

Automatic Detection and Prevention of Cyberbullying

Cynthia Van Hee*, Els Lefever*, Ben Verhoeven†, Julie Mennes*, Bart Desmet*,
Guy De Pauw†, Walter Daelemans† and Véronique Hoste*

*LT3 - Language and Translation Technology Team

Faculty of Arts and Philosophy, Ghent University, Belgium

Email: firstname.lastname@ugent.be

†CLiPS - Computational Linguistics Group

Faculty of Arts, University of Antwerp, Belgium

Email: firstname.lastname@uantwerpen.be

Abstract—The recent development of social media poses new challenges to the research community in analyzing online interactions between people. Social networking sites offer great opportunities for connecting with others, but also increase the vulnerability of young people to undesirable phenomena, such as cybervictimization. Recent research reports that on average, 20% to 40% of all teenagers have been victimized online. In this paper, we focus on cyberbullying as a particular form of cybervictimization. Successful prevention depends on the adequate detection of potentially harmful messages. However, given the massive information overload on the Web, there is a need for intelligent systems to identify potential risks automatically. We present the construction and annotation of a corpus of Dutch social media posts annotated with fine-grained cyberbullying-related text categories, such as insults and threats. Also, the specific participants (harasser, victim or bystander) in a cyberbullying conversation are identified to enhance the analysis of human interactions involving cyberbullying. Apart from describing our dataset construction and annotation, we present proof-of-concept experiments on the automatic identification of cyberbullying events and fine-grained cyberbullying categories.

Keywords—*Cyberbullying prevention; Text classification; Dataset construction.*

I. INTRODUCTION

The rise of Web 2.0 applications has substantially affected communication and relationships in today's society. Forums or message boards, blogs and social networking platforms like Facebook, Twitter, Tumblr or WhatsApp have become an important means of communication, especially among teenagers. Although most of the time, a child's Internet use is perfectly safe and enjoyable, there are risks involved in online communication through social media. Like offline communities, online communities can be harmful. Youngsters can be confronted with threatening situations, such as cyberbullying, suicidal behavior or grooming by paedophiles. As a response to those threats, a number of national and cross-national child protective initiatives (e.g., The Suicide Prevention Centre (<http://www.preventiezelfdoding.be/>), Child Focus (<http://www.childfocus.be/>)) have been starting projects over the last few years to increase online child safety. In spite of these efforts, much undesirable or even hurtful content remains online.

This research focuses on cyberbullying, one of the problems that emerged with the growing popularity of social media and its rapid adoption into our daily lives. Social media typically possess a number of features that make them a convenient way for cyberbullies to target their victims, including

anonymity, lack of supervision and impact [1]. Whereas traditional bullying was originally limited to school yards and youth movements, cyberbullying can continue at home. Cyberbullies can reach their victim through technological devices, such as mobile phones and laptops at any time of the day. Moreover, online content is exposed to a large audience and is difficult to remove. A message can be re-posted, liked or shared, which substantially increases the impact of an offensive or hurtful message, even if it was posted only once [2]. Over the past years, cyberbullying has become an important problem. A recent study among 2,000 Flemish secondary school students revealed that 11% of them had been bullied at least once in the six months preceding the survey [3]. The large-scale EU Kids Online Report [4] revealed that 17% of 9- to 16-year-olds had been bothered or upset by something online in the past year. Juvonen et al. [5] found that no less than 72% of 12- to 17-year-olds encountered cyberbullying at least once within the year preceding the questionnaire. Tokunaga [6] found that cybervictimization rates among teenagers vary between 20% and 40% on average [1], [7], [8], [9]. The figures vary depending on location, interval and the conceptualizations researchers use to describe cyberbullying. All of them demonstrate, however, that online platforms are increasingly used for bullying and that cyberbullying is thus not a rare problem. Moreover, it poses a significant threat to a teenager's mental and physical well-being with studies linking cyberbullying to depression, low self-esteem and school problems [10], [11], [12]. In extreme cases, its effects have even been linked to self-harm [10] and suicide [13]. Successful detection of cyberbullying is therefore of key importance to identify possibly threatening situations online and prevent them from escalating. Given the massive information overload on the Web, it has become unfeasible for humans to keep track of all conversations produced online. In order to manage this amount of information in an efficient way, there is an urgent need for intelligent techniques to signal harmful content automatically. This would allow for large-scale social media monitoring and early detection of harmful situations, such as cyberbullying, suicidality and sexually transgressive behavior (e.g., paedophilia). Recent research on the desirability of such detection systems found that a major part of the respondents favoured automatic monitoring on the condition that effective follow-up strategies are included and that privacy and autonomy are guaranteed [14].

