Cyberbullying Detection in Social Media Using Machine Learning Algorithms

1 Introduction

In recent years, with the rise of social media usage, new concerns regarding users' mental and physical safety have been introduced.

Offensive language, generated by the crowd over various social platforms, might bully or hurt the feelings of an individual or a community. Recently, the research community has investigated and developed different semi- or supervised approaches and training datasets to detect or prevent offensive monologues or dialogues automatically [1][2].

In this project, I plan to conduct a detailed analytical study of the best-performing approaches, features, and embeddings, reported by the state-of-the-art, for automatic detection of cyberbullying in social media.

2 Dataset

Davidson et al. [1] has compiled a corpus containing around 24783 tweets annotated the text by crowd sourcing. This dataset represents three classes of labels as "hate speech", "offensive language" and "neither".

They begin with a hate speech lexicon containing words and phrases identified by internet users as hate speech, compiled by Hatebase.org. Using the Twitter API they searched for tweets containing terms from the lexicon, resulting in a sample of tweets from 33,458 Twitter users. They extracted the time-line for each user, resulting in a set of 85.4 million tweets. From this corpus they then took a random sample of 25k tweets containing terms from the lexicon and had them manually coded by CrowdFlower (CF) workers. CF Workers were asked to label each tweet as one of the three categories.

Each data file contains 5 columns:

1. **Count**: Number of CF users who coded each tweet (min is 3, sometimes more users coded a tweet when judgments were determined to be unreliable by CF).
2. **HateSpeech**: Number of CF users who judged the tweet to be hate speech.
3. **Offensive_language**: Number of CF users who judged the tweet to be offensive.
4. **Neither**: Number of CF users who judged the tweet to be neither offensive nor non-offensive.
5. **Class**: Class label for majority of CF users. 0 - hate speech 1 - offensive language 2 - neither

In this project, HateSpeech and Offensive_language will be categorized as cyberbullying.

2.1 Data Preparation

Data Preparation is the first step for training binary classifiers. The strategies for data preparation, which need to be carefully conducted, are described as below:

[1] https://github.com/t-davidson/hate-speech-and-offensive-language/tree/master/data

- **Basic cleaning methods**: We need to clean the data as (i) extracting the pure text from the dataset, removing duplicates, and NANs (ii) transforming to lowercase (iii) expanding the abbreviations.
- **Slangs**: Given the micro-blogging style of Twitter, using slangs are very common. Slangs bring difficulties to text mining approaches, especially for those emerging and thus do not have an updated entry in any dictionaries. So, I plan to transform the text into a canonical form using the reference dictionary[2] for slangs and abbreviations.
- **Removing methods**: Since hashtags, user references, links, and emojis are typical on social media platforms, preprocessing of the data is essential to normalize the text.

2.2 Tokenization

To start any text analysis, we need to break down the text into smaller parts like paragraphs and sentences and then convert words into tokens. One can create customized tokenizers on sentence-level or word-level and then pass them into embedding methods.

2.3 Feature Engineering

To transform tweets to numerical representations to make it plausible for the machine to deal with the features, I plan to use three words embedding, such as:

- **TF-IDF**: One way to represent words into vectors is to count the occurrence of words seen in the whole documents.
- **Word2Vec**: It first constructs a vocabulary from the data of the training text and then learns to describe words in vector form.
- **FastText**: FastText represents a low-dimensional vector text that is generated by summing vectors corresponding to the words in the text.

3 Evaluation Method

Eight binary classifiers, reported by the state-of-the-art as the best performing, will be used in this project: i) Naïve Bayes, ii) Decision Tree, iii) Logistic Regression, iv) Random Forest, v) Ada Boost, vi) SVMs, vii) Gradient Boosting, and viii) Multi-Layer Perceptron.
Also, Precision, Recall, F1-Score, and AUC-Score will be reporeted as the evaluation metrics to see which classifire with which embedding perform best.

References

[1] S. Malmasi and M. Zampieri, "Detecting hate speech in social media," *arXiv preprint arXiv:1712.06427*, 2017.

[2] C. Van Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. De Pauw, W. Daelemans, and V. Hoste, "Automatic detection of cyberbullying in social media text," *PloS one*, vol. 13, no. 10, p. e0203794, 2018.

[2] https://github.com/goncalopereira/twitter-moods

Detecting Hate Speech in Social Media

Shervin Malmasi
Harvard Medical School
Boston, MA, United States
smalmasi@bwh.harvard.edu

Marcos Zampieri
University of Wolverhampton
Wolverhampton, United Kingdom
marcos.zampieri@uni-koeln.de

Abstract

In this paper we examine methods to detect hate speech in social media, while distinguishing this from general profanity. We aim to establish lexical baselines for this task by applying supervised classification methods using a recently released dataset annotated for this purpose. As features, our system uses character n-grams, word n-grams and word skip-grams. We obtain results of 78% accuracy in identifying posts across three classes. Results demonstrate that the main challenge lies in discriminating profanity and hate speech from each other. A number of directions for future work are discussed.

1 Introduction

Research on safety and security in social media has grown substantially in the last decade. A particularly relevant aspect of this work is detecting and preventing the use of various forms of abusive language in blogs, micro-blogs, and social networks. A number of recent studies have been published on this issue such as the work by Xu et al. (2012) on identifying cyber-bullying, the detection of hate speech (Burnap and Williams, 2015) which was the topic of a recent survey (Schmidt and Wiegand, 2017), and the detection of racism (Tulkens et al., 2016) in user generated content.

The growing interest in this topic within the research community is evidenced by several related studies presented in Section 2 and by two recent workshops: Text Analytics for Cybersecurity and Online Safety (TA-COS)[1] held in 2016 at LREC and Abusive Language Workshop (AWL)[2] held in 2017 at ACL.

[1]http://www.ta-cos.org/home
[2]https://sites.google.com/site/abusivelanguageworkshop2017/

In this paper we address the problem of hate speech detection using a dataset which contains English tweets annotated with three labels: (1) hate speech (HATE); (2) offensive language but no hate speech (OFFENSIVE); and (3) no offensive content (OK). Most studies on abusive language so far (Burnap and Williams, 2015; Djuric et al., 2015; Nobata et al., 2016) have been modeled as binary classification with only one positive and one negative classes (*e.g.* hate speech vs non-hate speech). As noted by Dinakar et al. (2011), systems trained on such data often rely on the frequency of offensive or non-socially acceptable words to distinguish between the two classes. Dinakar et al. (2011) stress that in some cases "the lack of profanity or negativity [can] mislead the classifier".

Indeed, the presence of profane content does not in itself signify hate speech. General profanity is not necessarily targeted towards an individual and may be used for stylistic purposes or emphasis. On the other hand, hate speech may denigrate or threaten an individual or a group of people without the use of any profanities.

The main aim of this paper is to establish a lexical baseline for discriminating between hate speech and profanity on this standard dataset. The corpus used here provides us with an interesting opportunity to investigate how well a system can detect hate speech from other content that is generally profane. This baseline can be used to determine the difficulty of this task, and help highlight the most challenging aspects which must be addressed in future work.

The rest of this paper is organized as follows. In Section 2 we briefly outline some previous work on abusive language detection. The data is presented in Section 3, along with a description of our computational approach, features, and evaluation methodology. Results are presented in Section 4, followed by a conclusion and future perspectives in Section 5.

2 Related Work

There have been several studies on computational methods to detect abusive language published in the last few years. One example is the work by Xu et al. (2012) who apply sentiment analysis to detect bullying in tweets and use Latent Dirichlet Allocation (LDA) topic models (Blei et al., 2003) to identify relevant topics in these texts.

A number of studies have been published on hate speech detection. As previously mentioned, to the best of our knowledge all of them rely on binary classification (*e.g.* hate speech vs non-hate speech). Examples of such studies include the work by Kwok and Wang (2013), Djuric et al. (2015), Burnap and Williams (2015), and by Nobata et al. (2016).

Due to the availability of suitable corpora, the overwhelming majority of studies on abusive language, including ours, have used English data. However, more recently a few studies have investigated abusive language detection in other languages. Mubarak et al. (2017) addresses abusive language detection on Arabic social media and Su et al. (2017) presents a system to detect and rephrase profanity in Chinese. Hate speech and abusive language datasets have been recently annotated for German (Ross et al., 2016) and Slovene (Fišer et al., 2017) opening avenues for future work in languages other than English.

3 Methods

Next we present the Hate Speech Detection dataset used in our experiments. We applied a linear Support Vector Machine (SVM) classifier and used three groups of features extracted for these experiments: surface n-grams, word skip-grams, and Brown clusters. The classifier and features are described in more detail in Section 3.2 and Section 3.3 respectively. Finally, Section 3.4 discusses evaluation methods.

3.1 Data

In these experiments we use the aforementioned Hate Speech Detection dataset created by Davidson et al. (2017) and distributed via CrowdFlower.[3] The dataset features 14,509 English tweets annotated by a minimum of three annotators.

Individuals in charge of the annotation of this dataset were asked to annotate each tweet and categorize them into one of three classes:

1. (HATE): contains hate speech;
2. (OFFENSIVE): contains offensive language but no hate speech;
3. (OK): no offensive content at all.

Each instance in this dataset contains the text of a tweet[4] along with one of the three aforementioned labels. The distribution of the texts across the three classes is shown in Table 1.

Class	Texts
HATE	2,399
OFFENSIVE	4,836
OK	7,274
Total	14,509

Table 1: The distribution of classes and tweets in the Hate Speech Detection dataset.

All the texts are preprocessed to lowercase all tokens and to remove URLs and emojis.

3.2 Classifier

We use a linear SVM to perform multi-class classification in our experiments. We use the LIBLINEAR[5] package (Fan et al., 2008) which has been shown to be very efficient for similar text classification tasks. For example, the LIBLINEAR SVM implementation has been demonstrated to be a very effective classifier for Native Language Identification (Malmasi and Dras, 2015), temporal text classification (Zampieri et al., 2016a), and language variety identification (Zampieri et al., 2016b).

3.3 Features

We use two groups of surface features in our experiments as follows:

- Surface n-grams: These are our most basic features, consisting of character n-grams (of order 2–8) and word n-grams (of order 1–3). All tokens are lowercased before extraction of n-grams; character n-grams are extracted across word boundaries.

- Word Skip-grams: Similar to the above features, we also extract 1-, 2- and 3-skip word bigrams. These features are were chosen to approximate longer distance dependencies between words, which would be hard to capture using bigrams alone.

[3] https://data.world/crowdflower/hate-speech-identification

[4] Each tweet is limited to a maximum of 140 characters.

[5] http://www.csie.ntu.edu.tw/%7Ecjlin/liblinear/

3.4 Evaluation

To evaluate our methods we use 10-fold cross-validation. For creating the folds, we employ stratified cross-validation aiming to ensure that the proportion of classes within each partition is equal (Kohavi, 1995).

We report our results in terms of accuracy. The results obtained by our methods are compared against a majority class baseline and an oracle classifier.

The oracle takes the predictions by all the classifiers in Table 2 into account. It assigns the correct class label for an instance if at least one of the the classifiers produces the correct label for that instance. This approach establishes the *potential* or *theoretical* upper limit performance for a given dataset. Similar analysis using oracle classifiers have been previously applied to estimate the theoretical upper bound of shared tasks datasets in Native Language Identification (Malmasi et al., 2015) and similar language and language variety identification (Goutte et al., 2016).

