Summer 6-15-2016

DETECTING CYBERBULLYING IN ONLINE COMMUNITIES

Uwe Bretschneider
Martin-Luther-University Halle-Wittenberg, uwe.bretschneider@wiwi.uni-halle.de

Ralf Peters
Martin-Luther-University Halle-Wittenberg, ralf.peters@wiwi.uni-halle.de

DETECTING CYBERBULLYING IN ONLINE COMMUNITIES

Research

Bretschneider, Uwe, Martin-Luther-University Halle-Wittenberg, Halle, Germany, uwe.bretschneider@wiwi.uni-halle.de

Peters, Ralf, Martin-Luther-University Halle-Wittenberg, Halle, Germany, ralf.peters@wiwi.uni-halle.de

Abstract

Online communities are platforms enabling their users to interact over the web. In particular, they are popular among adolescents as a tool to discuss topics of mutual interest. However, offending communication is a growing issue in these online environments. In its basic form, the process of sending messages over electronic media to cause psychological damage to a victim is called online harassment. In a more severe form, cyberbullying is the process of sending offending messages several times to the same victim by the same offender. In this work, we propose an approach to detect cyberbullies and their victims. Identifying and aiding victims received only brief attention in existing work. We introduce a harassment graph to capture multiple message exchanges comprising cyberbullying cases. We show that our approach is able to precisely detect cyberbullies and their victims. Additionally, we propose metrics to measure the severity of online harassment and cyberbullying cases in terms of quantitative aspects. In particular, the metrics allow to identify victims of severe cyberbullying cases and might be used as an early indicator to provide fast and selective aid by administrators. We further propose use cases for our approach in online communities to tackle the problem of cyberbullying.

Keywords: cyberbullying, online harassment, offending communication, online communities

1 Introduction

Online communities are platforms that enable their users to interact over the web. Common forms of online communities are, for example, forums, discussion boards and social networks. They are popular among adolescents as a tool to discuss topics of mutual interest (Lenhart, 2015). More than 90% of teenagers are online on a daily basis including over 20% that are almost constantly online to stay up to date (Lenhart, 2015). Since these platforms allow unrestricted and often anonymous exchange of content, they are vulnerable for abuse, especially in form of offending communication. Offending communication includes online harassment and cyberbullying, which are growing issues in online environments that involve user interaction (Jones, 2013). Online harassment is the process of sending messages over electronic media to cause psychological harm to a victim. If the same person sends such messages several times to the same victim, the process is called cyberbullying (Tokunaga, 2010).

Online harassment and cyberbullying may lead to serious consequences like depression for the victims (Patchin and Hinduja, 2013; Tokunaga, 2010; Li, 2007). In extreme cases, the consequences can be even more severe, especially for adolescents. Particularly, two cases of suicide attracted public attention in 2014. A 14-year-old girl from Italy committed suicide after being offended on the social network Ask.fm by several anonymous users (BBC News, 2014). Another suicide was caused by a Facebook user threatening repeatedly a 17-year-old boy to make him stay away from his former girlfriend (Dailymail, 2014). In both cases the message exchange caused psychological damage to the victims. More seriously, it was not recognized that the victims require aid to endure this kind of damage. As a recent decision from the

European Court of Human Rights shows, the operators of an online community might be legally responsible for psychological damage that is caused by the publication of content (European Court of Human Rights, 2015). Consequently, a lot of online communities like Facebook and YouTube introduced community standards to approach the problem of offending communication. They encourage victims of online harassment and cyberbullying to report such cases. Administrators manually review these reported cases. However, due to the vast amount of messages within online communities this task is cumbersome and time-consuming.

Current research offers approaches to detect online harassment automatically. Most of these methods focus on an isolated analysis of the corresponding messages excluding the message context. However, the detection of cyberbullying requires the analysis of interrelated messages and the identification of the involved roles (Xu et al., 2012) as a bully sends its victim offending messages multiple times. Consequently, an extension is required to adapt existing methods for the detection of online harassment to the problem of cyberbullying detection. Additionally, current methods to detect online harassment predominantly rely on bag-of-words or n-gram models, which have limited capabilities to detect the persons referenced in online harassment messages (Chen et al., 2012; Bretschneider et al., 2014). Therefore, the identification of the involved roles in cyberbullying is even more difficult.

In this work, we extend prior research by concentrating on the detection of cyberbullying and its involved actors, especially the victims. Our contribution is threefold: First, we introduce a graph-based model to structure observed offending communication between users of an online community. We build this graph by analyzing the message context, which includes all messages exchanged between these users. Second, we propose a method to detect cyberbullying in online communities based on the identification of online harassment and the identified actors in the graph model. In contrast to existing approaches, we also focus on the detection of victims. Furthermore, we propose metrics based on the harassment graph to measure the severity of online harassment and cyberbullying cases. Third, we develop two annotated datasets including the referenced victims to evaluate our approach. Additionally, the datasets might be used as a benchmark for further research.

The paper is structured as follows. Section 2 presents an overview of existing research on cyberbullying detection. In section 3 the proposed method is explained in detail. We evaluate our proposed method in section 4. Practical applications and limitations are discussed in section 5 and section 6. Finally, the results of this work are summarized in section 7.

2 Related Work

The detection of cyberbullying consists of the detection of online harassment messages and the involved persons, especially the offender and his victim (Xu et al., 2012). Consequently, the detection of cyberbullying relies above all on approaches to identify online harassment. Since online harassment detection is an emerging research field, there is only a limited amount of work available. Existing work predominantly employs lexicon and machine learning approaches.

