# Detecting Online Harassment in Social Networks

*Completed Research Paper*

**Uwe Bretschneider**
Martin-Luther-University
Halle-Wittenberg
Universitätsring 3
D-06108 Halle (Saale)
uwe.bretschneider@wiwi.uni-halle.de

**Thomas Wöhner**
Martin-Luther-University
Halle-Wittenberg
Universitätsring 3
D-06108 Halle (Saale)
thomas.woehner@wiwi.uni-halle.de

**Ralf Peters**
Martin-Luther-University
Halle-Wittenberg
Universitätsring 3
D-06108 Halle (Saale)
ralf.peters@wiwi.uni-halle.de

## Abstract

*Online Harassment is the process of sending messages for example in Social Networks to cause psychological harm to a victim. In this paper, we propose a pattern-based approach to detect such messages. Since user generated texts contain noisy language we perform a normalization step first to transform the words into their canonical forms. Additionally, we introduce a person identification module that marks phrases which relate to a person. Our results show that these preprocessing steps increase the classification performance. The pattern-based classifier uses the information provided by the preprocessing steps to detect patterns that connect a person to profane words. This technique achieves a substantial improvement compared to existing approaches. Finally, we discuss the portability of our approach to Social Networks and its possible contribution to tackle the abuse of such applications for the distribution of Online Harassment.*

**Keywords:** Online Harassment, Cyber Bullying

## Introduction

Web 2.0 applications enable users to publish content and connect with each other. In particular Social Networks and Social Broadcast Services like Facebook and Twitter enjoy huge popularity. Especially young people use them as a tool to maintain their relations to friends, classmates or fellow students (Easley and Kleinberg 2010). Since these platforms allow an unfiltered and sometimes anonymous exchange of content, new problems arise as well. Such problems are the missing protection of private information within user profiles, the questionable authenticity of users and the possibility of sending or broadcasting spam and offending messages.

Offending communication has already been an issue at schools and colleges in the form of Harassment and Bullying. With the rise of Social Networks the problem has been extended to Online Harassment and Cyber Bullying (Li 2007). Online Harassment is the process of sending messages over electronic media to

cause psychological harm to a victim. If such messages are sent several times by the same person to the same victim the process is called Cyber Bullying (Tokunaga 2010). Online Harassment and Cyber Bullying are growing in relevance and may lead to serious consequences like depression for the victims (Aponte and Richards 2013; Tokunaga 2010; Li 2007; Campbell 2005). Particularly the case of the 14 years old girl Nadia from Italy which committed suicide after being harassed on the Social Network Ask.fm attracted public attention (BBC News 2014). In this work we concentrate on Online Harassment methods since they are part of Cyber Bullying detection.

A victim has only limited options to defend himself. An offending message can be deleted if it is sent directly or if it is posted on the victim's profile. However, this requires the victim to read the message which might already inflict psychological harm. Depending on the reaction time of the victim the message might also be read by others before it is deleted. If the message is broadcasted to several receivers like on Twitter, the victim cannot delete it directly. Some Social Networks implement a reporting function to delete such messages by an administrator and possibly suspend the corresponding account. However, an offender can easily create a new account to bypass such restrictions. Another problem is that a high fraction of the victims isolate themselves and do not report such cases (Li 2007). Furthermore the barrier to send offending messages is lower in Social Networks compared to direct interaction. The victim cannot react in a direct way since the offender is possibly far away or anonymous. Finally, the reach of messages published online extend the reach of direct communication, especially if they are posted on a victims profile or send as broadcast message like on Twitter (Campbell 2005).