4 Results

We start by investigating the efficacy of our features for this task. We fist train a single classifier, with each of them using a type of feature. Subsequently we also train a single model combining all of our features into single space. These are compared against the majority class baseline, as well as the oracle. The results of these experiments are listed in Table 2.

Feature	Accuracy (%)
Majority Class Baseline	50.1
Oracle	91.6
Character bigrams	73.6
Character trigrams	77.2
Character 4-grams	**78.0**
Character 5-grams	77.9
Character 6-grams	77.2
Character 7-grams	76.5
Character 8-grams	75.8
Word unigrams	77.5
Word bigrams	73.8
Word trigrams	67.4
1-skip Word bigrams	74.0
2-skip Word bigrams	73.8
3-skip Word bigrams	73.9
All features combined	77.5

Table 2: Classification results under 10-fold cross-validation.

The majority class baseline is quite high due to the class imbalance in the data. The oracle achieves an accuracy of 91.6% showing that none of our features are able to correctly classify a substantial portion of our samples.

We note that character n-grams perform well here, with 4-grams achieving the best performance of all features. Word unigrams also perform well, while performance degrades with bigrams, trigrams and skip-grams. However, the skip-grams may be capturing longer distance dependencies which provide complementary information to the other feature types. In tasks relying on stylistic information, it has been shown that skip-grams capture information that is very similar to syntactic dependencies (Malmasi and Cahill, 2015, §5).

Finally, the combination of all features does not achieve the performance of a character 4-grams model and causes a large dimensionality increase, with a total of 5.5 million features. It is not clear if this model is able to correctly capture the diverse information provided by the three feature types since we include more character n-gram models than word-based ones.

Next we analyze the rate of learning for these features. A learning curve for the classifier that yielded the best performance overall, character 4-grams, is shown in Figure 1.

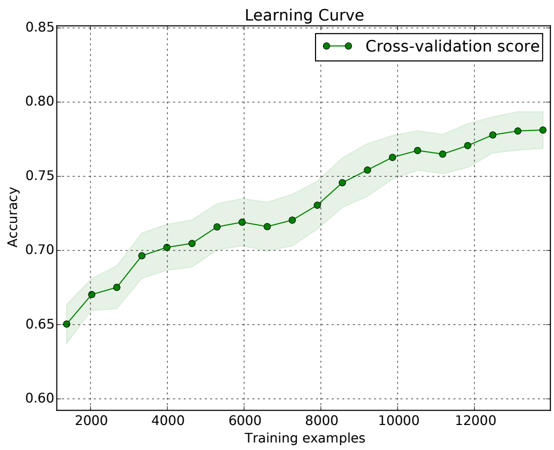

Figure 1: Learning curve for a character 4-gram model, with standard deviation highlighted. Accuracy does not plateau with the maximal data size.

We observe that accuracy increased continuously as the amount of training instances increased, and the standard deviation of the results between the cross-validation folds decreased. This suggests that the use of more training data is likely to provide even higher accuracy. It should be noted, however, that accuracy increases at a much slower rate after 15,000 training instances.

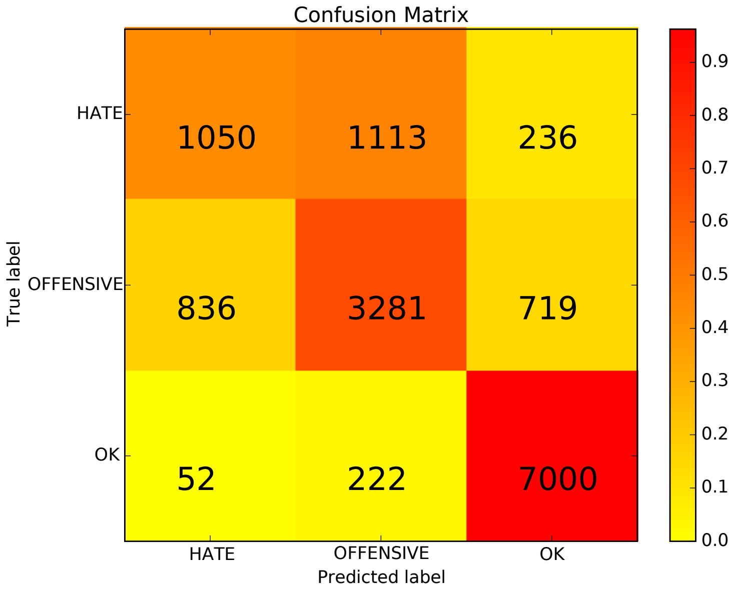

Figure 2: Confusion matrix of the character 4-gram model for our 3 classes. The heatmap represents the proportion of correctly classified examples in each class (this is normalized as the data distribution is imbalanced). The raw numbers are also reported within each cell. We note that the HATE class is the hardest to classify and is highly confused with the OFFENSIVE class.

Finally, we also examine a confusion matrix for the character 4-gram model, as shown in Figure 2. This demonstrates that the greatest degree of confusion lies between hate speech and generally offensive material, with hate speech more frequently being confused for offensive content. A substantial amount of offensive content is also misclassified as being non-offensive. The non-offensive class achieves the best result, with the vast majority of samples being correctly classified.

5 Conclusion

In this paper we applied text classification methods to distinguish between hate speech, profanity, and other texts. We applied standard lexical features and a linear SVM classifier to establish a baseline for this task. The best result was obtained by a character 4-gram model achieving 78% accuracy. The results presented in this paper showed that distinguishing profanity from hate speech is a very challenging task.

This was to the best of our knowledge one of the first experiments to detect hate speech on social media in a scenario including non-hate speech profanity. Previous work so far (e.g. Burnap and Williams (2015) and Djuric et al. (2015)) dealt with the distinction between hate speech and socially acceptable texts in a binary classification setting. In binary classification, Dinakar et al. (2011) note that the frequency of offensive words helps classifiers to distinguish between hate speech and socially acceptable texts.

We see a few directions in which this work could be expanded such as the use of more robust ensemble classifiers, a linguistic analysis of the most informative features, and error analysis of the misclassified instances. These aspects are presented in more detail in the next section.

5.1 Future Work

In future work we would like to investigate the performance of classifier ensembles and meta-learning for this task. Previous work has applied these techniques to a number of comparable text classification tasks, achieving success in competitive shared tasks. Examples of recent applications include automatic triage of posts in mental health forums (Malmasi et al., 2016b), detection of lexical complexity (Malmasi et al., 2016a), Native Language Identification (Malmasi and Dras, 2017), and dialect identification (Malmasi and Zampieri, 2017).

Another direction to pursue is the careful analysis of the most informative features for each class in this dataset. Our initial exploitation of the most informative words unigrams and bigrams suggests that coarse and obscene words are very informative for both HATE and OFFENSIVE words which confuses the classifiers. For HATE we observed a prominence of words targeting ethnic and social groups. Finally, an interesting outcome that should be investigated in more detail is that many of the most informative bigrams for the OK feature grammatical words. A more detailed analysis of these features could lead to more robust feature engineering methods.

An error analysis could also help us better understand the challenges in this task. This could be used to provide insights about the classifiers' performance as well as any underlying issues with the annotation of the Hate Speech Detection dataset which, as pointed out by Ross et al. (2016), is far from trivial. Figure 2 confirms that, as expected, most confusion occurs between HATE and OFFENSIVE texts. However, we also note that a substantial amount of offensive content is misclassified as being non-offensive. The aforementioned error analysis can provide insights about this.

Acknowledgments

We would like to thank the anonymous RANLP reviewers who provided us valuable feedback to increase the quality of this paper.

We further thank the developers and the annotators who worked on the Hate Speech Dataset for making this important resource available.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet* 7(2):223–242.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of ICWSM*.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *The Social Mobile Web*. pages 11–17.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, pages 29–30.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research* 9:1871–1874.

Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2017. Legal Framework, Dataset and Annotation Schema for Socially Unacceptable On-line Discourse Practices in Slovene. In *Proceedings of the Workshop Workshop on Abusive Language Online (ALW)*. Vancouver, Canada.

Cyril Goutte, Serge Léger, Shervin Malmasi, and Marcos Zampieri. 2016. Discriminating Similar Languages: Evaluations and Explorations. In *Proceedings of Language Resources and Evaluation (LREC)*. Portoroz, Slovenia.

Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*. volume 14, pages 1137–1145.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Shervin Malmasi and Aoife Cahill. 2015. Measuring Feature Diversity in Native Language Identification. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Denver, Colorado.

Shervin Malmasi and Mark Dras. 2015. Large-scale Native Language Identification with Cross-Corpus Evaluation. In *Proceedings of NAACL-HLT 2015*. Association for Computational Linguistics, Denver, Colorado.

Shervin Malmasi and Mark Dras. 2017. Native Language Identification using Stacked Generalization. *arXiv preprint arXiv:1703.06541* .

Shervin Malmasi, Mark Dras, and Marcos Zampieri. 2016a. Ltg at semeval-2016 task 11: Complex word identification with classifier ensembles. In *Proceedings of SemEval*.

Shervin Malmasi, Joel Tetreault, and Mark Dras. 2015. Oracle and Human Baselines for Native Language Identification. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Denver, Colorado.

Shervin Malmasi and Marcos Zampieri. 2017. German Dialect Identification in Interview Transcriptions. In *Proceedings of the Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Shervin Malmasi, Marcos Zampieri, and Mark Dras. 2016b. Predicting Post Severity in Mental Health Forums. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*.

Hamdy Mubarak, Darwish Kareem, and Magdy Walid. 2017. Abusive Language Detection on Arabic Social Media. In *Proceedings of the Workshop Workshop on Abusive Language Online (ALW)*. Vancouver, Canada.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pages 145–153.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proceedings of the Workshop on Natural Language Processing for Computer-Mediated Communication (NLP4CMC)*. Bochum, Germany.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection Using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Association for Computational Linguistics*. Valencia, Spain, pages 1–10.

Huei-Po Su, Chen-Jie Huang, Hao-Tsung Chang, and Chuan-Jie Lin. 2017. Rephrasing Profanity in Chinese Text. In *Proceedings of the Workshop Workshop on Abusive Language Online (ALW)*. Vancouver, Canada.

Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. A Dictionary-based Approach to Racism Detection in Dutch Social Media. In *Proceedings of the Workshop Text Analytics for Cybersecurity and Online Safety (TA-COS)*. Portoroz, Slovenia.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*. Association for Computational Linguistics, pages 656–666.

Marcos Zampieri, Shervin Malmasi, and Mark Dras. 2016a. Modeling language change in historical corpora: the case of Portuguese. In *Proceedings of Language Resources and Evaluation (LREC)*. Portoroz, Slovenia.

Marcos Zampieri, Shervin Malmasi, Octavia-Maria Sulea, and Liviu P Dinu. 2016b. A Computational Approach to the Study of Portuguese Newspapers Published in Macau. In *Proceedings of Workshop on Natural Language Processing Meets Journalism (NLPMJ)*. pages 47–51.