Dadvar [15], Dinakar et al. [16] and Reynolds et al. [17] describe some of the first forays into the automatic detection of cyberbullying. To the best of our knowledge, however, we present the first study on recognizing cyberbullying events

in social media content by means of a fine-grained textual annotation of the corpus, in addition to implementing a binary distinction (cyberbullying versus non-cyberbullying).

The main objective of this research is to gain insight into the linguistic characteristics of cyberbullying by collecting and annotating an adequate dataset. This will allow us to explore text characteristics (or *features*) that are potentially useful in distinguishing between cyberbullying and non-cyberbullying content. For the annotation of the data, we consider fine-grained categories related to cyberbullying, such as insults and threats [18]. Such a fine-grained distinction provides insight into various types of cyberbullying and the degree to which they are alarming (e.g., expressions of a threat are considered more alarming than a single insult). Moreover, typical roles in a cyberbullying event are annotated (i.e., bully, victim, bystander). This way, cyberbullying incidents can be reconstructed through its participants, which may provide clearer insight into the severity of the incident. For instance, cyberbullying incidents where bystanders defend the victim or discourage the bully from continuing might not be as alarming as those where a victim stands alone and feels powerless when faced with a bully. Finally, we investigate the feasibility of automatically recognizing potentially offensive or harmful messages in Dutch user-generated content. Such an automatic system could serve as a first filter that reduces the amount of incoming messages for human moderators. Several users are targeted here: child protection agencies, social care organizations, such as the Suicide Prevention Centre, as well as parents and teachers.

The remainder of the paper is structured as follows: in Section II, a brief literature review of studies that have focused on cyberbullying detection is presented. Our experimental corpus is described in Section III, as well as the data collection and annotation. Section IV gives an overview of the experimental setup and results. Finally, we draw conclusions and formulate directions for future research in Section V.

II. RELATED RESEARCH

Cyberbullying has been a widely covered research topic over the past few years, especially in the realm of social sciences. Studies have focused on the conceptualization of cyberbullying and the occurrence of the phenomenon [19], [20], [21]. Additionally, different types of cyberbullying have been identified [22], [23], [24] and the consequences of cyberbullying have been investigated [9], [10], [25]. More recently, studies have focused on the use of NLP techniques for the detection and prevention of cyberbullying. Yin et al. [26] applied a supervised machine learning approach for the automatic detection of cyberharassment. They combined local tf-idf features with sentiment features and features capturing the similarity between several posts and obtained an F-score of 0.44. Dadvar [15] applied a hybrid approach combining supervised machine learning models with an expert system that incorporates knowledge from a sociological and psychological point of view (e.g., identifying characteristics of potential bullies on social networks) to recognize cyberbullying. They showed that combining user information and expert views with lexical features, yields fairly good results ($F = 0.64$). Reynolds et al. [17] applied rule-based learning to develop a model for detecting cyberbullying based on textual features (e.g., the

number of curse words in a message) and compared its performance to a bag-of-words model (i.e., based on a matrix of all the words that occur in the training corpus). They found that the rule-based method outperformed the bag-of-words model, achieving a recall of 78.5%. Dinakar et al. [27] conducted text classification experiments on a YouTube corpus. Using supervised machine learning and bag-of-words features, they built topic-sensitive classifiers to determine whether the topic of an insulting document is of a sensitive nature (i.e., sexuality, intelligence or race). In all of the aforementioned studies, however, cyberbullying detection is approached as a binary classification task (cyberbullying versus non-cyberbullying) without taking into account specific forms of cyberbullying such as threats, exclusions or insults. Moreover, these studies mainly focused on the detection of offensive posts written by a harasser, without specifying whether and how posts from victims and bystanders were considered. However, recent studies in the domain of automatic role assignment have emphasized the importance of community detection and role identification to enhance the analysis of online conversations [28].