Lexicon approaches utilize wordlists containing known profane words to match them against a given text. In their naïve form, they classify a text as online harassment, if it contains offending words. The classification performance varies considerably depending on the wordlist used (Sood et al., 2012; Kontostathis et al., 2013). Kontostathis et al. (2013) observe that large wordlists improve the recall while reducing the precision. In contrast, smaller wordlists containing mainly severe offending words lead to the opposite effect. The performance is considerably improved by searching for user identifiers and offending words in combination (Chen et al., 2012; Bretschneider et al., 2014). Machine learning approaches are able to learn classification rules automatically by analyzing pre-classified training examples. A preprocessing step transforms a given text into an n-dimensional vector containing the features characterizing the text. These features are comprised of the words itself, n-grams, part-of-speech (POS) tags and other characteristics or a combination of these. Machine learning approaches often achieve slightly better classification performance compared to lexicon approaches (Kontostathis et al., 2013;

Sood et al., 2012; Dinakar et al., 2012). However, due to the sparse amount of online harassment messages and the lack of annotated datasets, it can be cumbersome to collect an adequate amount of training data (Kontostathis et al., 2013; Sood et al., 2012).

The above-mentioned approaches use bag-of-words or n-gram models to structure texts. The bag-of-words model disregards the sequence of the words and structures them in an isolated manner within a multiset. The n-gram model retains the sequence of n consecutive words to preserve a small proportion of the context. However, these approaches have limited capabilities to model relations between persons and profane words explicitly as such relations are expressed by a potentially large sequence of words. Consequently, the identification of the referenced victim is more difficult. Chen et al. (2012) overcome this restriction by introducing a finer grained text model based on a dependency graph used by the Stanford natural language processing toolkit. Thereby, relations between words within a sentence are accessible. The authors propose a lexicon approach extended by grammatical rules that leverage these relations to detect online harassment directed to a person. The parser needs to provide a comprehensive dependency graph for the approach to work, which is more difficult, if the text contains slang and abbreviations or has no clean sentence structure. Furthermore, the parser is not able to capture relations between words that are outside of a sentence. The authors remove punctuation marks to bypass this restriction. However, Chen et al. (2012) base their research on a different definition of offensive communication. They classify sentences as offensive that contain at least one strong offensive word or a combination of a weak offensive word and a person identifier. Consequently, the identification of a victim is not necessarily required and thus not the main focus of their work. In contrast, online harassment requires by definition a reference to a victim.

Although Cohen et al. (2014) underline the importance to aid victims of cyberbullying, this aspect is often excluded in automated analyses. There is only limited work available that focuses on the detection of the involved roles in cyberbullying cases. Xu et al. (2012) use a sequence labeling approach to identify different roles involved in online harassment cases. They first identify online harassment messages called bullying traces in social media with a machine learning approach. Bullying traces are isolated messages that express an experience with bullying or cyberbullying. After identifying these messages, role labeling for the author of the message and the mentioned entities within the message is applied. Xu et al. (2012) achieve good classification performance results for the role labeling of the author. Nevertheless, the classification performance for the mentioned persons within the text, especially the victim, is moderate. The role labeling is performed in an isolated manner ignoring multiple references across several messages to the same victim. Furthermore, victims are not uniquely identified, especially if the reference to the victim is expressed implicitly, for example, by using personal pronouns. Dadvar and de Jong (2012) detect online harassment messages by incorporating user metadata and the presence of profane words in a machine learning approach. The authors plan to extend this approach to identify bullies across several social networks. However, the identification of the victims is not yet part of their classification process. Hosseinmardi et al. (2014) examine cyberbullying behavior in the social network ask.fm by analyzing a substantial number of user profiles and messages and deriving an interaction graph. The interaction graph contains the users of the social network as nodes and the likes a user grants to another user as edges. In contrast to other approaches, the graph contains victims of cyberbullying by capturing the number of offending messages a user has received. However, Hosseinmardi et al. (2014) define cyberbullying as the process of sending a message containing at least one negative word to a user profile within this network. Thus, repeated interaction is not considered. Furthermore, they do not distinguish between negative content directed against the user profile and negative content without a certain target. Yet, their analysis reveals that users react differently to negative messages in dependence of their social isolation. Social isolation is measured by the number of likes a user grants to and receives from other users. As a key finding, they discover that users receiving a substantial amount of negative messages without any positive support in form of likes by others, grant less likes in return. As the positive support increases, these users tend to be more active. Hosseinmardi et al. (2014) assume that socially isolated users are more vulnerable to cyberbullying and thus require special attention.

Contrary to existing online harassment and cyberbullying detection approaches, we employ a pattern-based method as described in Bretschneider et al. (2014). This approach treats a text document as a sequence of words to preserve their order. In contrast to existing methods based on bag-of-words models, the sequence model allows to access the context of a detected profane word, which is important to detect the referenced victim of an offending statement. We choose the approach from Bretschneider et al. (2014) as it is able to detect an offending passage and allows us to search for the referenced victim within the context of this passage. In addition, we leverage a unique identifier, the username of the underlying online community, to recognize victims of multiple offending messages in different communication processes.

3 Proposed Method to Detect Cyberbullying

In this section, we describe our method to detect cyberbullying cases. After presenting the system architecture, we describe the tasks of username detection and the construction of the harassment graph in detail.

3.1 System Architecture

Our proposed method is based on the online harassment classification described in Bretschneider et al. (2014). We extend this approach by introducing three additional processing steps to identify cyberbullies and their victims. The resulting system architecture is depicted in figure 1.

