related Research

Cyberbullying is a widely covered research topic in the realm of social sciences and psychology. A fair amount of research has been done on the definition and occurrence of the phenomenon (Livingstone et al., 2010; Hinduja and Patchin, 2012; Slonje and Smith, 2008), the identification of different forms of cyberbullying (O'Sullivan and Flanagin, 2003; Vandebosch and Cleemput, 2009; Willard, 2007) and the consequences of cyberbullying (Cowie, 2013; Price and Dalgleish, 2010; Smith et al., 2008). By contrast, the number of studies that focus on the annotation and automatic detection of cyberbullying is limited.

Yin et al. (2009) applied a supervised machine learning approach to the automatic detection of cyberharassment by representing each post in their corpus by local tf-idf features, sentiment features and features capturing the similarity between posts, assuming that posts which are significantly different from their neighbors are more likely to contain cyberbullying. By combining all features, they obtain an F-score of 0.44. Dinakar et al. (2012) conducted text classification experiments on YouTube data. They adopted a bag-of-words supervised machine learning classification approach to identify the sensitive topic for a cyberbullying post (i.e. sexuality, intelligence or race and culture) and report an averaged F-score of 0.63. Reynolds et al. (2011) compared a rule-based model to a bag-of-words model for detecting cyberbullying posts and found that rule-based learning with a number of lexical features (e.g. the number of curse words in a post) outperformed the bag-of-words model. Dadvar et al.

[1] http://lsedesignunit.com/EUKidsOnline

(2014) combined the potential of machine learning algorithms with information from social studies for the automatic recognition of cyberbullying. User information and expert views were used in addition to textual features, which resulted in a classification performance of F = 0.64. Nahar et al. (2014) applied a fuzzy SVM algorithm for cyberbullying detection. They implemented a number of lexical features (e.g. the number of swearwords and capitalized words), sentiment features and features based on metadata (e.g. the user's age and gender) and report an F-score of 47%. In all of the aforementioned studies, cyberbullying detection was approached as a binary classification task (cyberbullying -vs- non-cyberbullying). In this paper, specific forms of cyberbullying like threats and insults are taken into account as fine-grained categories. Moreover, we aim to detect cyberbullying events and therefor consider posts from harassers as well as from victims and bystanders. We present two sets of experiments in which we explore 1) the detection of cyberbullying posts regardless of the author's role (i.e. *cyberbullying events*) and 2) the identification of fine-grained text categories related to cyberbullying.

The remainder of this paper is organized as follows: Section 3 describes our experimental corpus as well as the data collection and annotation. Section 4 gives an overview of the experimental setup. The results are discussed in Section 5 and we formulate conclusions and directions for future research in Section 6.

3 Dataset Construction and Annotation

3.1 Data Collection

The data was collected from Ask.fm[2], a social networking site where users can ask and answer questions to each other, with the option of doing so anonymously. Typically, Ask.fm data consists of question-answer pairs published on a user's profile. We retrieved the data using GNU Wget[3] and crawled a number of randomly chosen seed sites. Although the seed profiles were chosen to be of a user with Dutch as mother-tongue, the crawled data contained a fair amount of non-Dutch data (12,954 posts). The non-Dutch posts were filtered out, which resulted in our experimental corpus containing 85,485 Dutch posts.

3.2 Data Annotation

To operationalize the task of automatic cyberbullying detection, we developed and tested a fine-grained annotation scheme detailed in Van Hee et al. (2015), and applied it to our corpus. To provide the annotators with some context, all posts were presented within their original conversation when possible. The annotation scheme describes two levels of annotation. Firstly, the annotators were asked to indicate, at the post level, whether a post is part of a cyberbullying event. This was done with a harmfulness score on a three-point scale, with 0 signifying that the post does not contain indications of cyberbullying, 1 that the post contains indications of cyberbullying, although they are not severe, and 2 that the post contains serious indications of cyberbullying (e.g. physical threats or incitements to commit suicide). When a post is considered to be part of a cyberbullying event (i.e. its score is 1 or 2), annotators identify the author's role (i.e. harasser, victim or bystander). Two types of bystanders are distinguished in this annotation scheme: 1) bystanders who help the victim and discourage the harasser from continuing his actions (i.e. *bystander-defender*) and 2) bystanders who do not initiate, but take part in the actions of the harasser (i.e. *bystander-assistant*).