The current research focuses on the detection of cyberbullying *events*, which include 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 events (i.e., cyberbullying posts irrespective of the author's role) and 2) the classification of more fine-grained categories related to cyberbullying, such as threats and insults.

III. DATASET CONSTRUCTION AND ANNOTATION

The availability of suitable data represents an important challenge in research on cyberbullying. However, a suitable dataset is needed for building representative models for cyberbullying detection. This section describes the construction of a Dutch corpus of social media messages containing both cyberbullying and non-cyberbullying content.

A. Data Collection

We constructed a corpus by collecting data from the social networking site Ask.fm (<http://ask.fm>), by receiving donations and by setting up simulation experiments with volunteer youngsters. In total, 91,370 Dutch posts were collected.

Ask.fm A substantial part of our corpus was collected from the social networking site Ask.fm where users can create profiles and ask questions and answer them, with the option of doing so anonymously. Typically, Ask.fm data consists of question-answer pairs published on a user's profile. The data was retrieved by crawling a number of seed sites using the GNU Wget software (<https://www.gnu.org/software/wget>). After filtering out non-Dutch content this resulted in 85,462 posts. As the posts containing cyberbullying were underrepresented in the corpus, we started two initiatives to complement the dataset:

Donations Firstly, we launched a media campaign in which people were asked to donate evidence of personal cases of cyberbullying. This resulted in a rather small but highly topical set of messages including Facebook hate pages, message board posts and chat conversations.

Simulations Secondly, a series of simulation experiments were set up in which volunteer teenagers were asked to participate in a cyberbullying simulation on a social network

by means of a role-playing game. A social networking platform was designed that is comparable to Facebook using SocialEngine (<http://www.socialengine.com>).

TABLE I. DATA DISTRIBUTION FOR THE FINE-GRAINED TEXT CATEGORIES RELATED TO CYBERBULLYING.

Category	Positive Instances	Harmfulness Score		
		0	1	2
Threat/blackmail	204	-	137	67
Insult	4,265	381	3,796	88
Curse/exclusion	1,111	-	1,009	102
Defamation	162	-	160	2
Sexual talk	495	398	4	93
Defense	2,226	-	2,087	139
Encouragements to the harasser	42	-	41	1

B. Data Annotation

In order to keep track of harmful user-generated content, we developed a fine-grained annotation scheme for the analysis of textual cyberbullying which is detailed in Van Hee et al. [18] and applied it to our corpus. To provide the annotators with some context, all posts were presented within their original conversation where possible. The annotation scheme describes two levels of annotation. First, the annotators were asked to indicate, at the post level, whether a post is part of a cyberbullying event. This was done by assigning a harmfulness score to the post 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. When a post was considered to be part of a cyberbullying context (i.e., it was given a harmfulness score of 1 or 2), the annotators indicated the author’s role in the cyberbullying event. In addition to victim and harasser, two types of bystanders are distinguished in our annotation scheme: 1) bystander-defenders, who help the victim and discourage the harasser from continuing his actions and 2) bystander-assistants, who do not initiate, but take part in the actions of the harasser. Secondly, at the subsentence level, the annotators were tasked with the identification of fine-grained text categories related to cyberbullying, even if the post was not considered harmful. For instance, in the sentence “Hey bitches, zin in een filmpje vanavond?” (*Hi bitches, anyone in for a movie tonight?*), *bitches* should be annotated as an insulting word. More concretely, they identified all text spans corresponding to one of the categories described in the annotation scheme. All annotations were done using the brat rapid annotation tool [29]. Table II presents the fine-grained cyberbullying categories and some example annotations of our dataset in brat.