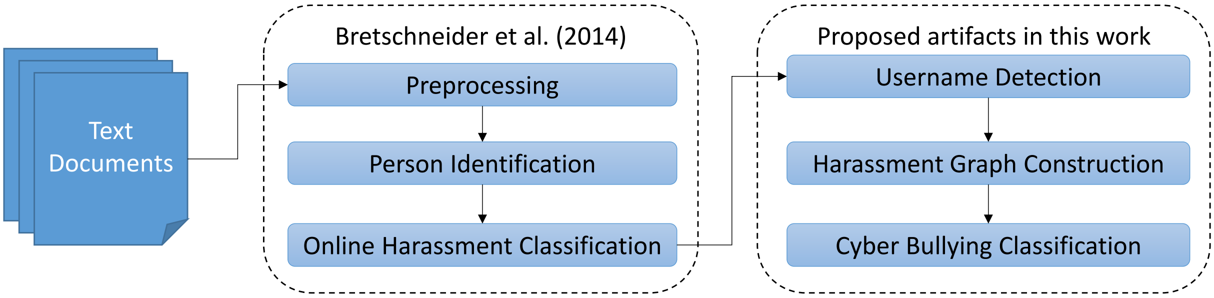
Text Documents
Bretschneider et al. (2014)
Preprocessing
Person Identification
Online Harassment Classification
Proposed artifacts in this work
Username Detection
Harassment Graph Construction
Cyber Bullying Classification

Figure 1. System architecture of the proposed cyberbullying detection approach.

The preprocessing step decomposes unstructured text into its components. These tokens are organized in a sequence to preserve their order and context. Additionally, the preprocessing module annotates these tokens with part-of-speech tags indicating their grammatical type, for example noun, verb or adjective. Furthermore, we utilize a normalization module that corrects spelling mistakes and resolves abbreviations or slang, which are typical for user generated content (Sood et al., 2012). In a next step, the person identification marks tokens that reference persons, i.e. usernames or personal pronouns. Finally, the online harassment classifier searches for relations between these person references and offending words by matching harassment patterns. Bretschneider et al. (2014) introduce seven harassment patterns that are able to match various ways of expressing online harassment within a sequence of words.

Approaches to detect online harassment treat offending messages in an isolated manner. Thus, they are not sufficient for cyberbullying detection as a cyberbully sends several interrelated messages to the same victim. Consequently, we propose three additional steps to detect interrelated online harassment messages that form cyberbullying cases. First, we introduce a module to identify the usernames and their roles involved in a cyberbullying case. We distinguish two roles, the cyberbully and the referenced victim. To correctly assign these roles, we employ the username as a unique identifier. We introduce a module that detects users referenced in the plaintext of a message by leveraging its context. In a second step, we build a directed graph containing the identified users as nodes and their roles indicated by directed edges representing the online harassment messages sent from a user to a victim. This way, we can map victims of offending communication and the corresponding offenders including the number of messages they sent. Finally, the cyberbullying classification step analyzes the resulting harassment graph. The module classifies users harassing other users multiple times as cyberbullies. In addition, we are able to identify victims that are harassed by multiple users including the amount of online harassment messages they received. Furthermore, we propose metrics to measure the severity of online harassment and cyberbullying cases.

3.2 Username Detection

Usernames are an inherent part of most online communities used as a unique identifier for online personas. We leverage usernames to identify the roles involved in online harassment and cyberbullying cases. The detection of the author of an offending message is trivial as his username is part of the message metadata. In contrast, the detection of referenced victims within the plaintext of a message is more complicated.

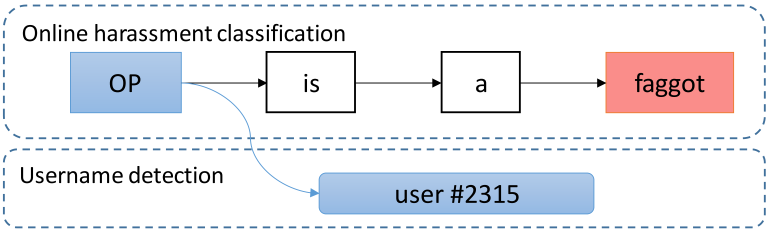

Figure 2. Online harassment classification and username detection.

As shown in figure 2, the pattern-based classifier detects offending passages within the plaintext containing a person reference and an offending word or phrase. Often, these references are stated implicitly, for example, by using personal pronouns. The usage of implicit references is typical for the progress of a discussion comprised of several interrelated messages. However, they are not sufficient as a unique identifier. The username detection resolves implicit references to explicit references expressed by a username. We propose the strategies listed in table 1 to perform this task. To aid this process, we collect usernames of authors from the message metadata distinguishing users occurring in a current discussion. Every implicit reference that is not resolvable with these strategies is disregarded and ignored in the following steps.

Strategy	Example
Reference to original poster	OP is <offending word>
Preceding reference to user	@<user> … you are <offending word>
Quote	[<user> said: …] you are <offending word>
Follow-up message	

Table 1. Username detection strategies.

The first strategy accounts for the abbreviation “OP” or its full form “original poster” that might be used directly in the offending passage as shorthand for the author publishing the first message of a discussion. Figure 2 shows an example from our dataset that contains such an abbreviation. We extend the person identification preprocessing step to detect such special words and recognize them as a person reference. Simultaneously, we are able to improve the online harassment classification performance as these types of references are not considered by Bretschneider et al. (2014). The username detection resolves then the reference by determining the corresponding author from the first message. The following strategy resolves implicit references in offending passages expressed by personal and reflexive pronouns. The corresponding explicit reference might be present in the context of the message. The sequence text model allows us to search for such an explicit reference within the sequence of words. If a username or a reference to the “OP” can be found, the implicit reference is resolved accordingly.