Secondly, at the subsentence level, the annotators were tasked with the identification of fine-grained text categories related to cyberbullying. More concretely, they identified all text spans corresponding to one of the categories described in the annotation scheme. For our experiments we focussed on the cyberbullying-related text categories that are described below.

- **Threat/Blackmail:** expressions containing physical or psychological threats or indications of blackmail (e.g. *My fist is itching to punch you so hard in the face*).

- **Insult:** expressions containing abusive, degrading or offensive language that are meant to insult the addressee (e.g. *You're a sad little fuck*).

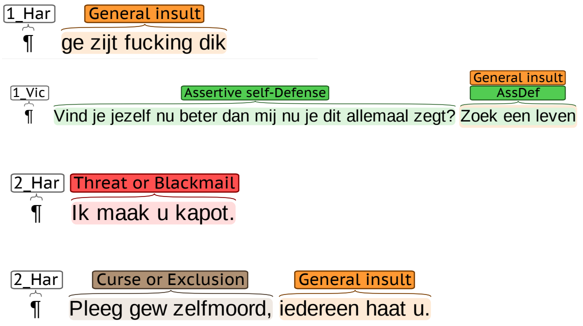
- **Curse/Exclusion:** expressions of a wish that some form of adversity or misfortune will befall the victim and expressions that exclude the victim from a conversation or a social group (e.g. *Just kill yourself*).

[2] http://ask.fm

[3] https://www.gnu.org/software/wget

- **Defamation:** expressions that reveal confident or defamatory information about the victim to a large public or expressions that ridicule the victim in public (e.g. *She's a whore and she'll influence you to be one too*).
- **Sexual talk:** expressions with a sexual meaning that are possibly harmful (e.g. *Post a naked pic, now!!*).
- **Defense:** expressions in support of the victim, expressed by the victim himself or by a bystander (e.g. *Shut up about my sister, she is not a slut!*)
- **Encouragement to the harasser:** expressions in support of the harasser (e.g. *Haha, you're so right, he's a nobody*)

We refer to our technical report for a complete overview of the annotation guidelines, including practical remarks and notes. All annotations were done using the brat rapid annotation tool (Stenetorp et al., 2012). Below are given some annotation examples from our dataset.



¶ ge zijt fucking dik

¶ Vind je jezelf nu beter dan mij nu je dit allemaal zegt? Zoek een leven

¶ Ik maak u kapot.

¶ Pleeg gew zelfmoord, iedereen haat u.

As shown in the annotation examples, the author's role and harmfulness score are indicated on the pilcrow sign preceding each post. The example posts contain a general insult (*Ge zijt fucking dik*, "you are fucking fat"), a defense (*Vind je jezelf nu beter dan mij nu je dit allemaal zegt? Zoek een leven*, "Do you think that saying this makes you a better person than I am? Get a life"), a threat (*Ik maak u kapot*, "I will destroy you") and a curse (*Pleeg gew zelfmoord*, "Just kill yourself").

In total, 85,485 Dutch posts were annotated by two annotators. To demonstrate the validity of our guidelines, inter-annotator agreement scores were calculated using Kappa (Cohen, 1960) and F-score[4] on a subset of the corpus (~6,500 posts). Kappa scores for the fine-grained categories range from substantial (0.69) to moderate (0.19), except for the category *Defamation*, whose identification seems to be rather difficult.

Annotation	**Kappa**	**F-score**
Cyberbullying -vs-non-cyberbullying	0.69	0.69
Author's role	0.65	0.63
Threat/Blackmail	0.52	0.53
Insult	0.66	0.68
Curse/Exclusion	0.19	0.20
Defamation	0	0
Sexual Talk	0.53	0.54
Defense	0.57	0.58
Encouragement to the harasser	0.21	0.21

Table 1: Inter-annotator agreement scores for the annotation of cyberbullying events, the author's role, and the fine-grained categories.

3.3 Experimental Corpus

The resulting experimental corpus of 85,485 Dutch posts features a skewed class distribution with the large majority of posts not referring to any cyberbullying event. In total, there were 5,685 cyberbullying events (i.e. posts containing at least one of the categories mentioned below), which corresponds to the ratio 1:15. As a cyberbullying event are considered all posts that are given a harmfulness score of 1 or 2.