PLOS ONE

RESEARCH ARTICLE

Automatic detection of cyberbullying in soci media text

Cynthia Van Hee [1☯*], Gilles Jacobs [1☯], Chris Emmery [2], Bart Desmet [1], Els Lefever [1], Ben Verhoeven [2], Guy De Pauw [2], Walter Daelemans [2], Véronique Hoste [1]

1 Department of Translation, Interpreting and Communication - Faculty of Arts and Philosophy, Ghent University, Ghent, Belgium, 2 Department of Linguistics - Faculty of Arts, University of Antwerp, Antwerp, Belgium

☯These authors contributed equally to this work.
* cynthia.vanhee@ugent.be

Citation: Van Hee C, Jacobs G, Emmery C, Desmet B, Lefever E, Verhoeven B, et al. (2018) Automatic detection of cyberbullying in social media text. PLoS ONE 13(10): e0203794. https://doi.org/10.1371/journal.pone.0203794

Editor: Hussein Suleman, University of Cape Town, SOUTH AFRICA

Received: February 6, 2017

Accepted: August 28, 2018

Published: October 8, 2018

Data Availability Statement: Because the actual posts in our corpus could contain names or other identifying information, we cannot share them publicly in a repository. They can, however be obtained upon request, for academic purposes solely and via gillesm.jacobs@ugent.be or cynthia.vanhee@ugent.be. The replication data are available through the Open Science Framework repository https://osf.io/rgqw8/ with DOI 10.17605/OSF.IO/RGQW8. This replication dataset allows interested researchers to download 1) the feature vectors of the corpus underlying the experiments described in this paper, 2) the indices

Abstract

While social media offer great communication opportunities, they also increase the vulnerability of young people to threatening situations online. Recent studies report that cyberbullying constitutes a growing problem among youngsters. Successful prevention depends on the adequate detection of potentially harmful messages and the information overload on the Web requires intelligent systems to identify potential risks automatically. The focus of this paper is on automatic cyberbullying detection in social media text by modelling posts written by bullies, victims, and bystanders of online bullying. We describe the collection and fine-grained annotation of a cyberbullying corpus for English and Dutch and perform a series of binary classification experiments to determine the feasibility of automatic cyberbullying detection. We make use of linear support vector machines exploiting a rich feature set and investigate which information sources contribute the most for the task. Experiments on a hold-out test set reveal promising results for the detection of cyberbullying-related posts. After optimisation of the hyperparameters, the classifier yields an F_1score of 64% and 61% for English and Dutch respectively, and considerably outperforms baseline systems.

Introduction

Web 2.0 has had a substantial impact on communication and relationships in t Children and teenagers go online more frequently, at younger ages, and in mo (e.g. smartphones, laptops and tablets). Although most of teenagers' Internet u and the benefits of digital communication are evident, the freedom and anony enced online makes young people vulnerable with cyberbullying being one of t [1, 2].

Bullying is not a new phenomenon and cyberbullying has manifested itself a technologies have become primary communication tools. On the positive side, like blogs, social networking sites (e.g. Facebook), and instant messaging platf sApp) make it possible to communicate with anyone and at any time. Moreove place where people engage in social interaction, offering the possibility to esta corresponding to instances that were kept separately to test the experimental design (referred to as the "hold-out test set" in the paper), 3) a feature mapping dictionary that allows to trace all indices in the feature vector files back to the corresponding feature types (e.g. the feature indices 0 to 14,230 represent word 3-gram features). We also share the seed terms that were used to construct the corpora for our topic model features. Lastly, we provide an Excel spreadsheet presenting a results overview of all the tested systems. All of this information is made available for both the Dutch and English experiments.

Funding: The work presented in this paper was carried out in the framework of the AMiCA IWT SBO-project 120007 project to WD and VH, funded by the government Flanders Innovation & Entrepreneurship (VLAIO) agency; http://www.vlaio.be. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

relationships and maintain existing friendships [3, 4]. On the negative side how media increase the risk of children being confronted with threatening situation grooming or sexually transgressive behaviour, signals of depression and suicid cyberbullying. Users are reachable 24/7 and are often able to remain anonymo this makes social media a convenient way for bullies to target their victims out yard.

With regard to cyberbullying, a number of national and international initiativ launched over the past few years to increase children's online safety. Example (http://www.kivaprogram.net/), a Finnish cyberbullying prevention programme, *harcèlement*' campaign in France, Belgian governmental initiatives and helplin *be, veiligonline.be, mediawijs.be*) that provide information about online safety,

In spite of these efforts, a lot of undesirable and hurtful content remains onli a body of quantitative research on cyberbullying and observed cybervictimisat teenagers between 20% and 40%. [5] focused on 12 to 17 year olds living in th and found that no less than 72% of them had encountered cyberbullying at lea the year preceding the questionnaire. [6] surveyed 9 to 26 year olds in the Uni ada, the United Kingdom and Australia, and found that 29% of the respondents victimised online. A study among 2,000 Flemish secondary school students (ag revealed that 11% of them had been bullied online at least once in the six mon the survey [7]. Finally, the 2014 large-scale EU Kids Online Report [8] publishe 11 to 16 year olds had been exposed to hate messages online. In addition, you more likely to be exposed to cyberbullying as compared to 2010, which clearly that cyberbullying is a growing problem.

The prevalence of cybervictimisation depends on the conceptualisation used cyberbullying, but also on research variables such as location and the number the participants. Nevertheless, the above studies demonstrate that online platf ingly used for bullying, which is a cause for concern given its impact. As shown cyberbullying may negatively impact the victim's self-esteem, academic achiev tional well-being. [12] found that self-reported effects of cyberbullying include on school grades and feelings of sadness, anger, fear, and depression. In extre bullying could even lead to self-harm and suicidal thoughts.

These findings demonstrate that cyberbullying is a serious problem the cons which can be dramatic. Early detection of cyberbullying attempts is therefore o tance to youngsters' mental well-being. Successful detection depends on effec of online content, but the amount of information on the Web makes it practical for moderators to monitor all user-generated content manually. To tackle this p gent systems are required that process this information in a fast way and autor potential threats. This way, moderators can respond quickly and prevent threa from escalating. According to recent research, teenagers are generally in favou matic monitoring, provided that effective follow-up strategies are formulated, a and autonomy are guaranteed [13].

Parental control tools (e.g. *NetNanny*, https://www.netnanny.com/) already block uns or undesirable content and some social networks make use of keyword-based r tools (i.e. using lists of profane and insulting words to flag harmful content). Ho approaches typically fail to detect implicit and subtle forms of cyberbullying in explicit vocabulary is used. This creates the need for intelligent and self-learnin go beyond keyword spotting and hence improve the recall of cyberbullying det

The ultimate goal of this type of research is to develop models that could im monitoring for cyberbullying on social networks. We explore the automatic dete textual signals of cyberbullying, in which cyberbulying is approached as a comp non that can be realised in various ways (see the Annotation guidelines sectior overview). While the vast majority of the related research focuses on detecting 'attacks' (i.e. verbal aggression), the present study takes different types of cyb account, including more implicit posts from the bully, but also posts written by bystanders. This is a more inclusive conceptualisation for the task of cyberbully and should aid in moderation and prevention efforts by capturing different and signals of bullying.

To tackle this problem, we propose a machine learning method based on a li classifier [14, 15] exploiting a rich feature set. The contribution we make is two develop a complex classifier to detect *signals* of cyberbullying, which allows us ent types of cyberbullying that are related to different social roles involved in a event. Second, we demonstrate that the methodology is easily portable to othe vided there is annotated data available, by performing experiments on an Engl dataset.

The remainder of this paper is structured as follows: the next section presen of cyberbullying and its participant roles and provides an overview of the state cyberbullying detection. The *Data collection and annotation* section describes t struction and annotation. Next, we present the experimental setup and discuss tal results for English and Dutch. Finally, the *Conclusion and future research* se this paper and provides some perspectives for further research.

Related research

Both offline and online bullying are widely covered in the realm of social scienc ogy, and the increasing number of cyberbullying cases in recent years [16] has research efforts to detect cyberbullying automatically. In the following section, definition of cyberbullying and identify its participant roles and we provide a br automatic approaches to cyberbullying detection.

Cyberbullying definition and participant roles

A common starting point for conceptualising cyberbullying are definitions of tra *offline*) bullying, one of the most influential ones being formulated by [17]. The described bullying based on three main criteria, including i) intention (i.e. a bul inflict harm on the victim), ii) repetition (i.e. bullying acts take place repeatedly and iii) a power imbalance between the bully and the victim (i.e. a more power a less powerful victim). With respect to cyberbullying, a number of definitions a the above criteria. A popular definition is that of [18, p. 376], which describes c "an aggressive, intentional act carried out by a group or individual, using electr contact, repeatedly and over time, against a victim who cannot easily defend h However, opinion on the applicability of the above characteristics to cyberbully much divided [19], and besides theoretical objections, a number of practical lir been observed. Firstly, while [17] claims intention to be inherent to traditional much harder to ascertain in an online environment. Online conversations lack t face-to-face interaction like intonation, facial expressions and gestures, which r more ambiguous than real-life conversations. The receiver may therefore get th impression that they are being offended or ridiculed [20]. Another criterion for might not hold in online situations is the power imbalance between the bully ar This can be evident in real life (e.g. the bully is taller, stronger or older than the

is hard to conceptualise or measure online, where power may be related to tec
anonymity or the inability of the victim to escape from the bullying [19, 21]. Al
for the bully are inherent characteristics of the Web: once defamatory or confi
tion is made public through the Internet, it is hard to remove.

Finally, while arguing that repetition distinguishes bullying from single acts o
[17] himself states that such a single aggressive action can be considered bully
circumstances. Accordingly, [21] claim that repetition in cyberbullying is proble
ationalise, as it is unclear what the consequences are of a single derogatory m
page. A single act of aggression or humiliation may cause continued distress a
for the victim if it is shared or liked by a large audience [21]. [22, p. 26] compa
"snowball effect": one post may be repeated or distributed by other people so
out of the control of the initial bully and has larger effects than was originally i

Given these arguments, a number of less 'strict' definitions of cyberbullying
by among others [2, 5, 6], where a power imbalance and repetition are not dee
conditions for cyberbullying.

The above paragraphs demonstrate that defining cyberbullying is far from tr
ing prevalence rates (see the Introduction section) confirm that a univocal defi
phenomenon is still lacking in the literature [2]. Based on existing conceptualis
define cyberbullying as *content that is published online by an individual and th*
hurtful against a victim. Based on this definition, an annotation scheme was de
signal textual characteristics of cyberbullying, including posts from bullies, as w
from victims and bystanders.

Cyberbullying research also involves the identification of its participant roles
among the first to define the roles in a bullying situation. Based on surveys am
involved in real-life bullying situations, they defined six participant roles: victim
the target of repeated harassment), bullies (i.e. who are the initiative-taking pe
assistants of the bully (i.e. who encourage the bullying), reinforcers of the bully
force the bullying), defenders (i.e. who comfort the victim, take their side or try
lying) and outsiders (i.e. who ignore or distance themselves from the situation)
addition to the bully and victim, the researchers distinguish four bystanders (i.e
reinforcers, defenders and outsiders). [25], however, do not distinguish betwee
and assistants of the bullying. Their typology includes victims, bullies and three
bystanders: i) bystanders who participate in the bullying, ii) bystanders who he
the victim and iii) bystanders who ignore the bullying. The cyberbullying roles t
fied in our annotation scheme are based on existing bullying role typologies, gi
tional bullying roles are applicable to cyberbullying as well [26, 27]. More detai
different roles that we take into account are provided in the Data collection and
section.