Typically, online communities offer a quote function to allow users to reference statements from others within the progress of the discussion. The quote strategy tries to find the author of a quoted text snippet, if such a quote is present before the offending passage. The quoted text snippet is matched against the preceding messages until its origin is found to determine the corresponding author. Finally, users might refer to directly preceding messages as an immediate response. Since a discussion contains messages in a chronological order, this temporal relation can be leveraged to identify follow-up messages. We assume, that users that actively participate in a discussion directly answer within the same day. Furthermore, a day is often the smallest unit displayed in the message timestamp. Thus, we resolve the implicit reference to the author of a directly preceding message, if the message is published within this time frame.

3.3 Harassment Graph Construction

The harassment graph is a directed graph containing harassment messages, their authors and the victims addressed in these messages. As an example, a snippet of the harassment graph resulting from the annotation process described in the subsequent chapter is depicted in figure 3.

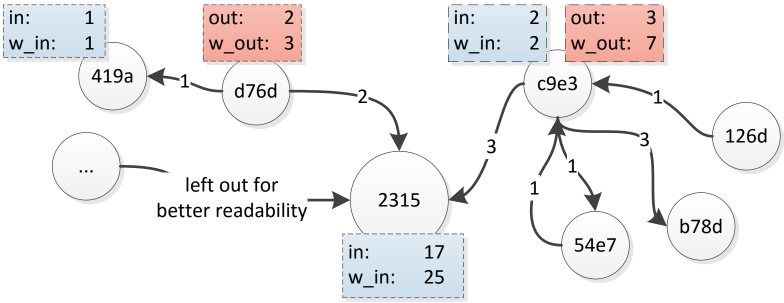

Figure 3. Harassment graph.

Formally, the harassment graph *G = (V, E)* is comprised of a set of nodes or vertices *V = (1, …, n)* and a set of directed edges *E = (1, …, k)*. The nodes represent users uniquely identified by their usernames. We anonymize the usernames by computing a hash value displaying the first four characters for better readability in the figure. The directed edges indicate the source and the target of an online harassment message. The edges are represented in an adjacency matrix X of the size $n \times n$ that contains binary variables. Each cell x_{ij} with a value of 1 represents a directed connection between the user i and j. Additionally, the number of online harassment messages sent from a user i to j are represented as weights in the weight matrix W of the size $n \times n$.

The classifier analyzes the harassment graph to identify cyberbullying cases. A cyberbullying case includes one offender sending at least two online harassment messages to a victim (Tokunaga, 2010). The harassment graph contains this information, as offenders are the source nodes of directed edges with a weight value greater than 1. In contrast to existing work, the harassment graph also allows us to focus on the identification of victims threatened by multiple offenders. For example, figure 3 depicts such a case where user "2315" is offended by multiple other nodes. To the best of our knowledge, current literature offers no definition for such cases. Thus, we define the problem of victim classification in analogy to cyberbullying as identifying victims offended by at least two distinct users.

Victim metric	Formula	Cyberbully metric	Formula
Indegree	$in(i) = \sum_{j=1}^{n} x_{ij}$	Outdegree	$out(i) = \sum_{j=1}^{n} x_{ji}$
Weighted indegree	$w\_in(i) = \sum_{j=1}^{n} x_{ij} * w_{ij}$	Weighted outdegree	$w\_out(i) = \sum_{j=1}^{n} x_{ji} * w_{ji}$

Table 2. Metrics to measure online harassment and cyberbullying severity.

In addition to existing research, we propose four metrics based on graph and social network analysis to further quantify the severity of online harassment and cyberbullying cases. The metrics are summarized in table 2. To apply them exclusively to cyberbullying cases, only directed edges with a weight greater than 1 are considered.

We employ the degree prestige metric (Wasserman and Faust, 1994) and the weighted indegree to indicate the psychological strain a victim has to bear. In social network analysis, the degree prestige is used to measure the popularity of a user. However, this interpretation reverses in the scenario of online harassment. For example, in figure 3 user "2315" has an indegree of 17, which means he faces 17 offenders (omitted in the figure for better readability).Thus, large values indicate severe cases including a large number of offenders referring to a single victim. Yet, the number of offenders alone does not cover repeated incidents caused by only a few offenders. Therefore, we employ the weighted indegree to account for such scenarios. This metric combines the number of offenders and the number of online harassment messages sent to a victim. User "2315" has a weighted indegree of 25 and thus received 25 online harassment messages in total.

In a similar way, we use the outdegree and weighted outdegree to quantify the aggressiveness of offending users. The outdegree measures the number of victims a user addresses. For example, user "c9e3" is referring to three victims resulting in an outdegree value of the same amount. A large value is an indicator for aggressive offenders in general as they refer to several victims. This assumption is supported more strongly, if the cases occur in different topics over a large time frame and the proportion of online harassment messages compared to neutral messages of this author is substantially large. However, these aspects are not yet considered and might be subject to further research. In analogy to the weighted indegree, the weighted outdegree considers the number of victims addressed by the offender and the total number of outgoing online harassment messages. If the weighted outdegree is substantially larger than the outdegree, it indicates that the offender focuses on only a few victims. User "c9e3", for example, has an outdegree of 3 and a weighted outdegree of 7. He seems to focus on the users "2315" and "b78d" as he directed six of seven online harassment messages against them.