Category	**# Positive posts**	**Ratio**
Threat/Blackmail	204	~1:418
Insult	4,276	~1:19
Curse/Exclusion	1,110	~1:76
Defamation	162	~1:527
Sexual talk	498	~1:171
Defense	2,218	~1:37
Encouragements to the harasser	42	~1:2,034

Table 2: Data distribution for the fine-grained text categories related to cyberbullying.

In what relates to the fine-grained cyberbullying categories, we can infer from Table 2 that insults are the most frequent type of cyberbullying activity in our data, followed by defense statements and curse/exclusion posts. *Encouragements to the harasser* is the least represented category, with a ratio of 1:2,034. It should be noted that in case the annotators had too little context at their disposal to discern encouragements by bystanders from bul-

[4] F-score is an evaluation measure that is the weighted average of precision and recall.

lying acts by bullies, they annotated the post as a bullying act.

Table 3 presents the different roles in the annotated bullying posts: the role of bully features in more than half of the annotated posts, followed by the victim role in about 30% of the posts. The bystander role in its two different subroles accounts for about 10% of the experimental corpus.

Author's role	Harmfulness	# Posts
Harasser	1	3085
Harasser	2	181
Victim	1	1671
Victim	2	129
Bystander-defender	1	546
Bystander-defender	2	23
Bystander-assistant	1	49
Bystander-assistant	2	1

Table 3: Data distribution for the different author roles in cyberbullying events.

4 Experiments

This section describes the experiments that were conducted to gain insight into the detection and fine-grained classification of cyberbullying events.

4.1 Experimental setup

Two sets of experiments were conducted. Firstly, we explored the detection of cyberbullying posts regardless of the harmfulness score (i.e. we considered posts that were given a score of 1 or 2) and the author's role. The second set of experiments focuses on a more complex task, the identification of fine-grained text categories related to cyberbullying (see Section 3.2). To this end, a binary classifier was built for each category.

We used Support Vector Machines (SVM) as the classification algorithm since they have proven to work well for high-skew text classification tasks similar to the ones under investigation (Desmet and Hoste, 2014). We used linear kernels and experimentally determined the optimal cost value c to be 1. All experiments were carried out using Pattern (De Smedt and Daelemans, 2012a), a Python package for data mining, natural language processing and machine learning. As preprocessing steps, we applied tokenization, PoS-tagging and lemmatization to the data using the LeTs Preprocess Toolkit (van de Kauter et al., 2013).

4.2 Features

We experimentally tested whether cyberbullying events and fine-grained categories related to cyberbullying can be recognized by lexical markers in a post. To this end, all posts were represented by a number of standard NLP features including bag-of-words features and sentiment lexicon features:

- **Word n-gram bags-of-words:** binary features indicating the presence of word unigrams and bigrams.
- **Character n-gram bag-of-words:** binary features indicating the presence of character trigrams (without crossing word boundaries), to provide some abstraction from the word level.
- **Sentiment lexicon features:** four numeric features representing the number of positive, negative, and neutral lexicon words (averaged over text length) and the overall post polarity (i.e. the sum of the values of identified sentiment words averaged over text length)[5]. The features were calculated based on existing sentiment lexicons for Dutch (De Smedt and Daelemans, 2012b; Jijkoun and Hofmann, 2009).

5 Results

We implemented different experimental set-ups with various feature groups and hence determined the informativeness of each feature group for the current classification tasks. We explored the contributiveness of the following feature groups in isolation: word unigram bag-of-words (which can be considered as the baseline approach), word bigram bag-of-words, character trigram bag-of-words, and sentiment lexicon features. In addition, all feature groups were combined (*full system*). The results obtained for the cyberbullying event detection and the more fine-grained classification task are described in Section 5.1 and Section 5.2, respectively. A general discussion of the results can be found in Section 5.3.

5.1 Cyberbullying Event Classification

For the detection of cyberbullying events, the best results are obtained by combining all features groups (F = 55.39%). Considering the scores of

[5] To increase the lexicon coverage, lemmas were taken into account.

	Word unigrams	**Word bigrams**	**Character trigrams**	**Sentiment lexicon**	**Full system**
Cyberbully event	47.94	24.31	33.18	6.35	**55.39**

Table 4: F-scores (percentages) obtained for the binary classification of cyberbullying events when using the feature groups in isolation and combined *(full system)*.