As suggested by Aponte and Richards (2013) these problems can be addressed by software systems which are able to block or mark Online Harassment messages. Software systems are more efficient than personnel due to the vast amount of messages in Social Networks. However, there is a lack of effective methods to realize an automatic detection for Social Networks (Kontostathis et al. 2013; Dinakar et al. 2012). Emerging approaches try to adapt methods from the research area of sentiment analysis. Sentiment analysis offers methods to classify texts regarding their contained sentiment into positive or negative (Tsytsaru and Palpanas 2012; Pang and Lee 2008). The detection of Online Harassment can be interpreted as a special problem of sentiment classification since such messages contain negative sentiment. We found, that sentiment analysis methods incorrectly classify a large amount of messages as Harassment (false positives). Since Online Harassment messages express a harmful statement related to a person and these methods are not able to detect relations between sentiments and persons, they are not suitable for Online Harassment detection. Moreover, messages in Social Networks contain noisy language including spelling errors, word variations and slang (Sood et al. 2012). Therefore, the methods for Online Harassment detection must also be robust against noisy language. Even though research on normalization of texts exists, these methods have not yet applied in the context of Online Harassment detection.

In order to address these problems, we extend current research on Online Harassment detection. Our contribution is twofold: Firstly, we introduce a new preprocessing step that indentifies phrases referring to a person. Secondly, we propose a pattern-based approach to identify Online Harassment. It is based on the detection of profane words and their links to recognized persons expressed by typical patterns. A text is interpreted as sequence-based model to match such patterns. We evaluate our method on the basis of a labeled Twitter dataset.

The rest of this paper is organized as follows: Section 2 introduces an overview of related work on Online Harassment detection. In section 3 the proposed method is presented in detail. Furthermore, we describe the development of the datasets which we construct from Twitter data. Section 4 contains the evaluation method. Each module of our proposed method is evaluated separately to distinguish between their effects on the classification results. Section 5 discusses the results and their portability into Social Networks. In section 6 limitations of the proposed approach are described. Finally, section 7 summarizes the results of this work and points out open problems for future research.

## Related Work

Since the research field of Online Harassment detection is still emerging, there is only a limited amount of work available. Currently three Online Harassment techniques approaches exist: wordlist-based, machine learning and rule-based approaches.

The first approach is based on wordlists containing known profane words. A document is interpreted as a bag-of-words model which is matched against the wordlist. The document is classified as Online Harassment if a match is found. Since the bag-of-words model treats all words in an isolated manner, these approaches are not able to explicitly model relations between persons and profane words. Furthermore, the classification performance varies considerably depending on the wordlist used (Sood et al. 2012). The work of Kontostathis et al. (2013) reveals that large wordlists result in the detection of a high percentage of Online Harassment messages while smaller wordlists result in less misclassifications.

The second approach is based on machine learning methods. These methods are able to learn classification rules automatically by detecting patterns in Online Harassment messages. They require manually annotated training data to learn such rules. However, due to the sparse amount of Online Harassment messages it can be cumbersome to collect an adequate amount of training data (Kontostathis et al. 2013; Sood et al. 2012). Machine learning approaches achieve slightly better classification performance than wordlist-based approaches (Kontostathis et al. 2013; Sood et al. 2012; Dinakar et al. 2012; Dinakar et al. 2011). However, they also treat input documents as a bag-of-words model sharing the limitations of wordlist-based approaches (Kontostathis et al. 2013).

The third approach is based on rule engines to analyze semantic relations within documents. Wordlist and machine learning techniques rely on explicitly formulated statements in a text. Dinakar et al. (2012) investigate the effect on the performance by incorporating a knowledge database and a rule engine in the classification process. Online Harassment content which is built upon implicit knowledge can be detected by such methods. For example, the sexually discriminatory message sent to a male: “why did you stop wearing makeup?” (Dinakar et al. 2012). Such techniques require thorough construction of knowledge databases. For the problem of detecting sexuality related harassment alone, Dinakar et al. (2012) construct around 200 assertions. These assertions allow the rule engine to infer conclusions whether a given statement is sexual harassment.

In addition to scientific methods, first commercial systems like XRayData[1] have already been introduced to monitor certain Social Network accounts. Parents can use such tools to intervene in potentially harmful conversations. These systems assume that a human will analyze messages marked by the system and draw a final conclusion. However, they are only able to protect a predefined set of certain users. Furthermore, there are no investigations regarding the classification performance of such tools yet.