In total, 85,462 Dutch posts were annotated by two annotators. To demonstrate the validity of our guidelines, inter-annotator agreement scores were calculated using Kappa [30] on a subset of the corpus. The Kappa score for the identification of cyberbullying events is 0.69. Kappa scores for the categories *Threat*, *Insult*, *Defense*, *Sexual Talk* and *Threat* range from moderate to substantial (i.e., from 0.52 to 0.66). They are low, however, for the categories *Defamation*, *Encouragements* and *Curse*, the identification of which seems to be rather difficult.

C. Experimental Corpus

For our preliminary experiments, we focused on the Ask.fm dataset. As shown in Table I, the experimental corpus features a heavily skewed class distribution with the large majority of posts not being part of any cyberbullying event. Regarding the occurrence of the fine-grained categories, we observe that insults are the most frequent type of cyberbullying activity in our corpus, followed by defense statements and curses/exclusions. Encouragements to the harasser is the least represented category. In this respect, it is worth mentioning that in case the annotators had too little context at their disposal to discern encouragements by bystanders from bullying acts by bullies, they annotated the post as a bullying act.

For each category, the number of instances marked with a harmfulness score of 0, 1 and 2 is given. As can be inferred from the table, 381 insults were identified in a non-cyberbullying context (e.g., insults as a ‘socially accepted’ way of addressing each other among friends). A major part of the category *Sexual talk* received a harmfulness score of zero, which means that these instances contained harmless sexual talk. Utterances considered sexual harassment were assigned a score of 1 or 2. If we consider the different roles in the annotated bullying events, we observe that the role of bully features in more than half of the annotated instances, followed by the victim role in about 30% of the instances. The bystander role in its two different subroles accounts for about 10% of the experimental corpus. These figures show that by focusing only on offensive posts (i.e., typical posts from a bully), as most studies on cyberbullying detection have done, about half of the relevant posts are ignored.

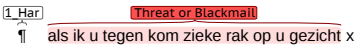
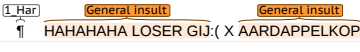
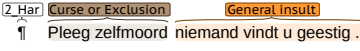
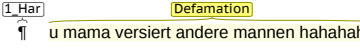
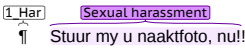
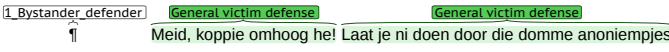

IV. EXPERIMENTS

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

A. Experimental setup

We explored the feasibility of automatic classification of cyberbullying events (i.e., a binary classifier was developed for distinguishing cyberbullying from non-cyberbullying posts) and more fine-grained text categories related to cyberbullying. To this end, binary classifiers were built for each of these categories (see Table I for an overview of the fine-grained categories). For our experiments, we used Support Vector Machines (SVM) as the classification algorithm, since they have been proven to work well for high-skew text classification tasks similar to the ones under investigation [31]. We used linear kernels and experimentally determined the optimal cost value c to be 1. All experiments were carried out using Pattern [32]. As preprocessing steps, we applied tokenization, PoS-tagging and lemmatization to the data using the LeT’s Preprocess Toolkit [33]. In supervised learning, a machine learning algorithm takes a set of training instances (of which the label is known) and seeks to build a model that generates a desired prediction for an unseen instance. To enable the model construction, all instances are represented as a vector of features (i.e., inherent characteristics of the data) that contain information that is potentially useful for distinguishing cyberbullying from non-cyberbullying content. For our experiments, we implemented two types of lexical features: bag-of-word features and polarity features based on

TABLE II. DEFINITIONS AND BRAT ANNOTATION EXAMPLES OF THE FINE-GRAINED TEXT CATEGORIES RELATED TO CYBERBULLYING.