	Word unigrams	**Word bigrams**	**Character trigrams**	**Sentiment lexicon**	**Full system**
Threat/blackmail	5.42	0.78	2.48	0.14	**19.84**
Sexual talk	15.42	2.40	10.32	0.91	**35.18**
Insult	47.62	19.44	32.13	4.91	**56.32**
Curse/exclusion	20.06	4.76	9.68	0.96	**33.46**
Defense	22.45	8.17	10.38	2.01	**35.09**
Defamation	1.05	0.36	0.23	0.10	**7.41**
Encouragement	**0.12**	0.10	0.07	0.01	0.00

Table 5: F-scores (percentages) obtained for the classification of fine-grained text categories related to cyberbullying when using the feature groups in isolation and combined *(full system)*.

the separate feature groups, we find that word unigram bag-of-words *(b-o-w)* features are the most contributive features, followed by character trigram b-o-w features. Sentiment lexicon features perform the least well for this task. As shown in Table 6, the system performs better in terms of precision than recall.

5.2 Fine-Grained Classification

In line with the cyberbullying event classification, the performance of the fine-grained classifiers benefits from combining all feature groups. F-scores for the fine-grained classification vary rather strongly, reaching up to 56.32% for the *Insult* category. Just as for the cyberbullying event detection, the most contributive feature groups are the word unigram and character trigram b-o-w features, whereas the sentiment lexicon features are the least informative for the classifier. Table 5 shows that the classification of some fine-grained categories related to cyberbullying is more difficult than that of others: the insults classifier obtains an F-score of 56.32%, whereas the best classification performance for *Encouragement* and *Defamation* remains at 0.12% and 7.41%, respectively. In addition to data scarcity (e.g. only 42 positive posts for the *Encouragement* category), the large discrepancies in performance are presumably due to the extent to which a category is lexicalized. Except for these last two groups, most fine-grained categories also show a good balance between precision and recall (see Table 6).

	Recall	**Precision**
Cyberbully event classification		
Cyberbully event	51.46	59.96
Fine-grained classification		
Threat/Blackmail	25.00	16.45
Sexual talk	36.35	34.09
Insult	53.60	59.33
Curse/Exclusion	32.34	34.65
Defense	31.74	39.22
Defamation	9.88	5.93
Encouragement	0	0

Table 6: Full system performance by means of recall and precision.

5.3 General Discussion

As can be inferred from Table 4 and Table 5, using the feature groups in isolation is insufficient for cyberbullying detection. This is especially clear from the sentiment lexicon features. The poor performance of sentiment features in isolation is in line with the findings of Yin et al. (2009). They argue that the sentiment word coverage is limited by the occurrence of spelling errors in social media content. Furthermore, some cyberbullying posts are hurtful even when they do not contain explicit negative language. Inversely, a post may be very negative while devoid of any form of cyberbullying. Although our experiments show that sentiment lexicon features are not very informative when used in isolation, we believe that they should not be discarded for future work as they may be beneficial to the classification performance when used in a combined feature set.

In this paper, we mainly focussed on lexical (bag-of-words) features. A major limitation of bag-of-words features is that they often result in sparse feature vectors: a large part of the n-grams in the training data only occur in one or two posts. To reduce feature sparseness, we explored the effect of filtering the n-gram features based on their PoS-tags. Hence we only considered nouns, verbs, adjectives and adverbs for the extraction of word unigram and bigram bag-of-word features. However, this filtering decreased the classification performance by 10% on average. The insults classifier suffered the largest drop (16%). A plausible explanation for this drop is that, by considering only words with simplified PoS-tags, pronouns (e.g. *you*), interjections (e.g. *haha*), foreign words (e.g. *putain*), and misspelled words (e.g. *uglyy*) are discarded although they might be relevant for distinguishing between cyberbullying and non-cyberbullying posts.

Although our results show that there is room for improvement, the scores obtained for the binary distinction between cyberbullying and non-cyberbullying are in line with state-of-the-art approaches to automatic cyberbullying detection (e.g. Dadvar et al., 2014; Dinakar et al., 2012). Reynolds et al. (2011) worked with data that is similar to ours (i.e. question-answer pairs) and made use of lexical features including the number of 'bad' words in a post. They obtained an accuracy of 78.5% when the positive posts were overrepresented (their actual presence multiplied by 10) in the training corpus. When the normal distribution was kept, however, the accuracy remained at 53.82%.