Contrary to these approaches we propose a method that treats a document as a sequence of words to preserve their order. This allows us to focus on the identification of references between profane words and potential victims. We specify patterns that express typical links between such words to improve the classification performance. In contrast to rule-based approaches, our approach relies on a small set of patterns reducing the effort compared to the maintenance of a knowledge database.

# Proposed method

In this section we first introduce the system architecture of our proposed method. The associated modules are described in detail in the subsequent sections including a prototypical implementation using the example of Twitter.

## *System architecture*

The goal of the proposed method is the automatic detection of Online Harassment messages in Social Networks. Two requirements arise from the definition of Online Harassment: Firstly, the method has to identify content that might cause psychological harm. Secondly, the method has to detect links between such content and references to a person. Additionally, according to Sood et al. (2012), the method has to be robust against noisy language. To meet these requirements we introduce the architecture shown in Figure 1 which maps each requirement to a module. The modules are further organized in a three step process consisting of text preprocessing, person identification and classification. This modular architecture allows us to evaluate and exchange each module separately.

[1] http://xraydata.com/

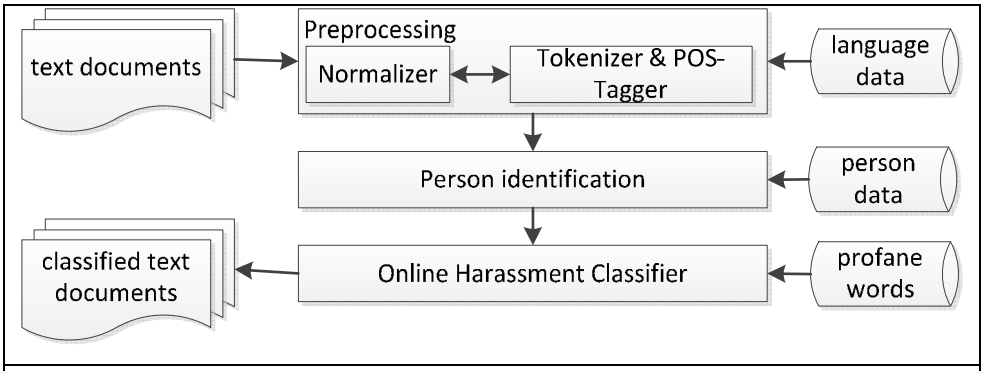


**Figure 1. System architecture**

In a first step the text documents are preprocessed. A minimal requirement for a wordlist-based classifier comprises a tokenization of the unstructured text into word chunks. In addition these chunks might be annotated by their part-of-speech (POS) tag, for example as noun, verb or adjective. While current work focuses on these preprocessing steps, we investigate an additional step and its effect on the classification performance. We integrate a normalization module that transforms noisy text consisting of spelling mistakes, slang and abbreviations into a canonical form. The canonical form of a word corresponds to its form found in a reference dictionary. Using this module we address the requirement stated by Sood et al. (2012) to handle dynamically changing and noisy language of Web 2.0 applications.

After preprocessing, the tokens are annotated by the person identification module. Existing work covers only partially the requirements stated in the definition of Online Harassment. The pure presence of a profane word is not sufficient to classify a message as Online Harassment. The purpose of the person identification step is to incorporate the requirement of addressing a victim within a document. For this purpose it identifies and marks words or phrases that refer to a person using POS tags in combination with language data and data from the corresponding Social Network, i.e. usernames.

Finally, the Online Harassment classifier uses the information from the preceding steps to solve the binary classification problem. In contrast to existing research, the document is only classified as Online Harassment if a link between the profane word and the word that relates to the victim exists. As stated by Sood et al. (2012) it is necessary to incorporate the context of profane words to achieve better classification results compared to a bag-of-words model. In order to improve the performance of the classification, we propose a pattern-based approach which treats a text document as a sequence of words. The sequence model preserves the order of the words and allows for the analysis of such links.