Category	Brat annotation example	Translation
Threat/blackmail Expressions containing physical or psychological threats, or indications of blackmail.		<i>I'll smash you in the face when I see you x</i>
Insult Expressions containing abusive, degrading or offensive language that are meant to insult the addressee.		<i>HAHAHAHA YOU LOSER :(X POTATO HEAD</i>
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.		<i>Just commit suicide, nobody thinks you're funny...</i>
Defamation Expressions that reveal confident, embarrassing or defamatory information about the victim to a large public.		<i>Your mom is flirting with other men hahaha</i>
Sexual talk Expressions with a sexual meaning that are possibly harmful.		<i>Send me a naked picture of yourself, now!!</i>
Defense Expressions in support of the victim, expressed by the victim himself or by a bystander.		<i>Cheer up girl, don't let those stupid anons make you feel bad</i>
Encouragements to the harasser Expressions in support of the harasser.		<i>Indeed, she shouldn't be alive !!</i>

existing sentiment lexicons, resulting in a set of ~300.000 features in total. Bag-of-words features represent a corpus as an unordered set (or 'bag') of word or character sequences.

- **Word unigram and bigram bags-of-words:** binary features indicating the presence of word unigrams (i.e., a single word) and bigrams (i.e., a sequence of two words).
- **Character trigram bag-of-words:** binary features indicating the presence of character trigrams (without crossing word boundaries). A character-based bag-of-words representation is useful as it provides some abstraction from the word level and is more robust to variation in spelling or grammar.
- **Sentiment lexicon features:** polarity features that might be useful to provide insight into the polarity orientation of cyberbullying posts. To increase the lexicon coverage, lemmas were taken into account. The features are based on existing sentiment lexicons for Dutch [34], [35]:
 - The number of positive, negative and neutral lexicon words found in the text (averaged over text length).
 - The overall post polarity (i.e., the sum of the values of identified sentiment words, averaged over text length).

B. Results

This section presents the results of our preliminary experiments. Two classification tasks were carried out: cyberbullying

event detection and the classification of fine-grained classification text categories related to cyberbullying. Evaluation was done using 10-fold cross-validation. As the evaluation metric we used F-score, which is the weighted average of the classifier's precision (i.e., the fraction of retrieved instances that are relevant) and recall (i.e., the ratio of the number of relevant instances that are retrieved). For the classification of cyberbullying events, our classifier obtains an F-score of 55.39%. F-scores for the fine-grained classification of cyberbullying vary considerably. As shown in Figure 1, the *Insult* classifier yields an F-score of 56.32%, whereas the classification performance for the categories *Encouragement* and *Defamation* is significantly lower with F-scores of 0.12% and 7.41%, respectively. In addition to data scarcity (e.g., only 42 positive instances for the *Encouragement* category), the large discrepancies in performance are presumably due to the extent to which a category is lexicalized. For instance, insults are generally highly lexicalized, whereas threats are often expressed in an implicit way.

As shown in Figure 2, the identification of cyberbullying events performs better in terms of precision than recall. Generally, the fine-grained cyberbullying categories show a good balance between precision and recall. Our experiments show satisfactory preliminary results, especially for the classification of bully events and insults. The best classification performance is obtained for fine-grained categories that are explicitly lexicalized (e.g., insults, sexual talk, defensive statements). This intuitively makes sense as we made use of lexical features to represent the data. The figures also show a correlation between

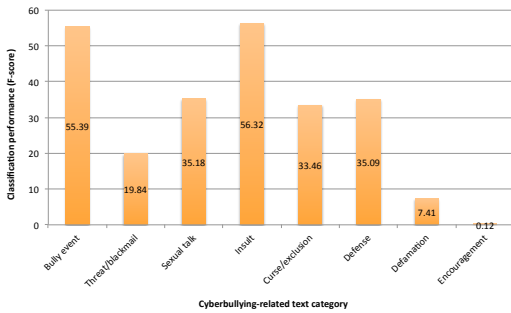


FIGURE 1. CLASSIFICATION RESULTS FOR THE IDENTIFICATION OF CYBERBULLYING EVENTS AND FINE-GRAINED CYBERBULLYING CATEGORIES, REPORTED AS 10-FOLD CROSS-VALIDATED F-SCORE ON THE POSITIVE CLASS (PERCENTAGES).