Nevertheless, all of the above-mentioned studies mainly focus on the detection of cyberbullying posts that contain insults or curses. The focus of our work is on the detection of cyberbullying events (i.e. posts from victims and bystanders as well as posts from the harasser). Moreover, we aim to detect fine-grained categories related to cyberbullying.

6 Conclusions and future work

As cyberbullying often has an implicit and subtle nature, its detection is not a trivial task. We show promising initial results for the identification of cyberbullying events and the fine-grained classification of insults. For the experiments presented in this paper, we relied on lexical features to gain insight into the difficulty and learnability of the detection and fine-grained classification of cyberbullying. We conclude that especially this fine-grained classification is a very challenging task, which is hindered by data sparseness on the one hand and by the degree to which the categories are lexicalized on the other hand.

The ultimate goal of automatic cyberbullying detection is to reduce manual monitoring efforts on social media. As we want to send as much online threats as possible to the moderator of the network, recall optimization will be a prior focus for further research. We will also explore to what extent author role information can be used to enhance the detection of cyberbullying events. Moreover, implicit realizations of cyberbullying are hard to recognize as they are devoid of lexical cues including profanity. Therefore, we will explore the use of more advanced features (e.g. syntactic patterns, semantic information) in addition to lexical features. Additionally, we will examine feature selection techniques to decrease vector sparseness and hence avoid the introduction of noise. Finally, social media texts tend to deviate from the linguistic norm, which reduces the effectiveness of more complex features. Another direction for future work will therefore be orthographic normalization of the data as a preprocessing step.

All experiments in this paper were conducted on a Dutch dataset. Nevertheless, a set of similar experiments will be carried out on an English dataset that is currently under construction.

Acknowledgments

The work presented in this paper was carried out in the framework of the AMiCA (IWT SBO-project 120007) project, funded by the government agency for Innovation by Science and Technology (IWT).

References

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Helen Cowie. 2013. Cyberbullying and its impact on young people's emotional health and well-being. *The Psychiatrist*, 37(5):167–170.

Maral Dadvar, Dolf Trieschnigg, and Franciska de Jong. 2014. Experts and Machines against Bullies: A Hybrid Approach to Detect Cyberbullies. In *Advances in Artificial Intelligence*, pages 275–281. Springer International Publishing.

Tom De Smedt and Walter Daelemans. 2012a. Pattern for Python. *Journal of Machine Learning Research*, 13:2063–2067.

Tom De Smedt and Walter Daelemans. 2012b. "Vreselijk mooi!" ("Terribly Beautiful!"): A Subjectivity Lexicon for Dutch Adjectives. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3568–3572, Istanbul, Turkey.

Francine Dehue, Catherine Bolman, and Trijntje Vollink. 2006. Cyberbullying: Youngster's Experiences and Parental Perception. *CyberPsychology*, 4(2):148–169.

Bart Desmet and Véronique Hoste. 2014. Recognising suicidal messages in Dutch social media. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 830–835, Reykjavik, Iceland.

Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying. *ACM Transactions on Interactive Intelligent Systems*, 2(3):18:1–18:30.

Julian J. Dooley and Donna Cross. 2010. Cyberbullying versus face-to-face bullying: A review of the similarities and differences. *Journal of Psychology*, 217:182–188.

Dorothy Wunmi Grigg. 2010. Cyber-Aggression: Definition and Concept of Cyberbullying. *Australian Journal of Guidance and Counselling*, 20:143–156, 12.

Sameer Hinduja and Justin W. Patchin. 2006. Bullies Move Beyond the Schoolyard: A Preliminary Look at Cyberbullying. *Youth Violence And Juvenile Justice*, 4(2):148–169.

Sameer Hinduja and Justin W. Patchin. 2012. Cyberbullying: Neither an epidemic nor a rarity. *European Journal of Developmental Psychology*, 9(5):539–543.

Valentin Jijkoun and Katja Hofmann. 2009. Generating a Non-English Subjectivity Lexicon: Relations That Matter. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–405, Stroudsburg, PA, USA.

Jaana Juvonen and Elisheva F. Gross. 2008. Extending the school grounds?-Bullying experiences in cyberspace. *Journal of School Health*, 78(9):496–505.

Sara Kiesler, Jane Sigel, and W.Timothy McGuire. 1984. Social psychological aspects of computer-mediated communication. *American Psychologist*, 39(10):1123–1134.