4 Method and Evaluation

We discussed the proposed system architecture to detect cyberbullying cases in the previous section. This section is intended to assess the effectiveness of the artifacts contained in this architecture.

4.1 Dataset

To evaluate our proposed artifacts, we require an annotated dataset containing online harassment cases including the usernames of the offender and the victim. To the best of our knowledge, there are not yet any reference datasets containing this information. Consequently, we collect two datasets by downloading the general forum of the popular online games World of Warcraft[1] (dataset 1) and League of Legends[2] (dataset 2). We select these forums because of their popularity among adolescents. Adolescents, in particular, are vulnerable against online harassment and cyberbullying (Li, 2007). Furthermore, we evaluate the approach on two different datasets to prevent overfitting.

Since the amount of online harassment messages is typically sparse (Kontostathis et al., 2013; Sood et al., 2012) and the total amount of messages in these forums is substantially large, we preselect topics containing potential online harassment messages by searching for offending words contained in the wordlist from noswearing.com. This way, we selected 20 topics for each dataset. The annotation is performed by three human experts labeling each message in these topics. Online harassment cases are annotated as a tuple of the form: *(offender, victim, message)*. We only include tuples in the final dataset,

[1] http://eu.battle.net/wow/en/forum/872818/

[2] http://na.leagueoflegends.com/board/

if there is a consensus between at least two of the three annotators excluding the remaining tuples. The resulting dataset 1 contains 16975 messages with 137 harassment cases and dataset 2 contains 17354 messages with 207 harassment cases. To measure the inter-annotator agreement, we employ Fleiss' Kappa. As there is a substantial amount of neutral messages that would distort this measurement, we only consider cases judged as online harassment by at least one annotator. As a consequence, the measurement is more realistic. We measure a Fleiss' Kappa value of 0.51 for dataset 1 indicating moderate agreement and for dataset 2 a value of 0.72 indicating substantial agreement. These results emphasize the difficulty to identify offending passages, especially for borderline cases. In dataset 2, we observe more severe statements resulting in larger agreement between the annotators. We anonymized the usernames by employing a hash function on each username for the purpose of the publication. We provide access to the datasets under the URL http://ub-web.de/research/.

4.2 Method

Since the cyberbullying and victim classification process consist of three consecutive steps, we evaluate each step separately. First, we evaluate the online harassment classification. As evaluation metrics we employ precision, recall and f1-measure as recommended in (Sokolova and Lapalme, 2009). Precision measures the ratio of correctly classified instances to all instances classified as online harassment. Recall measures the ratio of correctly classified instances to all instances that really are online harassment. The f1 value is the harmonic mean between precision and recall. Second, we evaluate the username detection. We measure the amount of correctly detected usernames of victims among all detected online harassment cases. Third, we evaluate the detection of cyberbullies and the detection of victims of multiple offenders as described in the previous chapter. We measure precision, recall and f1 for both classification tasks.

To compare our results achieved in the online harassment classification step with the pattern-based approach from Bretschneider et al. (2014), we implement a baseline classifier as described in Chen et al. (2012). However, Chen et al. (2012) base their work on a different definition of offensive communication resulting in limited comparability to our results. Each message containing at least one strong offending word regardless of contained person references is classified as offending. While this is correct in their evaluation, it is not specific enough for our definition of online harassment, which necessarily requires a person reference. Since there is no public implementation of the Lexical Syntactic Feature (LSF) framework available, we followed their descriptions to implement the classifier. Additionally, we apply a support vector machine using the software RapidMiner as this approach performed moderately well in existing work (Kontostathis et al., 2013; Sood et al., 2012; Dinakar et al., 2012). We evaluate different configurations for the SVM (kernels: polynomial, dot, radial, anova and epachnenikov) and present the best result in terms of f1 in the evaluation section. To account for the substantially skewed class distribution, we activate the balance cost option in RapidMiner to adjust the settings accordingly. As machine learning approaches require training data, we split the dataset into training and test data. We follow the suggestions from Witten et al. (2011) applying a 3-fold cross validation for the evaluation. To ensure that the class distribution in each fold represents the class distribution of the whole dataset, we employ stratification.

4.3 Evaluation

The evaluation results for the online harassment classification are listed in table 3. The classification task is a difficult problem as the measurements demonstrate. The moderate overall results are associated with the characteristics of offending language in general and online harassment in particular. While online harassment is sparse in nature, offending language that is not necessarily directed against a person is fairly common in our datasets. Dataset 1 contains 4.17% offending messages marked by the LSF-like classifier from Chen et al. (2012), while only 0.81% of the messages are annotated as online harassment. Capturing the minor difference between offending language and online harassment by machine learning approaches requires adequate features and training examples. However, the imbalanced class ratio makes the collection of training data more difficult. In contrast to the results from Chen et al. (2012) and Bretschneider et al. (2014), we were not able to reproduce the substantially high f1 values achieved in their respective evaluations. Both classifiers were evaluated on short messages from Twitter (Bretschneider et al., 2014) and YouTube (Chen et al., 2012). The messages in our datasets are substantially longer containing statements with various referenced objects (i.e. companies or game-related characters).

	Dataset 1			Dataset 2		
Classifier	Precision	Recall	F1	Precision	Recall	F1
LSF-like classifier (baseline 1)	10.73%	55.47%	17.99%	13.56%	64.53%	22.41%
Machine learning (baseline 2)	14.35%	67.65%	23.68%	22.78%	55.39%	32.29%
Pattern-based	59.17%	52.21%	55.47%	74.10%	60.29%	66.49%

Table 3. Evaluation of online harassment classification.