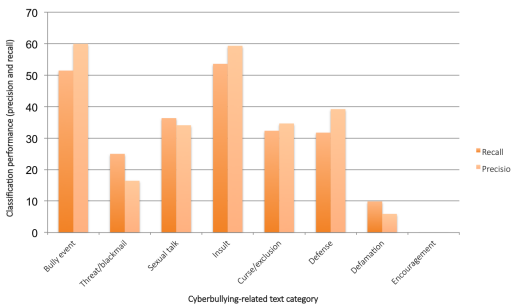


FIGURE 2. CLASSIFICATION RESULTS REPORTED BY MEANS OF PRECISION AND RECALL (PERCENTAGES).

the classification performance and the representation of the fine-grained category in our dataset. We therefore believe that the classification performance might benefit from extending the training corpus. The score obtained for the detection of cyberbullying events is 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. [17] worked with data that is similar to ours (i.e., question-answer pairs) and reported an accuracy of 78.5% when the positive posts were overrepresented in the training corpus. However, the classification accuracy was lower (53.82%) when the model was applied to the original corpus where the distribution of the positive posts was left unchanged.

V. CONCLUSIONS AND FUTURE WORK

Web 2.0 offers a multitude of ways to communicate with peers. Both positive and negative experiences are abundant on the Web and children and youngsters are vulnerable groups in harmful online communication. In this paper, we constructed a Dutch dataset of social media messages containing cyberbullying and proposed and evaluated a methodology for adequate annotation of this data. Additionally, we explored the feasibility of automatic cyberbullying detection. Our initial results show that cyberbullying detection is not a trivial task, especially not when focusing on more fine-grained categories.

As the ultimate goal of automatic cyberbullying detection is to reduce manual monitoring efforts on social media, recall optimization will be the prior focus for further research as we want to flag as many online threats as possible for the

moderator of a network. We will do a thorough qualitative analysis of the classification results to gain insight into the linguistic realization of cyberbullying and more specifically a series of fine-grained categories related to cyberbullying. We will also explore to what extent author role information can be used to enhance cyberbullying detection. A shallow error analysis revealed that implicit realizations of cyberbullying are fairly hard to recognize, as they are devoid of lexical cues such as 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. Social media texts tend to deviate from the linguistic norm, which reduces the effectiveness of both lexical and more complex features. Another direction for future work will therefore be orthographic normalization of the data as a preprocessing step [36]. Finally, we will investigate the integration of techniques such as cost-sensitive learning, data resampling or one-class learning to tackle the severe class imbalance.

ACKNOWLEDGMENT

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

- [1] S. Hinduja and J. W. Patchin, "Bullies Move Beyond the Schoolyard: A Preliminary Look at Cyberbullying," *Youth Violence And Juvenile Justice*, vol. 4, 2006, pp. 148–169.
- [2] J. J. Dooley and D. Cross, "Cyberbullying versus face-to-face bullying: A review of the similarities and differences," *Journal of Psychology*, vol. 217, 2010, pp. 182–188, ISSN: 0044-3409.
- [3] K. Van Cleemput, S. Bastiaensens, H. Vandebosch, K. Poels, G. Deboutte, A. DeSmets, and I. De Bourdeaudhuij, "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)," University of Antwerp & Ghent University, Tech. Rep., 2013.
- [4] "EU Kids Online: findings, methods, recommendations." 2014, URL: <http://eprints.lse.ac.uk/60512/> [accessed: 2015-07-30].
- [5] J. Juvonen and E. F. G., "Extending the school grounds?-Bullying experiences in cyberspace," *Journal of School Health*, vol. 78, 2008, pp. 496–505, ISSN: 1746-1561.
- [6] R. S. Tokunaga, "Following You Home from School: A Critical Review and Synthesis of Research on Cyberbullying Victimization," *Computers in Human Behavior*, vol. 26, 2010, pp. 277–287, ISSN: 0747-5632.
- [7] F. Dehue, C. Bolman, and T. Vollink, "Cyberbullying: Youngster's Experiences and Parental Perception," *CyberPsychology*, vol. 4, 2006, pp. 148–169.
- [8] Q. Li, "New Bottle but Old Wine: A Research of Cyberbullying in Schools," *Computers in Human Behavior*, vol. 23, 2007, pp. 1777–1791, ISSN: 0747-5632.
- [9] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett, "Cyberbullying: its nature and impact in secondary school pupils," *Journal of Child Psychology and Psychiatry*, vol. 49, 2008, pp. 376–385.
- [10] M. Price and J. Dalglish, "Cyberbullying: Experiences, Impacts and Coping Strategies as Described by Australian Young People," *Youth Studies Australia*, vol. 29, 2010, pp. 51–59, ISSN: 1038-2569.
- [11] V. Šléglová and A. Černá, "Cyberbullying in Adolescent Victims: Perception and Coping," *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, vol. 5, 2011, ISSN: 1802-7962. [Online]. Available: <http://cyberpsychology.eu/view.php?cisloclanku=2011121901&article=4>