Bystanders and -to a lesser extent- victims are often overlooked in the relate
result, these studies can be better characterised as verbal aggression detectio
retrieving bully attacks. By taking bystanders into account, we capture differen
tle signals of a bullying episode. Note that while in this work we did not include
of the participant roles as such, they are essential to the conceptualisation of t
tion task.

Detecting and preventing cyberbullying

As mentioned earlier, although research on cyberbullying detection is more lim
studies on the phenomenon, some important advances have been made in rec

what follows, we present a brief overview of the most important natural langua approaches to cyberbullying detection, but we refer to the survey paper by [28 detailed overview.

Although some studies have investigated the effectiveness of rule-based mo dominant approach to cyberbullying detection involves machine learning. Most learning approaches are based on supervised [30, 30–32] or semi-supervised le former involves the construction of a classifier based on labelled training data, supervised approaches rely on classifiers that are built from a training corpus c small set of labelled and a large set of unlabelled instances. Semi-supervised m used to handle data sparsity, a typical issue in cyberbullying research. As cybe tion essentially involves the distinction between bullying and non-bullying post is generally approached as a binary classification task where the positive class by instances containing (textual) cyberbullying, while the negative class is dev signals.

A key challenge in cyberbullying research is the availability of suitable data, sary to develop models that characterise cyberbullying. In recent years, only a have become publicly available for this particular task, such as the training set the context of the CAW 2.0 workshop (http://caw2.barcelonamedia.org), a MyS myspace.com) [34] and Formspring (http://www.formspring.me) cyberbullying tated with the help of Mechanical Turk [29], and more recently, the Twitter Bul dataset [35]. Many studies have therefore constructed their own corpus from s websites that are prone to bullying content, such as YouTube [30, 32], Twitter gram [38], MySpace [31, 34], FormSpring [29, 39], Kaggle [40] and ASKfm [41] bottleneck of data availability, cyberbullying detection approaches have been s implemented over the past years and the relevance of automatic text analysis ensure child safety online has been recognised [42].

Among the first studies on cyberbullying detection are [29–31], who explored tive power of *n*-grams (with and without tf-idf weighting), part-of-speech inform first and second pronouns), and sentiment information based on (polarity and p lexicons for this task. Similar features were not only exploited for coarse-graine ing detection, but also for the detection of more fine-grained cyberbullying cate Despite their apparent simplicity, content-based features (i.e. lexical, syntactic information) are very often exploited in recent approaches to cyberbullying det 43]. In fact, as observed by [28], more than 41 papers have approached cyberb tion using content-based features, which confirms that this type of information the task.

More and more, however, content-based features are combined with semant derived from topic model information [44], word embeddings and representatio [43, 45]. More recent studies have also demonstrated the added value of user- tion for the task, more specifically by including users' activities (i.e. the numbe social network, their age, gender, location, number of friends and followers, an 46, 47]. A final feature type that gains increasing popularity in cyberbullying de work-based features, whose application is motivated by the frequent use of soc for the task. By using network information, researchers aim to capture social re participants in a conversation (e.g. bully versus victim), and other relevant info the popularity of a person (i.e. which can indicate the power of a potential bully network, the number of (historical) interactions between two people, and so on instance used network-based features to take the behavioural history of a pote account. [49] detected cyberbullying in tweets and included network features i

Olweus' [17] bullying conditions (see supra). More specifically, they measured t
imbalance between a bully and victim, as well as the bully's popularity based o
graphs and the bully's position in the network.

As mentioned earlier, social media are a commonly used genre for this type
recently, researchers have investigated cyberbullying detection in multi-modal
specific platforms. For instance [38] explored cyberbullying detection using mu
extracted from the social network Instagram. More precisely, they combined te
derived from the posts themselves with user metadata and image features anc
integrating the latter enhanced the classification performance. [37] also detect
in different data genres, including ASKfm, Twitter, and Instagram. They took ro
into account by integrating bully and victim scores as features, based on the oc
bully-related keywords in their sent or received posts.

With respect to the datasets used in cyberbullying research, it can be observ
are often composed by keyword search (e.g. [43, 44]), which produces a biasec
tive (i.e. bullying) instances. To balance these corpora, negative data are often
background corpus or data resampling [50] techniques are adopted [33, 47]. Fo
data were randomly crawled across ASKfm and no keyword search was used to
data. Instead, all instances were manually annotated for the presence of bullyi
our corpus contains a realistic distribution of bullying instances.

When looking at the performance of automatic cyberbullying, we see that sc
greatly and do not only depend on the implemented algorithm and parameter s
also on a number of other variables. These include the metrics that are used to
tem (i.e. micro- or macro-averaged F_1, precision, recall, AUC, etc.), the corpus genre (i.
book, Twitter, ASKfm, Instagram) and class distribution (i.e. balanced or unbala
annotation method (i.e. automatic annotations or manual annotations using cro
or by experts) and, perhaps the most important distinguishing factor, the conce
cyberbullying that is used. More concretely, while some approaches identify se
[30] or insulting language [29], others propose a more comprehensive approac
different types of cyberbullying [41] or by modelling the bully-victim communic
involved in a cyberbullying incident [37].

The studies discussed in this section demonstrated the variety of approaches
used to tackle cyberbullying detection. However, most of them focused on cybe
'attacks', or posts written by a bully. Moreover, it is not entirely clear if differen
cyberbullying were taken into account (e.g. sexual intimidation or harassment,
cal threats), in addition to derogatory language or insults. In the present study,
is considered a complex phenomenon comprising different forms of harmful on
iour, which are described in more detail in our annotation scheme [23]. Purposi
manual monitoring efforts on social networks, we developed a system that auto
detects signals of cyberbullying, including attacks from bullies, as well as victim
reactions, the latter of which are generally overlooked in related research.

Most similar to this research is the work by [44], [43, 45], who investigated b
posted by different author roles (e.g. bully, victim, bystander, assistant, defend
accuser, reinforcer). However, they collected tweets using the keywords *bully,*
ing. As a result, their corpus contained many reports or testimonials of cyberbu
1), instead of actual cyberbullying. Moreover, their method implies that cyberb
that are devoid of such keywords are not included in the training corpus.

1. *"Some tweens got violent on the n train, the one boy got off after blows 2 t*
   him cryin as he walkd away: (bullying not cool" [44, *p. 658*]

What clearly distinguishes these works from the present is that their concept
cyberbullying is not explained. It is, in other words, not clear which type of post
ered bullying and which are not. In the present research, we identify different t
and all are included in the positive class of our experimental corpus.

For this research, English and Dutch social media data were annotated for fi
forms of cyberbullying, based on the actors involved in a cyberbullying inciden
nary experiments for Dutch [41, 51], we currently present an optimised cyberb
tion method for English and Dutch and hereby show that the proposed method
easily be applied to different languages, provided that annotated data are avai

Data collection and annotation

To be able to build representative models for cyberbullying, a suitable dataset
section describes the construction of two corpora, English and Dutch, containin
posts that are manually annotated for cyberbullying according to our fine-grain
scheme. This allows us to cover different forms and participants (or *roles*) invol
bullying event.

Data collection

Two corpora were constructed by collecting data from the social networking sit
where users can create profiles and ask or answer questions, with the option o
ymously. ASKfm data typically consists of question-answer pairs published on a
The data were retrieved by crawling a number of seed profiles using the GNU W
(http://www.gnu.org/software/wget/) in April and October, 2013. After language
(i.e. non-English or non-Dutch content was removed), the experimental corpora
113,698 and 78,387 posts for English and Dutch, respectively.

Data annotation

Cyberbullying has been a widely covered research topic recently and studies h
direct and indirect types of cyberbullying, implicit and explicit forms, verbal an
cyberbullying, and so on. This is important from a sociolinguistic point of view,
what cyberbullying involves is also crucial to build models for automatic cyberb
tion. In the following paragraphs, we present our data annotation guidelines [2
different types and roles related to the phenomenon.

Types of cyberbullying

Cyberbullying research is mainly centered around the conceptualisation, occur
vention of the phenomenon [1, 52, 53]. Sociolinguistic studies have identified d
of cyberbullying [12, 54, 55] and compared these types with forms of tradition
bullying [20]. Like traditional bullying, direct and indirect forms of cyberbullyin
identified. Direct cyberbullying refers to actions in which the victim is directly i
sending a virus-infected file, excluding someone from an online group, insulting
ing), whereas indirect cyberbullying can take place without awareness of the vi
or publishing confidential information, spreading gossip, creating a hate page o
working sites) [20].

The present annotation scheme describes some specific textual categories r
bullying, including threats, insults, defensive statements from a victim, encoura
harasser, etc. (see the Data collection and annotation section for a complete o

these forms were inspired by social studies on cyberbullying [7, 20] and manua
cyberbullying examples.

Roles in cyberbullying

Similarly to traditional bullying, cyberbullying involves a number of participants
well-defined roles. Researchers have identified several roles in (cyber)bullying
Although traditional studies on bullying have mainly concentrated on bullies ar
the importance of bystanders in a bullying episode has been acknowledged [56
can support the victim and mitigate the negative effects caused by the bullying
on social networking sites, where they hold higher intentions to help the victim
conversations [58]. [25] distinguish three main types of bystanders: i) bystande
pate in the bullying, ii) who help or support the victim and iii) those who ignore
Given that passive bystanders are hard to recognise in online text, only the for
included in our annotation scheme.

Annotation guidelines

To operationalise the task of automatic cyberbullying detection, we elaborated
tation scheme for cyberbullying that is strongly embedded in the literature and
our corpora. The applicability of the scheme was iteratively tested. Our final gu
fine-grained annotation of cyberbullying are described in a technical report [23
of the scheme was to indicate several types of textual cyberbullying and verba
severity, and the author participant roles. The scheme is formulated to be gene
limited to a specific social media platform. All messages were annotated in con
sented within their original content or conversation event) when available.

Essentially, the annotation scheme describes two levels of annotation. Firstly
were asked to indicate, at the message or post level, whether the text under in
related to cyberbullying. If the message was considered harmful and thus conta
tions of cyberbullying, annotators identified the author's participant role. Based
ture on role-allocation in cyberbullying episodes [25, 59], four roles are distingu
annotation scheme, including victim, bully, and two types of bystanders.

1. Harasser or bully: person who initiates the bullying.

2. Victim: person who is harassed.

3. Bystander-defender: person who helps the victim and discourages the haras
tinuing his actions.

4. Bystander-assistant: person who does not initiate, but helps or encourages t

Secondly, at the sub-sentence level, the annotators were tasked with the ide
fine-grained text categories related to cyberbullying. In the literature, different
bullying are identified [12, 54, 55] and compared with traditional bullying [20].
forms, the annotation scheme describes a number of textual categories that ar
to a cyberbullying event, such as threats, insults, defensive statements from a
agements to the harasser, etc. Most of the categories are related to direct form
(as defined by [25]), while one is related to *outing* [25], an indirect form of cyb
namely *defamation*. Additionally, a number of subcategories were defined to m
tion scheme as concrete and distinctive as possible (e.g., *discrimination* as a su
insult). All cyberbullying-related categories in the scheme are listed below, and
post for each category is presented in Table 1.