Both baseline classifiers, the LSF-like and the machine learning approach (anova kernel), can only achieve moderate results in terms of precision and f1. A low precision value indicates a high amount of false positives resulting in falsely classified cyberbullies and victims in the subsequent steps. Thus, they increase the amount of work for administrators to manually examine the corresponding messages correcting these results. A high recall value indicates that a large amount of the actual offending messages are detected. By definition, cyberbullies write at least two offending messages to the same victim (Tokunaga, 2010) and thus, they can only be identified, if at least two cases are detected correctly. Consequently, it is important to detect a large amount of offending messages to cover preferably the complete amount of the messages sent by an offender. The pattern-based online harassment classifier achieves moderate recall values and precision values making it more suitable for the given task.

	Dataset 1		Dataset 2	
Strategy	Correctly transformed	Incorrectly transformed	Correctly transformed	Incorrectly transformed
Explicit references	1	0	8	0
Reference to original poster	2	0	7	2
Preceding reference to user	5	4	7	1
Quote	44	2	88	0
Follow-up message	7	1	6	2
Total amount	**59**	**7**	**116**	**5**
Relative amount (in %)	**89.39**	**10.61**	**95.87**	**4.13**

Table 4. Evaluation of username detection.

Table 4 contains the evaluation results for the username detection grouped by strategies. As the results show, users rarely use explicit references to refer to other users in an offending statement. Instead, they frequently use quotes. Quotes often can be resolved accurately by parsing the html code or applying text matching techniques. In contrast, follow-up messages and preceding references to other users are prone to errors, especially, if multiple references exist in the context. However, the approach is able to resolve a reasonable amount of indirect references used in the detected online harassment cases. In dataset 1 are 7 and in dataset 2 are 5 incorrectly transformed cases. Thus, the username detection is able to resolve 89.39% (dataset 1) and 95.87% (dataset 2) references correctly. However, there are 6 references in dataset 1 and 2 references in dataset 2 that could not be resolved as no strategy was applicable. By further analyzing these unresolved cases, we observe that most of these references are implicit. Especially in mutual discussions, users tend to omit explicit references as they are clear due to the ongoing discussion and its content. This way, users refer to each other even though their messages might be spread over the discussion. The approach is not able to analyze such references as they are expressed in a semantic way and thus none of the strategies is applicable. These cases might be addressed in further research.

	Cyberbully classification			Victim classification		
Dataset	Precision	Recall	F1	Precision	Recall	F1
Dataset 1	87.5%	53.85%	66.67%	66.67%	53.33%	59.26%
Dataset 2	93.33%	56%	70%	71.43%	57.69%	63.83%

Table 5. Evaluation of cyberbullying and victim classification.

Finally, table 5 summarizes the classification results for the cyberbully and victim classification. Considering the three consecutive steps performed to classify cyberbullies and victims, the achieved f1 values for cyberbully classification are considerably high and reasonable for victim classification. To correctly identify a cyberbullying case, at least two online harassment messages from the same author to the same correctly resolved victim need to be identified. As the results indicate, we are able to detect a fair amount of the actual cyberbullies and victims. In addition, the precision values indicate a low false positive rate. The low recall values mainly result from errors during the preceding steps. As we are interested in a measurement for the complete process, we consider in the calculation of the evaluation metrics for cyberbullying and victim classification errors during these preceding steps. Consequently, false negatives during online harassment classification and unresolved usernames reduce the recall value of cyberbullying classification.

By further analyzing the correctly identified cases, we observe that we were able to detect the severe cases measured by the metrics introduced in the previous section. This relation between the number of correctly detected cyberbullies and their weighted outdegree, respectively the number of correctly detected victims and their weighted indegree is depicted in figure 3. We choose the weighted out- and indegree to account for the total number of online harassment messages sent and received.

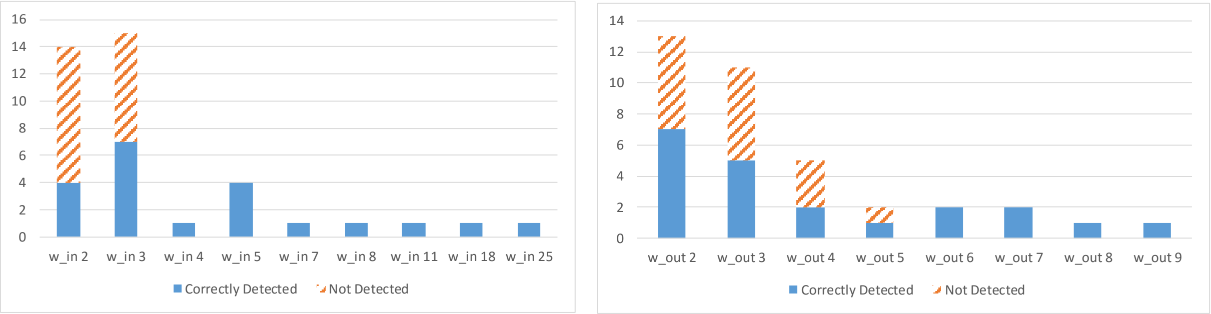

Figure 4. Correctly detected cyberbullies by weighted outdegree (left) and correctly detected victims by weighted indegree (right).