Qing Li. 2007. New Bottle but Old Wine: A Research of Cyberbullying in Schools. *Computers in Human Behavior*, 23(4):1777–1791.

Sonia Livingstone, Leslie Haddon, Anke Görzig, and Kjartan Ólafsson. 2010. Risks and safety on the internet: The perspective of European children. Initial Findings.

Vinita Nahar, Sanad Al-Maskari, Xue Li, and Chaoyi Pang. 2014. Semi-supervised Learning for Cyberbullying Detection in Social Networks. In *ADC.Databases Theory and Applications*, pages 160–171. Springer International Publishing.

Dan Olweus. 1996. Bullying at School: Knowledge Base and an Effective Intervention Program. *Annals of the New York Academy of Sciences*, 794:265–276.

Patrick B. O'Sullivan and Andrew J. Flanagin. 2003. Reconceptualizing 'flaming' and other problematic messages. *New Media & Society*, 5(1):69–94.

Megan Price and John Dalgleish. 2010. Cyberbullying: Experiences, Impacts and Coping Strategies as Described by Australian Young People. *Youth Studies Australia*, 29(2):51–59.

Kelly Reynolds, April Kontostathis, and Lynne Edwards. 2011. Using Machine Learning to Detect Cyberbullying. In *Proceedings of the 2011 10th International Conference on Machine Learning and Applications and Workshops*, ICMLA '11, pages 241–244, Washington, DC, USA. IEEE Computer Society.

Robert Slonje and Peter K. Smith. 2008. Cyberbullying: Another main type of bullying? *Scandinavian Journal of Psychology*, 49(2):147–154.

Peter K. Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. 2008. Cyberbullying: its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry*, 49(4):376–385.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, pages 102–107, Avignon, France.

Robert S. Tokunaga. 2010. Following You Home from School: A Critical Review and Synthesis of Research on Cyberbullying Victimization. *Computers in Human Behavior*, 26(3):277–287.

Katrien Van Cleemput, Sara Bastiaensens, Heidi Vandebosch, Karolien Poels, Gie Deboutte, Ann DeSmet, and Ilse De Bourdeaudhuij. 2013. Zes jaar onderzoek naar cyberpesten in Vlaanderen, België en daarbuiten: een overzicht van de bevindingen. (Six years of research on cyberbullying in Flanders, Belgium and beyond: an overview of the findings.) (White Paper). Technical report, University of Antwerp & Ghent University.

Marjan van de Kauter, Geert Coorman, Els Lefever, Bart Desmet, Lieve Macken, and Véronique Hoste. 2013. LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit. *Computational Linguistics in the Netherlands Journal*, 3:103–120.

Cynthia Van Hee, Ben Verhoeven, Els Lefever, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2015. Guidelines for the Fine-Grained Analysis of Cyberbullying, version 1.0. Technical Report LT3 15-01, LT3, Language and Translation Technology Team–Ghent University.

Kathleen Van Royen, Karolien Poels, Walter Daelemans, and Heidi Vandebosch. 2015. Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability. *Telematics and Informatics*, 32(1):89–97.

Heidi Vandebosch and Katrien Van Cleemput. 2009. Cyberbullying among youngsters: profiles of bullies and victims. *New Media & Society*, 11(8):1349–1371.

Heidi Vandebosch, Katrien Van Cleemput, Dimitri Mortelmans, and Michel Walrave. 2006. Cyberpesten bij jongeren in Vlaanderen: Een studie in opdracht van het viWTA (Cyberbullying among youngsters in Flanders: a study commissoned by the viWTA). Brussels: viWTA.

Veronika Šléglová and Alena Černá. 2011. Cyberbullying in Adolescent Victims: Perception and Coping. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 5(2).

Nancy E. Willard. 2007. *Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress*. Research Publishers LLC, 2nd edition.

Michele L. Ybarra and Kimberly J. Mitchell. 2008. How Risky are Social Networking Sites? A Comparison of Places Online Where Youth Sexual Soliciation and Harassment Occurs. *Paediatrics*, 121:350–357.

Dawei Yin, Brian D. Davison, Zhenzhen Xue, Liangjie Hong, April Kontostathis, and Lynne Edwards. 2009. Detection of Harassment on Web 2.0. In *Proceedings of the Content Analysis in the Web 2.0 (CAW2.0)*, Madrid, Spain.