Table 1. Definitions and brat annotation examples of more fine-grained text categories related to c

Annotation category	Annotation example
Threat/blackmail	[I am going to find out who you are & I swear you are going to regret it.] THREAT
Insult	[Kill yourself] CURS [you fucking mc slut!!!!] GEN. INSULT [NO ONE LIKES YOU!!!!!] GEN. INSULT [You are an ugly useless little whore!!!!] GEN. INSULT
Curse/Exclusion	[Fuck you.] GEN. INSULT [Now shush I don’t wanna hear anything.] CURSE OR EXCLUSION
Defamation	[She slept with her ex behind his girlfiends back and she and him had broken up.] DEFAMATION
Sexual Talk	[Naked pic of you now.] SEXUAL HARASSMENT
Defense	[I would appreciate if you dindn’t talk shit about my bestfriend.] GEN. VICTIM DEFENSE [He has enough to deal with already.
Encour. to har.	[She is a massive slut] GEN. INSULT [i agree with you @user she is!] ENCOUR. HARASSER [LOL AT HER mate, im on your side] ENCOUR. HARASSER

https://doi.org/10.1371/journal.pone.0203794.t001

- Threat/blackmail: expressions containing physical or psychological threats or blackmail.
- Insult: expressions meant to hurt or offend the victim.
  - General insult: general expressions containing abusive, degrading or offens that are meant to insult the addressee.
  - Attacking relatives: insulting expressions towards relatives or friends of the
  - Discrimination: expressions of unjust or prejudicial treatment of the victim. discrimination are distinguished (i.e. sexism and racism). Other forms of dis should be categorised as general insults.
- Curse/exclusion: expressions of a wish that some form of adversity or misfort the victim and expressions that exclude the victim from a conversation or a s
- Defamation: expressions that reveal confident or defamatory information abo to a large public.
- Sexual Talk: expressions with a sexual meaning or connotation. A distinction between innocent sexual talk and sexual harassment.
- Defense: expressions in support of the victim, expressed by the victim himse bystander.
  - Bystander defense: expressions by which a bystander shows support for the courages the harasser from continuing his actions.
  - Victim defense: assertive or powerless reactions from the victim.
- Encouragement to the harasser: expressions in support of the harasser.
- Other: expressions that contain any other form of cyberbullying-related beha ones described here.

It is important to note that the categories were always indicated in text, even in which they occurred was not considered harmful, for instance in the post “hi for a movie?”, “bitches” was annotated as an insult while the post itself was no cyberbullying.

To provide the annotators with some context, all posts were presented withi
conversation when possible. All annotations were done using the brat rapid an
[60], some examples of which are presented in Table 1.

As can be deduced from the examples in the table, there were no restriction
form the annotations could take. They could be adjectives, noun phrases, verb
on. The only condition was that the annotation could not span more than one s
less than one word. Posts that were (primarily) written in another language tha
language (i.e. Dutch and English) were marked as such and required no furthe

We examined the validity of our guidelines and the annotations with an inte
agreement experiment that is described in the following section.

Annotation statistics

The English and Dutch corpora were independently annotated for cyberbullyin
linguists. All were Dutch native speakers and English second-language speake
strate the validity of our guidelines, inter-annotator agreement scores were ca
Kappa on a subset of each corpus. Inter-rater agreement for Dutch (2 raters) is
using Cohen's Kappa [61]. Fleiss' Kappa [62] is used for the English corpus (> 2
scores for the identification of cyberbullying are $\kappa = 0.69$ (Dutch) and $\kappa = 0.59$

As shown in Table 2, inter-annotator agreement for the identification of the r
grained categories for English varies from fair to substantial [63], except for *de*
appears to be more difficult to recognise. No encouragements to the harasser w
this subset of the corpus. For Dutch, the inter-annotator agreement is fair to su
for *curse* and *defamation*. Analysis revealed that one of both annotators often a
ter as an insult, and in some cases even did not consider it as cyberbullying-rel

In short, the inter-rater reliability study shows that the annotation of cyberbu
trivial and that more fine-grained categories like *defamation*, *curse* and *encour*
sometimes hard to recognise. It appears that defamations were sometimes har
from insults, whereas curses and exclusions were sometimes considered insult
The analysis further reveals that encouragements to the harasser are subject t
Some are straightforward (e.g. "I agree we should send her hate"), whereas ot
to the annotator's judgment and interpretation (e.g. "hahaha", "LOL").

Experimental setup

In this paper, we explore the feasibility of automatically recognising signals of
crucial difference with related research is that we do not only model bully 'atta
more implicit forms of cyberbullying and reactions from victims and bystander
one binary label 'signals of cyberbullying'), since these could likewise indicate t
ing is going on. The experiments described in this paper focus on the automati
such cyberbullying signals that need to be further investigated by human mod
applied in a real-life moderation loop.

The English and Dutch corpus contain 113,698 and 78,387 posts, respective
Table 3, the experimental corpus features a heavily imbalanced class distributi

Table 2. Inter-annotator agreement on the fine-grained categories related to cyberbullying.

	Threat	Insult	Defense	Sexual talk	Curse/exclusion	Defamation	Encouragements to the harasser
English	0.65	0.63	0.45	0.38	0.58	0.15	N/A
Dutch	0.52	0.66	0.63	0.53	0.19	0.00	0.21

https://doi.org/10.1371/journal.pone.0203794.t002

Table 3. Statistics of the English and Dutch cyberbullying corpus.

	Corpus size	Number(ratio) of bullying posts
English	113,698	5,375(4.73%)
Dutch	78,387	5,106(6.97%)

https://doi.org/10.1371/journal.pone.0203794.t003

large majority of posts not being part of cyberbullying. In classification, this cla
can lead to decreased performance. We apply cost-sensitive SVM as a possible
in optimisation to counter this. The cost-sensitive SVM reweighs the penalty pa
the error term by the inverse class-ratio. This means that misclassifications of t
positive class are penalised more than classification errors on the majority neg
Other pre-processing methods to handle data imbalance in classification includ
ing metrics and data resampling [64]. These methods were omitted as they we
too computationally expensive given our high-dimensional dataset.

For the automatic detection of cyberbullying, we performed binary classificat
ments using a linear kernel support vector machine (SVM) implemented in LIBL
by making use of Scikit-learn [66], a machine learning library for Python. The m
behind this is twofold: i) support vector machines (SVMs) have proven to work
similar to the ones under investigation [67] and ii) LIBLINEAR allows fast trainin
scale data which allow for a linear mapping (which was confirmed after a series
experiments using LIBSVM with linear, RBF and polynomial kernels).

The classifier was optimised for feature type (see the Pre-processing and fea
section) and hyperparameter combinations (see Table 4). Model selection was
10-fold cross validation in grid search over all possible feature types (i.e. group
tures, like different orders of *n*-gram bag-of-words features) and hyperparamet
tions. The best performing hyperparameters are selected by the F_1 score on the positive class. Th
winning model is then retrained on all held-in data and subsequently tested on
set to assess whether the classifier is over- or under-fitting. The hold-out set re
dom sample (10%) of all data. The folds were randomly stratified splits over th
distribution. Testing all feature type combinations is a rudimentary form of feat
and provides insight into which types of features work best for this particular ta

Feature selection over all individual features was not performed because of t
space (NL: 795,072 and EN: 871,296 individual features). [68], among other res
onstrated the importance of joint optimisation, where feature selection and hyp
optimisation are performed simultaneously, since the techniques mutually influ
other.

The optimised models are evaluated against two baseline systems: i) an uno
ear-kernel SVM (configured with default parameter settings) based on word *n*-g
and, ii) a keyword-based system that marks posts as positive for cyberbullying
word from existing vocabulary lists composed by aggressive language and prof

Table 4. Hyperparameters in grid-search model selection.

Hyperparameter	Values
Penalty of error term *C*	$1e^{\{-3,-2,..,2,3\}}$
Loss function	Hinge, squared hinge
Penalty: norm used in penalisation	'l1' ('least absolute deviations') or 'l2' ('least squares')
Class weight (sets penalty *C* of class *i* to weight*C*)	None or 'balanced', i.e. weight inversely proportional to cla frequencies

https://doi.org/10.1371/journal.pone.0203794.t004

Pre-processing and feature engineering

As pre-processing, we applied tokenisation, PoS-tagging and lemmatisation to t
the LeTs Preprocess Toolkit [69]. In supervised learning, a machine learning alg
set of training instances (of which the label is known) and seeks to build a mod
a desired prediction for an unseen instance. To enable the model construction,
are represented as a vector of features (i.e. inherent characteristics of the data
information that is potentially useful to distinguish cyberbullying from non-cybe
content.

We experimentally tested whether cyberbullying events can be recognised a
lexical markers in a post. To this end, all posts were represented by a number o
sources (or *features*) including lexical features like bags-of-words, sentiment le
and topic model features, which are described in more detail below. Prior to fea
some data cleaning steps were executed, such as the replacement of hyperlink
removal of superfluous white spaces, and the replacement of abbreviations by
(based on an existing mapping dictionary: http://www.chatslang.com/terms/abb
Additionally, tokenisation was applied before *n*-gram extraction and sentiment
ing, and stemming was applied prior to extracting topic model features.

After pre-processing of the corpus, the following feature types were extracte

- Word ***n***-gram bag-of-words: binary features indicating the presence of word u
  bigrams and trigrams.
- Character ***n***-gram bag-of-words: binary features indicating the presence of ch
  bigrams, trigrams and fourgrams (without crossing word boundaries). Charac
  provide some abstraction from the word level and provide robustness to the s
  tion that characterises social media data.
- Term lists: one binary feature derived for each one out of six lists, indicating t
  an item from the list in a post:
  - proper names: a gazetteer of named entities collected from several resourc
  - 'allness' indicators (e.g. "always", "everybody"): forms which indicate rheto
    ity [70] which can be helpful in identifying the often hyperbolic bullying lang
  - diminishers (e.g. "slightly", "relatively"): diminishers, intensifiers and negat
    all obtained from an English grammar describing these lexical classes and e
    ment lexicons (see further).
  - intensifiers (e.g. "absolutely", "amazingly")
  - negation words
  - aggressive language and profanity words: for English, we used the Google P
    (https://code.google.com/archive/p/badwordslist/downloads). For Dutch, a p
    ity lexicon was consulted (http://scheldwoorden.goedbegin.nl).
    Person alternation is a binary feature indicating whether the combination of
    second person pronoun occurs in order to capture interpersonal intent.
- Subjectivity lexicon features: positive and negative opinion word ratios, as we
  all post polarity were calculated using existing sentiment lexicons. For Dutch,
  of the Duoman [71] and Pattern [72] lexicons. For English, we included the Liu
  opinion lexicon [73], the MPQA lexicon [74], the General Inquirer Sentiment L

AFINN [76], and MSOL [77]. For both languages, we included the relative freq
68 psychometric categories in the Linguistic Inquiry and Word Count (LIWC) d
for English [78] and Dutch [79].