Since there is only a limited amount of work on the problem of cyberbullying classification available, we cannot directly compare our results. Although Xu et al. (2012) present the task of role labeling to classify the role of the author of a message and to identify the roles of the persons mentioned in the plaintext, they base their work on different definitions of these roles. Additionally, the evaluation is performed on a dataset with different characteristics containing short Twitter messages. The author role classification treats each message in an isolated manner deciding if the author is an offender. Thus, this problem is more similar to our online harassment classification task. Their person identification is a sequential tagging task annotating the roles of persons mentioned in a plaintext. Thus, it is comparable to our person identification and username detection task. Xu et al. (2012) achieve results of 53% precision and 42% recall emphasizing the difficulty of this task.

5 Practical Applications

Numerous online communities voluntarily committed themselves to introduce and enforce policies to maintain socially acceptable mutual communication. Such policies enable administrators to intervene in offending communication and punish the corresponding offender. However, due to the vast amount of messages published in online communities this task is labor-intensive. As a consequence, some online communities are not actively moderated and rely on their users to report abuses. However, a lot of victims isolate themselves to cope with online harassment or cyberbullying and thus do not report these cases (Li, 2007). The approach presented in this work can be utilized to aid human control instances and reduce their effort by automatically marking online harassment and cyberbullying cases.

Additionally, the harassment graph reveals further insight that an online harassment detection system alone cannot provide. First, if the system is implemented in the online community, additional information might be integrated in the harassment graph. Online harassment messages and their corresponding origin, i.e. a topic or a discussion, might be stored within the directed edges. This way, an administrator can easily navigate to the relevant section of the online community and evaluate the message context. Furthermore, the administrator can display all online harassment messages sent from or received by a certain node. Second, the introduced metrics help to estimate the severity of an online harassment or cyberbullying case. The offender metrics are an indicator to identify malicious users. The administrator might investigate all the published messages from this user and decide if further actions need to be taken. The victim metrics are an indicator for the severity of the perceived psychological damage. Instead of only punishing the offender, the administrator could aid the victim by asking about his condition. In online communities an offender might create new accounts to bypass message publication restrictions. Thus, protecting the victim might be more effective to avoid ongoing psychological damage. Furthermore, our approach is able to identify victims of severe online harassment cases with substantially high precision. As the case of the 14-year-old girl demonstrates, the identification of such cases is important, especially to intervene in an early stage.

Finally, the approach might be used to provide data for other research disciplines, i.e. social sciences, as proposed by Xu et al. (2012). The system is able to identify cyberbullying cases including the corresponding messages and involved users. The message context might be manually analyzed by experts to identify other roles like cyberbully assistants or bystanders.

6 Limitations

The severity of online harassment cases in terms of quality is not assessed by the proposed approach. Thus, the proposed metrics are entirely based on the quantity of offending messages instead of a combined measurement of quantity and quality. However, to assess the severity of online harassment cases is a challenging task even for expert annotators as the perceived severity varies among individuals (Tokunaga, 2010). Additionally, no temporal relation is considered. Online harassment messages sent in a small time frame to the same victim might cause more psychological damage than messages spanning over a larger period. Yet, the classification performance in terms of the achieved precision value is not sufficient for automated systems blocking malicious users or content. Falsely blocked users or messages due to low precision might cause frustration for the corresponding authors as they received unjust penalty. Precision might be further improved by introducing severity values and thus focusing on severe cases. Such cases might be blocked automatically while less severe cases might be reviewed by human control instances. Furthermore, we are not able to detect other roles involved in a cyberbullying case. Assistants, for example, intensify the severity of such cases as they support the involved bully (Xu et al., 2012). Finally, the detection of irony, sarcasm or longer statements paraphrasing online harassment is still an open problem. Currently, only the approach proposed by Dinakar et al. (2012) is capable of detecting paraphrased sexual harassment that alludes to characteristics of the opposite sex. Thus, the remaining cases need to be examined manually by personnel.

7 Conclusion

Offending communication is a growing issue in online environments that involve user interaction (Jones, 2013). In its basic form, the process of sending messages over electronic media to cause psychological damage to a victim is called online harassment. In a more severe form, the process of sending offending messages several times to the same victim by the same offender is called cyberbullying (Tokunaga, 2010). In this work we propose an approach to detect cyberbullies and their victims in online communities.

In current research, online harassment and cyberbullying cases are examined in an isolated manner detecting offending messages separately without focusing on the addressed victims. However, cyberbullying cases consist of several interrelated messages referring to the same victim. We extend current research to detect cyberbullying cases consisting of multiple message exchanges between the same users. First, we introduce the harassment graph to capture all offending messages as directed edges and the corresponding actors as nodes. We leverage the usernames already available in online communities to uniquely identify the involved actors. In discussions, users typically refer implicitly to each other, for example, by using personal pronouns as other users can derive the addressed user from the context of the discussion. However, implicit references are not sufficient to map them unambiguously to the harassment graph. Thus, we propose strategies to detect the usernames of referenced victims within the plaintext of a message as a second step. In a third step, we examine the resulting graph to classify cyberbullies. In contrast to existing research, we also focus on identifying victims of online harassment caused by several offenders referring to one victim. Identifying and aiding victims received only brief attention in existing work. Finally, we propose metrics to measure the severity of online harassment and cyberbullying cases in terms of quantitative aspects. The results show that our approach is able to detect the most severe cases accurately.

We introduce two labeled datasets to evaluate our approach as there are no reference datasets available that include information about the referenced victims. We provide access to the datasets as a benchmark and for further research[3]. The approach is evaluated in terms of precision, recall and f1-measure. Since there is only limited amount of work on cyberbullying detection available, we cannot directly compare our results to other approaches. We were able to achieve reasonable results for the cyberbullying classification task. Our approach yields 87.5% (dataset 1) and 93.33% (dataset 2) precision and 53.85% (dataset 1) and 56% (dataset 2) recall, which is substantially better than the achieved values of 53% precision and 42% recall from Xu at al. (2012) on their similar sequential tagging task. The results indicate that the presented approach might be used to aid administrators by identifying users involved in online harassment and cyberbullying cases within the vast amount of messages exchanged among the users of online communities. Administrators can easily view the corresponding messages and their context to decide if further actions need to be taken, especially if the victims might need assistance.