- [12] H. Vandebosch, K. Van Cleemput, D. Mortelmans, and M. Walrave, "Cyberpesten bij jongeren in Vlaanderen: Een studie in opdracht van het viWTA (Cyberbullying among youngsters in Flanders: a study commissioned by the viWTA). Brussels: viWTA," 2006, URL: http://ist.vito.be/nl/publicaties/rapporten/rapport_cyberpesten.html [accessed: 2015-07-30].
- [13] S. Hinduja and J. W. Patchin, "Bullying, cyberbullying, and suicide," *Archives of Suicide Research*, vol. 14, 2010, pp. 206–221, ISSN: 1381-1118.
- [14] K. Van Royen, K. Poels, W. Daelemans, and H. Vandebosch, "Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability," *Telematics and Informatics*, vol. 32, 2015, pp. 89–97, ISSN: 0736-5853.
- [15] M. Dadvar, D. Trieschnigg, and F. de Jong, *Experts and Machines against Bullies: A Hybrid Approach to Detect Cyberbullies*. Springer International Publishing, Jan. 2014, pp. 275–281, in Sokolova, M. and van Beek, P., *Advances in Artificial Intelligence*, ISBN: 978-3-319-06482-6.
- [16] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying," *ACM Transactions on Interactive Intelligent Systems*, vol. 2, 2012, pp. 1–30, ISSN: 2160-6455.
- [17] K. Reynolds, A. Kontostathis, and L. Edwards, "Using Machine Learning to Detect Cyberbullying," in *Proceedings of the 10th International Conference on Machine Learning and Applications and Workshops December 18–21, 2011, Honolulu, Hawaii*. IEEE Computer Society, Dec. 2011, pp. 241–244, IEEE, ISBN: 978-0-7695-4607-0, URL: <http://dx.doi.org/10.1109/ICMLA.2011.152> [accessed: 2015-07-30].
- [18] C. Van Hee, B. Verhoeven, E. Lefever, G. De Pauw, W. Daelemans, and V. Hoste, "Guidelines for the Fine-Grained Analysis of Cyberbullying, version 1.0," LT3, Language and Translation Technology Team–Ghent University, Tech. Rep. LT3 15-01, 2015.
- [19] S. Hinduja and J. W. Patchin, "Cyberbullying: Neither an epidemic nor a rarity," *European Journal of Developmental Psychology*, vol. 9, 2012, pp. 539–543.
- [20] S. Livingstone, L. Haddon, A. Görzig, and K. Ólafsson, "Risks and safety on the internet: The perspective of European children. Full findings," 2011, URL: <http://eprints.lse.ac.uk/33731/> [accessed: 2015-07-30].
- [21] R. Slonje and P. K. Smith, "Cyberbullying: Another main type of bullying?" *Scandinavian Journal of Psychology*, vol. 49, 2008, pp. 147–154.
- [22] P. B. O'Sullivan and A. J. Flanagan, "Reconceptualizing 'flaming' and other problematic messages," *New Media & Society*, vol. 5, 2003, pp. 69–94, ISSN: 14614448.
- [23] H. Vandebosch and K. Van Cleemput, "Cyberbullying among youngsters: profiles of bullies and victims," *New Media & Society*, vol. 11, 2009, pp. 1349–1371.
- [24] N. E. Willard, Ed., *Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress*. Research Publishers LLC, 2007, ISBN: 978-087822-537-8.
- [25] H. Cowie, "Cyberbullying and its impact on young people's emotional health and well-being," *The Psychiatrist*, vol. 37, 2013, pp. 167–170, ISSN: 1758-3209.