- Topic model features: by making use of the Gensim topic modelling library [8
LDA [81] and LSI [82] topic models with varying granularity (k = 20, 50, 100 a
trained on data corresponding to each fine-grained category of a cyberbullyin
threats, defamations, insults, defenses). The topic models were based on a ba
pus (EN: ± 1,200,000 tokens, NL: ± 1,400,000 tokens) scraped with the BootC
corpus toolkit. BootCaT collected ASKfm user profiles using lists of manually c
seed words that are characteristic of the cyberbullying categories.

When applied to the training data, this resulted in 871,296 and 795,072 feature
and Dutch, respectively.

Results

In this section, we present the results of our experiments to automatically dete
signals in an English and Dutch corpus of ASKfm posts. Ten-fold cross-validatio
formed in exhaustive grid search over different feature type and hyperparamet
(see the Experimental setup section). The *unoptimised word n-gram-based* clas
word-matching system serve as baselines for comparison. Precision, Recall and F_1 performanc
metrics were calculated on the positive class. We also report Area Under the Re
Curve (AUROC) scores, a performance metric that is more robust to data imbal
cision, recall and F score [84].

Table 5 gives us an indication of which feature type combinations score best
tribute most to this task. It presents the cross-validation and hold-out scores of

Table 5. Cross-validated and hold-out scores (%) according to different metrics (F_1, precision, recall, accuracy and area under the curve) for the English a
three best and worst combined feature type systems.

	Feature combination	Cross-validation scores					Hold-out scores				
		F_1	P	R	Acc	AUROC	F_1	P	R	Acc	AUROC
English											
Best three	B + C + D + E	64.26	73.32	57.19	96.97	78.07	63.69	74.13	55.82	97.21	77.47
	A + B + C	64.24	73.22	57.23	96.96	78.09	64.32	74.08	56.83	97.24	77.96
	A + C + E	63.84	73.21	56.59	96.94	77.78	62.94	72.82	55.42	97.14	77.24
Worst three	D	40.48	38.98	42.12	94.10	69.41	39.56	39.56	39.56	94.71	68.39
	A + D + E	38.95	31.47	51.10	92.37	72.76	40.71	33.87	51.00	93.49	73.22
	E	17.35	9.73	79.91	63.72	71.41	15.70	8.72	78.51	63.07	70.44
Baseline	word *n*-gram	58.17	67.55	51.07	96.54	74.93	59.63	69.57	52.17	96.57	75.50
	profanity	17.17	9.61	80.14	63.73	71.53	17.61	9.90	78.51	63.79	71.34
Dutch											
Best three	A + B + C + E	61.20	56.76	66.40	94.47	81.42	58.13	54.03	62.90	94.58	79.75
	A + B + C + D + E	61.03	71.55	53.20	95.53	75.86	58.72	67.40	52.03	95.62	75.21
	A + C + E	60.82	71.66	52.84	95.53	75.68	58.15	67.71	50.96	95.61	74.71
Worst three	D + B	32.90	29.23	37.63	89.91	65.61	30.16	34.72	26.65	92.61	61.73
	D	28.65	19.36	55.10	81.97	69.48	25.13	16.73	50.53	81.99	67.26
	B	24.74	21.24	29.61	88.16	60.94	17.99	23.15	14.71	91.98	55.80
Baseline	word *n*-gram	50.39	67.80	40.09	94.81	69.38	49.54	64.29	40.30	95.09	69.44
	profanity	28.46	19.24	54.66	81.99	69.28	25.13	16.73	50.53	81.99	67.26

https://doi.org/10.1371/journal.pone.0203794.t005

Table 6. Feature group mapping (Table 5).

A	word *n*-grams
B	subjectivity lexicons
C	character *n*-grams
D	term lists
E	topic models

https://doi.org/10.1371/journal.pone.0203794.t006

combinations, which are explained in the feature groups legend (Table 6). A to type combinations, each with 28 different hyperparameter sets have been test the results for the three best scoring systems by included feature types with op parameters. The maximum obtained F_1 score in cross-validation is 64.26% for English a 61.20% for Dutch and shows that the classifier benefits from a variety of featur results on the hold-out test set show that the trained systems generalise well o indicating little under- or overfitting. The simple keyword-matching baseline sy lowest performance for both languages even though it obtains high recall for b especially for English (80.14%), suggesting that profane language characterise lying-related posts. Feature group and hyperparameter optimisation provides a performance increase over the unoptimised word *n*-gram baseline system. The systems for each language do not differ a lot in performance, except the best s which trades recall for precision when compared to the runner-ups.

Table 7 presents the scores of the (hyperparameter-optimised) single feature to gain insight into the performance of these feature types when used individua the combined and single feature type sets reveals that word ***n***-grams, characte subjectivity lexicons prove to be strong features for this task. In effect, adding grams always improved classification performance for both languages. They ar vide robustness to lexical variation in social media text, as compared to word *n* subjectivity lexicons appear to be discriminative features, term lists perform ba own as well as in combinations for both languages. This shows once again (see baseline) that cyberbullying detection requires more sophisticated information profanity lists. Topic models seem to do badly for both languages on their own,

Table 7. Cross-validated and hold-out scores (%) according to different metrics (F_1, precision, recall, accuracy and area under the ROC curve) for English a Dutch single feature type systems.

	Feature type	Cross-validation scores					hold-out scores				
		F_1	P	R	Acc	AUROC	F_1	P	R	Acc	AUROC
English											
	word *n*-grams	60.09	60.49	59.69	96.22	78.87	58.35	57.12	59.64	96.27	78.79
	subjectivity lexicons	56.82	73.32	46.38	96.64	72.77	56.16	72.61	45.78	96.87	72.50
	character *n*-grams	52.69	58.70	47.80	95.91	73.06	53.33	62.37	46.59	96.43	72.65
	term lists	40.48	38.98	42.12	94.10	69.41	39.56	39.56	39.56	94.71	68.39
	topic models	17.35	9.73	79.91	63.72	71.41	15.70	8.72	78.51	63.07	70.44
Dutch											
	word *n*-grams	55.53	72.64	44.94	95.27	71.88	54.99	70.20	45.20	95.57	71.99
	subjectivity lexicons	54.34	54.12	54.56	93.97	75.65	51.82	50.61	53.09	94.09	74.90
	character *n*-grams	51.70	67.58	41.86	94.86	70.22	50.46	65.20	41.15	95.17	69.88
	term lists	28.65	19.36	55.10	81.97	69.48	25.13	16.73	50.53	81.99	67.26
	topic models	24.74	21.24	29.61	88.16	60.94	17.99	23.15	14.71	91.98	55.80

https://doi.org/10.1371/journal.pone.0203794.t007

combination with other features, they improve Dutch performance consistently
explanation for their varying performance in both languages would be that the
trained on the Dutch background corpus are of better quality than the English
a random selection of background corpus texts reveals that the English scrape
noisy data (i.e. low word-count posts and non-English posts) compared to the D
corpus.

A shallow qualitative analysis of the classification output provided insight int
classification mistakes.

Table 8 gives an overview of the error rates per cyberbullying category of th
ing and baseline systems. This could give an indication of the types of bullying
detect by the current classifier. All categories are always considered positive fo
(i.e. the error rate equals the false negative rate), except for *Sexual* and *Insult*
negative (in case of harmless sexual talk and 'socially acceptable' insulting lan
bitches, in for a movie?" the corresponding category was indicated, but the pos
annotated as cyberbullying) and *Not cyberbullying*, which is always negative. E
being lowest for the profanity baseline confirms that it performs particularly we
recall (at the expense of precision, see Table 5). When looking at the best syste
guages, we see that *Defense* is the hardest category to classify. This should no
the category comprises defensive posts from bystanders and victims, which co
sive language than cyberbullying attacks and are often shorter in length than t
tive defensive posts (i.e. a subcategory of *Defense*) which attack the bully are,
often correctly classified. There are not sufficient instances of the *Encourageme*
either language in the hold-out set to be representative. In both languages, thr
incidences of sexual harassment are most easily recognisable, showing (far) lo
than the categories *Defamation*, *Defense*, *Encouragements to the harasser*, an

A qualitative error analysis of the English and Dutch predictions reveals that
often contain aggressive language directed at a second person, often denoting

Table 8. Error rates (%) per cyberbullying subcategory on hold-out for English and Dutch systems.

	Category	Nr. occurrences in hold-out	Profanity baseline	Word ***n***-gram baseline	Best system
English					
	Curse	*n* = 109	14.68	30.28	24.77
	Defamation	*n* = 21	23.81	47.62	38.10
	Defense	*n* = 165	22.42	52.12	43.64
	Encouragement	*n* = 1	0.00	100.00	100.00
	Insult	*n* = 345	26.67	41.74	35.94
	Sexual	*n* = 165	63.80	21.47	21,47
	Threat	*n* = 12	8.33	41.67	25.00
	Not cyberbullying	*n* = 10,714	36.94	1.10	0.76
Dutch					
	Curse	*n* = 96	39.58	50.00	22.92
	Defamation	*n* = 6	100.00	66.67	33.33
	Defense	*n* = 200	52.50	63.50	46.00
	Encouragement	*n* = 5	40.00	60.00	40.00
	Insult	*n* = 355	43.38	47.89	28.17
	Sexual	*n* = 37	37.84	21.62	27.03
	Threat	*n* = 15	33.33	46.67	20.00
	Not cyberbullying	*n* = 7,295	15.63	1.23	3.07

https://doi.org/10.1371/journal.pone.0203794.t008

or containing sexual and profanity words. We see that misclassifications are oft containing just a few words and that false negatives often lack explicit verbal s lying (e.g. insulting or profane words) or are ironic (examples 2 and 3). Additior that cyberbullying posts containing misspellings or grammatical errors and inc are also hard to recognise as such (examples 4 and 5). The Dutch and English c all similar with respect to qualitative properties of classification errors.

2. *You might want to do some sports ahah x*
3. *Look who is there… my thousandth anonymous hater*, *congratulations*!
4. *ivegot 1 word foryou… yknow whatit is*? *! slut*
5. *One word for you*: *G—A—…*

In short, the experiments show that our classifier clearly outperforms both a based and word *n*-gram baseline. However, analysis of the classifier output rev negatives often lack explicit clues that cyberbullying is going on, indicating tha might benefit from irony recognition and integrating world knowledge to captu implicit realisations of cyberbullying.

Our annotation scheme allowed to indicate different author roles, which prov insight into the realisation of cyberbullying. Table 9 presents the error rates of the different author roles, being harasser, victim, and two types of bystanders. the error rates are high for *bystander assistant* and *victim*, but there are not su in the hold-out set of the former role for either language to be representative. I the *victim* class of 50.39% and 54% in English and Dutch respectively indicate hard to recognise by the classifier. A possible explanation for this could be that in our corpus either expressed powerlessness facing the bully (example 6) or e explicit aggressive language as well (example 7).

6. *Your the one going round saying im a cunt and a twat and im ugly. tbh all i ing up for myself*.
7. *You're fucked up saying I smell from sweat*, *because unlike some other pec day BITCH*

According to the figures, the most straightforward roles in detection are *byst* and *harasser*.

Table 9. Error rates (%) per cyberbullying participant role on hold-out for English and Dutch systems.