Further research might be conducted on improving the classification performance to create systems suitable for fully automated systems that restrict offending users. Currently, the harassment graph focuses on offenders and their victims. Other roles, for example cyberbully assistants, might be in the graph as well to gain further insight in the severity of such a case. As each online harassment message contains a timestamp, the message history is available in the online harassment graph. Thus, a dynamic analysis is possible, for example, to detect users that act as cyberbullies after being cyberbullied themselves. Further research might also extend the metrics in terms of qualitative aspects, for example, the severity of online harassment expressed by each message. At this time, only quantitative aspects, namely the number of offenders and messages are considered.

[3] http://ub-web.de/research/

References

BBC News (2014). “Cyberbullying suicide: Italy shocked by Amnesia Ask.fm case.” URL: http://www.bbc.com/news/world-europe-26151425 (visited on 09/14/2015).

Bretschneider, U., Wöhner, T. and Peters, R. (2014). “Detecting Online Harassment in Social Networks - Building a Better World through Information Systems.” In: *Proceedings of the International Conference on Information Systems 2014*. Auckland: New Zealand.

Chen, Y., Zhou, Y., Zhu, S. and Xu, H. (2012). “Detecting Offensive Language in Social Media to Protect Adolescent Online Safety.” In: *2012 International Conference on Privacy, Security, Risk and Trust (PASSAT)*. Amsterdam: Netherlands, 71–80.

Cohen, R., Rawat, R., Sun, W., Wang, D., Wexler, M., Lam, D.Y., Agarwal, N., Cormier, M., Jagdev, J., Jin, T., Kukreti, M., Liu, J. and Rahim, K. (2014). “Using computer technology to address the problem of cyberbullying.” In: *ACM SIGCAS Computers and Society* 44 (2), 52–61.

Dadvar, M. and de Jong, F. (2012). “Cyberbullying detection.” In: *Proceedings of the 21st International Conference on World Wide Web*. Lyon: France, 121–126.

Daily Mail Online (2015). “Teenage boy drowns himself in the sea after being trolled on Facebook by a former friend who was dating his ex-girlfriend.” URL: http://www.dailymail.co.uk/news/article-2638053/Teenage-boy-drowns-sea-trolled-Facebook-former-friend-dating-ex-girlfriend.html (visited on 09/14/2015).

Dinakar, K., Jones, B., Havasi, C., Lieberman, H. and Picard, R. (2012). “Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying.” In: *ACM Transactions on Interactive Intelligent Systems* 2 (3), 1–30.

European Court of Human Rights (2015). “Grand Chamber judgment Delfi AS v. Estonia - liability of Internet news portal for offensive online comments.” URL: http://hudoc.echr.coe.int/eng-press?i=003-5110487-6300958 (visited on 09/15/2015).

Hosseinmardi, H., Ghasemianlangroodi, A., Han, R., Lv, Q. and Mishra, S. (2014). “Towards Understanding Cyberbullying Behavior in a Semi-Anonymous Social Network.” In: *ASONAM: IEEE Computer Society 2014*. Beijing: China, 244–252.

Kontostathis, A., Reynolds, K., Garron, A. and Edwards, L. (2013). “Detecting cyberbullying.” In: *the 5th Annual ACM Web Science Conference*. Paris: France, 195–204.

Lenhart, A. (2015). “Teen, Social Media and Technology Overview 2015.” URL: http://www.pewinternet.org/files/2015/04/PI_TeensandTech_Update2015_0409151.pdf (visited on 09/14/2015).

Li, Q. (2007). “New bottle but old wine: A research of cyberbullying in schools.” In: *Computers in Human Behavior* 23 (4), 1777–1791.

Jones, L.M., Mitchell, K.J. and Finkelhor, D. (2013). “Online harassment in context: Trends from three Youth Internet Safety Surveys (2000, 2005, 2010).” In: *Psychology of Violence* 3 (1), 53–69.

Patchin, J.W. and Hinduja, S. (2013). “Cyberbullying among adolescents: implications for empirical research.” In: *The Journal of Adolescent Health: Official Publication of the Society for Adolescent Medicine* 53 (4), 431–432.

Sokolova, M. and Lapalme, G. (2009). “A systematic analysis of performance measures for classification tasks.” In: *Information Processing and Management* 45 (4), 427–437.

Sood, S.O., Churchill, E.F. and Antin, J. (2012). “Automatic identification of personal insults on social news sites.” In: *Journal of the American Society for Information Science and Technology* 63 (2), 270–285.

Tokunaga, R.S. (2010). “Following you home from school: A critical review and synthesis of research on cyberbullying victimization.” In: *Computers in Human Behavior* 26 (3), 277–287.

Wasserman, S. and Faust, K. (1994). “Social Network Analysis. Methods and Applications.” Cambridge: Cambridge University Press.

Witten, I. H., Frank, E. and Hall, M. A. (2011). “Data Mining: Practical Machine Learning Tools and Techniques.” San Francisco: Morgan Kaufmann Publishers.

Xu, J.-M., Jun, K.-S., Zhu, X. and Bellmore, A. (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*. Stroudsburg: USA, 656–666.