- [26] D. Yin, B. D. Davison, Z. Xue, L. Hong, A. Kontostathis, and L. Edwards, "Detection of Harassment on Web 2.0," in *Proceedings of the Content Analysis in the Web 2.0 (CAW2.0) April 21, 2009, Madrid, Spain*. CAW 2.0, Apr. 2009, pp. 1231–1238, CAW 2.0, URL: <http://wbox0.cse.lehigh.edu/~brian/pubs/2009/CAW2/harassment.pdf> [accessed: 2015-07-25].
- [27] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the Detection of Textual Cyberbullying," in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, July 17–21, 2011, Barcelona, Spain*. AAAI, Jul. 2011, pp. 11–17, AAAI, ISBN: 978-1-57735-505-2, URL: <http://dblp.uni-trier.de/db/conf/icwsm/smw2011> [accessed: 2015-07-21].
- [28] A. McCallum, X. Wang, and A. Corrada-Emmanuel, "Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email," *Journal of Artificial Intelligence Research*, vol. 30, 2007, pp. 249–272, ISSN: 1076-9757.
- [29] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii, "BRAT: A Web-based Tool for NLP-assisted Text Annotation," in *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics April, 23–27, 2012, Avignon, France*. Association for Computational Linguistics, Apr. 2012, pp. 102–107, ACL, ISBN: 978-1-937284-19-0, URL: <http://dl.acm.org/citation.cfm?id=2380921.2380942> [accessed: 2015-07-24].
- [30] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, 1960, pp. 37–46.
- [31] B. Desmet and V. Hoste, "Recognising suicidal messages in Dutch social media," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC) May 26-31, 2014, Reykjavik, Iceland*. European Language Resources Association (ELRA), May 2014, pp. 830–835, ELRA, ISBN: 978-2-9517408-8-4, URL: <http://www.lrec-conf.org/proceedings/lrec2014/index.html> [accessed: 2015-07-26].
- [32] T. De Smedt and W. Daelemans, "Pattern for Python," *Journal of Machine Learning Research*, vol. 13, 2012, pp. 2063–2067, ISSN: 1532-4435.
- [33] M. van de Kauter, G. Coorman, E. Lefever, B. Desmet, L. Macken, and V. Hoste, "LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit," *Computational Linguistics in the Netherlands Journal*, vol. 3, 2013, pp. 103–120, ISSN: 2211-4009.
- [34] T. De Smedt and W. Daelemans, "'Vreselijk mooi!' ('Terribly Beautiful!'): A Subjectivity Lexicon for Dutch Adjectives," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC) May 23–25, 2012, Istanbul, Turkey*. European Language Resources Association (ELRA), May 2012, pp. 3568–3572, ELRA, ISBN: 978-2-9517408-7-7, URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/312_Paper.pdf [accessed: 2015-07-30].
- [35] V. Jijkoun and K. Hofmann, "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 (EACL) March 30–April 3, 2009, Athens, Greece*. Association for Computational Linguistics, Apr. 2009, pp. 398–405, ACL, URL: <http://www.aclweb.org/anthology/E09-1046> [accessed: 2015-07-27].
- [36] S. Schulz, G. De Pauw, O. De Clercq, B. Desmet, V. Hoste, W. Daelemans, and L. Macken, "Multi-Modular Text Normalization of Dutch User-Generated Content," *ACM Transactions on Intelligent Systems and Technology*, 2015 (in press).