	Participant role	Nr. occurrences in hold-out	Profanity baseline	Word ***n***-gram baseline	Best system
English					
	Harasser	*n* = 328	20.43	48.48	43.60
	Bystander assistant	*n* = 2	50.00	100.00	100.00
	Bystander defender	*n* = 39	7.69	38.46	25.64
	Victim	*n* = 129	27.91	57.36	50.39
	Not cyberbullying	*n* = 10872	37.64	1.24	0.89
Dutch					
	Harasser	*n* = 261	47.13	56.70	29.89
	Bystander assistant	*n* = 6	50.00	66.67	50.00
	Bystander defender	*n* = 52	25.00	38.46	23.08
	Victim	*n* = 150	62.00	72.00	54.00
	Not cyberbullying	*n* = 7370	16.01	1.42	3.41

https://doi.org/10.1371/journal.pone.0203794.t009

Table 10. Overview of the most related cyberbullying detection approaches.

Reference	Classifier	Corpus	Bully rate	F_1 score
[44]	SVM	1,762 tweets	39%	77%
[43]	wvec+SVM	1,762 tweets	39%	78%
[45]	smSDA+SVM	7,321 tweets	29%	72%
[45]	smSDA+SVM	1,539 MySpace posts	26%	78%

https://doi.org/10.1371/journal.pone.0203794.t010

In the light of comparison with state-of-the-art approaches to cyberbullying,
that competitive results are obtained with regard to [30–32, 41]. However, the
differences with respect to data collection, sources, and conceptualisations of b
allow for direct comparison. Table 10 presents the experimental results obtaine
who, like the current study, approach the task as detecting posts from bullies a
victims and bystanders. Given their experimental setup (i.e. task description, d
classifier), their work can be considered most similar to ours so their results mi
as benchmarks. Also here, a number of crucial differences with the current app
observed: Firstly, their corpora were collected using the keywords “bully”, “bul
lied”, which may bias the dataset towards the positive class and ensures that r
calisations are present in the positive class. Second, it is not clear which types
(i.e. explicit and implicit bullying, threats, insults, sexual harassment) are inclu
tive class. Furthermore, as can be deduced from Table 10, the datasets are cor
than ours and show a more balanced class distribution (respectively 39% cyber
in [43] and [44], and 29%/26% in [45]) than the ratio of bullying posts in our co
Table 3: 5% for English, 7% for Dutch). Hence, any comparison should be made
due to these differences.

These studies obtain higher scores on similar task but vastly different datase
[45] shows a great improvement in classification performance using deep repre
learning with a semantic-enhanced marginalized denoising auto-encoder over t
gram and topic modelling features.

Conclusions and future research

The goal of the current research was to investigate the automatic detection of
related posts on social media. Given the information overload on the web, man
for cyberbullying has become unfeasible. Automatic detection of signals of cyb
would enhance moderation and allow to respond quickly when necessary.

Cyberbullying research has often focused on detecting cyberbullying ‘attack
overlook other or more implicit forms of cyberbullying and posts written by vict
bystanders. However, these posts could just as well indicate that cyberbullying
main contribution of this paper is that it presents a system to automatically de
cyberbullying on social media, including different types of cyberbullying, cover
bullies, victims and bystanders. We evaluated our system on a manually annot
ing corpus for English and Dutch and hereby demonstrated that our approach c
applied to different languages, provided that annotated data for these languag

A set of binary classification experiments were conducted to explore the feas
matic cyberbullying detection on social media. In addition, we sought to detern
information sources contribute most to the task. Two classifiers were trained o
and Dutch ASKfm corpus and evaluated on a hold-out test of the same genre. O
reveal that the current approach is a promising strategy for detecting signals o

on social media automatically. After feature and hyperparameter optimisation
maximum F_1 score of 64.32% and 58.72% was obtained for English and Dutch, re
The classifiers hereby significantly outperformed a keyword and an (unoptimise
baseline. A qualitative analysis of the results revealed that false positives ofte
cyberbullying or offenses through irony, the challenge of which will constitute
area for future work. Error rates on the different author roles in our corpus reve
cially victims are hard to recognise, as they react differently in our corpus, sho
erlessness facing the bully or reacting in an assertive and sometimes even agg

As shown in [45] deep representation learning is a promising avenue for this
therefore intent to apply deep learning techniques to improve classifier perfor

Another interesting direction for future work would be the detection of fine-g
bullying categories such as threats, curses and expressions of racism and hate
a cascaded model, the system could find severe cases of cyberbullying with hi
would be particularly interesting for monitoring purposes. Additionally, our dat
detection of participant roles typically involved in cyberbullying. When applied
support on online platforms, such a system enables feedback in function of the
bully, victim, or bystander).

Author Contributions

Conceptualization: Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet
Ben Verhoeven, Guy De Pauw, Walter Daelemans, Véronique Hoste.

Data curation: Cynthia Van Hee, Gilles Jacobs, Els Lefever, Ben Verhoeven, Guy

Funding acquisition: Walter Daelemans, Véronique Hoste.

Methodology: Cynthia Van Hee, Gilles Jacobs, Bart Desmet, Els Lefever, Walter
Véronique Hoste.

Project administration: Walter Daelemans, Véronique Hoste.

Resources: Cynthia Van Hee, Gilles Jacobs, Els Lefever.

Software: Cynthia Van Hee, Gilles Jacobs, Bart Desmet.

Supervision: Walter Daelemans, Véronique Hoste.

Writing – original draft: Cynthia Van Hee, Gilles Jacobs.

Writing – review & editing: Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart
Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, Véronique Hoste.

References

1. Livingstone S, Haddon L, Görzig A, Ólafsson K. Risks and safety on the internet: The perspective of European children. Initial Findings. London: EU Kids Online; 2010.
2. Tokunaga RS. Following You Home from School: A Critical Review and Synthesis of Research on Cyberbullying Victimization. Computers in Human Behavior. 2010; 26(3):277–287. https://doi.org/10.1016/j.chb.2009.11.014
3. Mckenna KY, Bargh JA. Plan 9 From Cyberspace: The Implications of the Internet for Personality and Social Psychology. Personality & Social Psychology Review. 1999; 4(1):57–75. https://doi.org/10.1207/S15327957PSPR0401_6
4. Gross EF, Juvonen J, Gable SL. Internet Use and Well-Being in Adolescence. Journal of Social Issues. 2002; 58(1):75–90. https://doi.org/10.1111/1540-4560.00249

5. Juvonen J, Gross EF. Extending the school grounds?—Bullying experiences in cyberspace. Journal of School Health. 2008; 78(9):496–505. https://doi.org/10.1111/j.1746-1561.2008.00335.x PMID: 18786042
6. Hinduja S, Patchin JW. Bullies Move Beyond the Schoolyard: A Preliminary Look at Cyberbullying. Youth Violence And Juvenile Justice. 2006; 4(2):148–169. https://doi.org/10.1177/1541204006286288
7. Van Cleemput K, Bastiaensens S, Vandebosch H, Poels K, Deboutte G, DeSmet A, et al. 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; 2013.
8. Livingstone S, Haddon L, Vincent J, Giovanna M, Ólafsson K. Net Children Go Mobile: The Uk report; 2014. Available from: http://netchildrengomobile.eu/reports. [Accessed 30th March 2018].
9. O'Moore M, Kirkham C. Self-esteem and its relationship to bullying behaviour. Aggressive Behavior. 2001; 27(4):269–283. https://doi.org/10.1002/ab.1010
10. Fekkes M, Pijpers FIM, Fredriks AM, Vogels T, Verloove-Vanhorick SP. Do Bullied Children Get Ill, or Do Ill Children Get Bullied? A Prospective Cohort Study on the Relationship Between Bullying and Health-Related Symptoms. Pediatrics. 2006; 117(5):1568–1574. https://doi.org/10.1542/peds.2005-0187 PMID: 16651310
11. Cowie H. Cyberbullying and its impact on young people's emotional health and well-being. The Psychiatrist. 2013; 37(5):167–170. https://doi.org/10.1192/pb.bp.112.040840
12. Price M, Dalgleish J. Cyberbullying: Experiences, Impacts and Coping Strategies as Described by Australian Young People. Youth Studies Australia. 2010; 29(2):51–59.
13. Van Royen K, Poels K, Daelemans W, Vandebosch H. Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability. Telematics and Informatics. 2014;.
14. Cortes C, Vapnik V. Support-Vector Networks. Machine Learning. 1995; 20(3):273–297. https://doi.org/10.1007/BF00994018
15. Chang CC, Lin CJ. LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology (TIST). 2011; 2(3):27:1–27:27.
16. Dalla Pozza V, Di Pietro A, Morel S, Emma P. Cyberbullying among Young People. European Parliament; 2016.
17. Olweus D. Bullying at School: What We Know and What We Can Do. 2nd ed. Wiley; 1993.
18. Smith PK, Mahdavi J, Carvalho M, Fisher S, Russell S, Tippett N. Cyberbullying: its nature and impact in secondary school pupils. Journal of Child Psychology and Psychiatry. 2008; 49(4):376–385. https://doi.org/10.1111/j.1469-7610.2007.01846.x PMID: 18363945
19. Vandebosch H, Van Cleemput K. Defining cyberbullying: a qualitative research into the perceptions of youngsters. Cyberpsychology and behavior: the impact of the Internet, multimedia and virtual reality on behavior and society. 2008; 11(4):499–503. https://doi.org/10.1089/cpb.2007.0042
20. Vandebosch H, Van Cleemput K. Cyberbullying among youngsters: profiles of bullies and victims. New Media & Society. 2009; 11(8):1349–1371. https://doi.org/10.1177/1461444809341263
21. Dooley JJ, Cross D. Cyberbullying versus face-to-face bullying: A review of the similarities and differences. Journal of Psychology. 2009; 217:182–188.
22. Slonje R, Smith PK, Frisén A. The Nature of Cyberbullying, and Strategies for Prevention. Compututers in Human Behavior. 2013; 29(1):26–32. https://doi.org/10.1016/j.chb.2012.05.024
23. Van Hee C, Verhoeven B, Lefever E, De Pauw G, Daelemans W, Hoste V. Guidelines for the Fine-Grained Analysis of Cyberbullying, version 1.0. LT3, Language and Translation Technology Team–Ghent University; 2015. LT3 15-01.
24. Salmivalli C, Lagerspetz K, Björkqvist K, Österman K, Kaukiainen A. Bullying as a group process: Participant roles and their relations to social status within the group. Aggressive Behavior. 1996; 22(1):1–15. https://doi.org/10.1002/(SICI)1098-2337(1996)22:1%3C1::AID-AB1%3E3.0.CO;2-T
25. Vandebosch H, Van Cleemput K, Mortelmans D, Walrave M. 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; 2006. Available from: https://wise.vub.ac.be/fattac/mios/Eindrapport%20cyberpesten%20viwta%202006.pdf. [Accessed 30th March 2018].
26. Dehue F, Bolman C, Völlink T. Cyberbullying: Youngsters' Experiences and Parental Perception. Cyberpsychology and Behavior. 2008; 11(2):217–223. https://doi.org/10.1089/cpb.2007.0008 PMID: 18422417
27. Salmivalli C, Pöyhönen V. In: Cyberbullying in Finland. 2nd ed. Wiley-Blackwell; 2012. p. 57–72.
28. Salawu S, he Y, Lumsden J. Approaches to Automated Detection of Cyberbullying: A Survey. IEEE Transactions on Affective Computing. forthcoming;p. 1.

29. Reynolds K, Kontostathis A, Edwards L. Using Machine Learning to Detect Cyberbullying. In: Proceedings of the 2011 10th International Conference on Machine Learning and Applications and Workshops. ICMLA'11. Washington, DC, USA: IEEE Computer Society; 2011. p. 241–244.
30. Dinakar K, Reichart R, Lieberman H. Modeling the Detection of Textual Cyberbullying. In: The Social Mobile Web. vol. WS-11-02 of AAAI Workshops. AAAI; 2011. p. 11–17.
31. Yin D, Davison BD, Xue Z, Hong L, Kontostathis A, Edwards L. Detection of Harassment on Web 2.0. In: Proceedings of the Content Analysis in the Web 2.0 (CAW2.0). Madrid, Spain; 2009.
32. Dadvar M. Experts and machines united against cyberbullying [PhD thesis]. University of Twente; 2014.
33. Nahar V, Al-Maskari S, Li X, Pang C. Semi-supervised Learning for Cyberbullying Detection in Social Networks. In: ADC.Databases Theory and ApplicationsSpringer International Publishing; 2014. p. 160–171.
34. Bayzick J, Kontostathis A, Edwards L. Detecting the Presence of Cyberbullying Using Computer Software. In: Proceedings of the 3rd International Web Science Conference. WebSci11; 2011.
35. Sui J. Understanding and Fighting Bullying with Machine Learning [PhD thesis]. Department of Computer Sciences, University of Wisconsin-Madison; 2015.
36. Galán-García P, Puerta JGdl, Gómez CL, Santos I, Bringas PG. Supervised machine learning for the detection of troll profiles in twitter social network: application to a real case of cyberbullying. Logic Journal of the IGPL. 2016; 24(1):42–53. Available from: http://dx.doi.org/10.1093/jigpal/jzv048.
37. Raisi E, Huang B. Cyberbullying Detection with Weakly Supervised Machine Learning. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017. ASONAM'17. New York, NY, USA: ACM; 2017. p. 409–416.
38. Hosseinmardi H. Survey of Computational Methods in Cyberbullying Research. In: Proceedings of the First International Workshop on Computational Methods for CyberSafety. CyberSafety'16. New York, NY, USA: ACM; 2016. p. 4–4.
39. Nandhini B Sri, Sheeba JI. Online Social Network Bullying Detection Using Intelligence Techniques. Procedia Computer Science. 2015; 45:485–492. https://doi.org/10.1016/j.procs.2015.03.085
40. Chavan VS, S SS. Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In: 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI); 2015. p. 2354–2358.
41. Van Hee C, Lefever E, Verhoeven B, Mennes J, Desmet B, De Pauw G, et al. Detection and fine-grained classification of cyberbullying events. In: Angelova G, Bontcheva K, Mitkov R, editors. Proceedings of Recent Advances in Natural Language Processing, Proceedings; 2015. p. 672–680.
42. Van Royen K, Poels K, Vandebosch H. Harmonizing freedom and protection: Adolescents' voices on automatic monitoring of social networking sites. Children and Youth Services Review. 2016; 64:35–41. https://doi.org/10.1016/j.childyouth.2016.02.024
43. Zhao R, Zhou A, Mao K. Automatic Detection of Cyberbullying on Social Networks Based on Bullying Features. In: Proceedings of the 17th International Conference on Distributed Computing and Networking. No. 43 in ICDCN'16. New York, NY, USA: ACM; 2016. p. 43:1–43:6.
44. Xu JM, Jun KS, Zhu X, Bellmore A. Learning from Bullying Traces in Social Media. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. NAACL HLT'12. Stroudsburg, PA, USA: Association for Computational Linguistics; 2012. p. 656–666.
45. Zhao R, Mao K. Cyberbullying Detection Based on Semantic-Enhanced Marginalized Denoising Auto-Encoder. EEE Transactions on Affective Computing. 2017; 8(3):328–339. https://doi.org/10.1109/TAFFC.2016.2531682
46. Huang Q, Singh VK, Atrey PK. Cyber Bullying Detection Using Social and Textual Analysis. In: Proceedings of the 3rd International Workshop on Socially-Aware Multimedia. SAM'14. New York, NY, USA: ACM; 2014. p. 3–6.
47. Al-garadi MA, Dewi Varathan K, Devi Ravana S. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. Computers in Human Behavior. 2016; 63:433–443. https://doi.org/10.1016/j.chb.2016.05.051
48. Squicciarini A, Rajtmajer S, Liu Yh, Griffin C. Identification and Characterization of Cyberbullying Dynamics in an Online Social Network. In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015. ASONAM'15. New York, NY, USA: ACM; 2015. p. 280–285.
49. Chatzakou D, Kourtellis N, Blackburn J, De Cristofaro E, Stringhini G, Vakali A. Mean Birds: Detecting Aggression and Bullying on Twitter. In: Proceedings of the 2017 ACM on Web Science Conference. WebSci'17. New York, NY, USA: ACM; 2017. p. 13–22.

50. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer PW. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research (JAIR). 2002; 16:321–357. https://doi.org/10.1613/jair.953

51. Van Hee C, Lefever E, Verhoeven B, Mennes J, Desmet B, De Pauw G, et al. Automatic detection and prevention of cyberbullying. In: Lorenz P, Bourret C, editors. International Conference on Human and Social Analytics, Proceedings. IARIA; 2015. p. 13–18.

52. Hinduja S, Patchin JW. Cyberbullying: Neither an epidemic nor a rarity. European Journal of Developmental Psychology. 2012; 9(5):539–543. https://doi.org/10.1080/17405629.2012.706448

53. Slonje R, Smith PK. Cyberbullying: Another main type of bullying? Scandinavian Journal of Psychology. 2008; 49(2):147–154. https://doi.org/10.1111/j.1467-9450.2007.00611.x PMID: 18352984

54. O'Sullivan PB, Flanagin AJ. Reconceptualizing 'flaming' and other problematic messages. New Media & Society. 2003; 5(1):69–94. https://doi.org/10.1177/1461444803005001908

55. Willard NE. Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress. 2nd ed. Research Publishers LLC; 2007.

56. Bastiaensens S, Vandebosch H, Poels K, Van Cleemput K, DeSmet A, De Bourdeaudhuij I. Cyberbullying on social network sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully. Computers in Human Behavior. 2014; 31:259–271. https://doi.org/10.1016/j.chb.2013.10.036

57. Salmivalli C. Bullying and the peer group: A review. Aggression and Violent Behavior. 2010; 15(2):112–120. https://doi.org/10.1016/j.avb.2009.08.007

58. Bastiaensens S, Vandebosch H, Poels K, Van Cleemput K, DeSmet A, De Bourdeaudhuij I. 'Can I afford to help?' How affordances of communication modalities guide bystanders' helping intentions towards harassment on social network sites. Behaviour & Information Technology. 2015; 34(4):425–435. https://doi.org/10.1080/0144929X.2014.983979

59. Salmivalli C, Voeten M, Poskiparta E. Bystanders Matter: Associations Between Reinforcing, Defending, and the Frequency of Bullying Behavior in Classrooms. Journal of Clinical Child & Adolescent Psychology. 2011; 40(5):668–676. https://doi.org/10.1080/15374416.2011.597090

60. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J. brat: a Web-based Tool for NLP-Assisted Text Annotation. In: Proceedings of the Demonstrations Session at EACL 2012. Avignon, France; 2012. p. 102–107.

61. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960; 20(1):37–46. https://doi.org/10.1177/001316446002000104

62. Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin. 1971; 76(5):378–382. https://doi.org/10.1037/h0031619

63. McHugh ML. Interrater reliability: the kappa statistic. Biochemia Medica. 2012; 22(3):276–282. https://doi.org/10.11613/BM.2012.031 PMID: 23092060

64. He H, Garcia EA. Learning from imbalanced data. IEEE Transactions on knowledge and data engineering. 2009; 21(9):1263–1284. https://doi.org/10.1109/TKDE.2008.239

65. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LIBLINEAR: A Library for Large Linear Classification. Journal of Machine Learning Research. 2008; 9:1871–1874.

66. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12:2825–2830.

67. Desmet B. Finding the online cry for help: automatic text classification for suicide prevention [PhD thesis]. Ghent University; 2014.

68. Hoste V. Optimization Issues in Machine Learning of Coreference Resolution [PhD thesis]. Antwerp University; 2005.

69. van de Kauter M, Coorman G, Lefever E, Desmet B, Macken L, Hoste V. LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit. Computational Linguistics in the Netherlands Journal. 2013; 3:103–120.

70. Osgood CE, Walker EG. Motivation and language behavior: A content analysis of suicide notes. The Journal of Abnormal and Social Psychology. 1959; 59(1):58. https://doi.org/10.1037/h0047078

71. Jijkoun V, Hofmann K. 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. Stroudsburg, PA, USA; 2009. p. 398–405.

72. De Smedt T, Daelemans W. "Vreselijk mooi!" ("Terribly Beautiful!"): A Subjectivity Lexicon for Dutch Adjectives. In: Proceedings of the Eight International Conference on Language Resources and Evaluation. LREC'12. Istanbul, Turkey; 2012. p. 3568–3572.

73. Hu M, Liu B. Mining and summarizing customer reviews. In: Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining. KDD04. ACM; 2004. p. 168–177.

74. Wilson T, Wiebe J, Hoffmann P. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. HLT'05. Association for Computational Linguistics; 2005. p. 347–354.

75. Stone PJ, Dunphy DCD, Smith MS, Ogilvie DM. The General Inquirer: A Computer Approach to Content Analysis. The MIT Press; 1966.

76. Nielsen FÅ. A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs. In: Rowe M, Stankovic M, Dadzie AS, Hardey M, editors. Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages. vol. 718 of CEUR Workshop Proceedings. CEUR-WS.org; 2011. p. 93–98.

77. Mohammad S, Dunne C, Dorr B. Generating High-coverage Semantic Orientation Lexicons from Overtly Marked Words and a Thesaurus. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2. EMNLP'09. Stroudsburg, PA, USA: Association for Computational Linguistics; 2009. p. 599–608.

78. Pennebaker JW, Francis ME, Booth RJ. Linguistic Inquiry and Word Count: LIWC 2001. Mahwah, NJ: Lawrence Erlbaum Associates; 2001.

79. Zijlstra H, Van Meerveld T, Van Middendorp H, Pennebaker JW, Geenen R. De Nederlandse versie van de 'linguistic inquiry and word count' (LIWC). Gedrag Gezond. 2004; 32:271–281.

80. Rehurek R, Sojka P. Software framework for topic modelling with large corpora. In: The LREC 2010 Workshop on new Challenges for NLP Frameworks. University of Malta; 2010. p. 45–50.

81. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research. 2003; 3:993–1022.

82. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. Journal of the American Society for Information Science. 1990; 41:391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6%3C391::AID-ASI1%3E3.0.CO;2-9

83. Baroni M, Bernardini S. BootCaT: Bootstrapping Corpora and Terms from the Web. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation. LREC'04; 2004. p. 1313–1316.

84. Fawcett T. An introduction to ROC analysis. Pattern recognition letters. 2006; 27(8):861–874. https://doi.org/10.1016/j.patrec.2005.10.010