We use the example of Twitter to implement our proposed approach in a Java program. Twitter is a popular Social Network allowing users to exchange messages directly or broadcast them to several receivers. We first develop an evaluation dataset to evaluate our method which is described in the next section.

## *Development of the datasets*

Since no dataset is available, we collected our own set of Twitter messages (Tweets) from the public stream between 2012-10-20 and 2012-12-30. The labeling process is accomplished by three annotators. A message is classified as Online Harassment if there is a consensus between at least two of the three annotators. Because the annotation of the messages is cumbersome, we classified the Tweets in their chronological order until a substantial amount of Online Harassment messages was found. For the annotation process we exclude non English, spam, empty and Re-Tweets (messages starting with “RT” or being completely enclosed by quotation marks). Re-Tweets are filtered to avoid duplicates and misclassification, because they forward a text written by another author.

The final dataset consists of 220 Online Harassment and 5162 neutral messages and is further denoted as main dataset. The sparse amount of 4.09% Online Harassment messages confirms findings of Kontostathis et al. (2013) and Sood et al. (2012) and thus represents a realistic proportion. A realistic

proportion is important, since neutral messages might contain profane words without expressing Online Harassment. Consequently, a lower amount of such messages might lead to a lower false positive rate of the classifier and thus distorting the performance measurements. However, we developed a second dataset with a similar amount of Online Harassment messages to provide an independent evaluation dataset. Since schools and colleges are our primary domain of interest, we collected the data by filtering the public stream data for tweets containing the words "school", "class", "college" and "campus". We labeled randomly selected tweets until a substantial amount of Online Harassment messages were found. We refer to the resulting data as school dataset which consists of 194 Online Harassment messages and 2599 neutral messages. We provide access to the datasets under the URL http://www.ub-web.de/research/index.html.

### *Word normalization*

The quality of user generated content in terms of correct speech varies within different Web 2.0 applications. While some applications like Wikipedia try to improve the text quality by offering an editing function to community members[2], other applications like Twitter do not allow a correction after a text is published. For the purpose of this work, we focus on the example of Twitter. Twitter limits the length of Tweets to 140 characters[3] which encourages users to use abbreviations and slang. The normalization process transforms such noisy words into their canonical forms. Wordlist and machine learning approaches benefit from this step because noisy profane words cannot be found in a dictionary unless all the noisy variations are stored as well. Explicitly storing all variations of profane words is laborious (Sood et al. 2012). In the same manner the person identification module benefits from normalized forms of personal pronouns as well.

To assess the relevance of a normalization step we investigate the main dataset in more detail. Our investigation is comprised by two steps. In a first step we determine whether a word is in its canonical form or an out-of-vocabulary (OOV) word. Every word is looked up in a reference vocabulary after removing common pre- and suffixes. If no match is found, the word is judged as OOV word (Jufarsky and Martin 2009). This decision is automated and part of the preprocessing step. However, the selection of an appropriate reference vocabulary is required first. Thus, we examine three common vocabularies regarding the resulting percentage of OOV words on our evaluation dataset. The results are summarized in Table 1.

| **Table 1. Vocabularies** | | |
|---|---|---|
| | Number of OOV words | Fraction of OOV words |
| Wordnet 3.0 | 19.915 | 37.6657% |
| Hunspell | 10.291 | 19.4636% |
| Moby project | 3.542 | 6.6991% |

**Table 1. Vocabularies**

Wordnet 3.0 (Fellbaum 1998) is popular among natural language processing applications (Jufarsky and Martin 2009) but is performing poorly compared to the other wordlists. Instead of covering a high fraction of all existing words, the main focus of Wordnet is to provide high quality syntactic and semantic information. The Hunspell[4] wordlist is used in machine translation tasks (Herrmann et al. 2011) and existing work regarding text normalization (Mosquera et al. 2012). The Moby project wordlist[5] comprises several publicly available vocabularies and is designed for applications that incorporate phonetic

[2] http://en.wikipedia.org/wiki/Wikipedia:About

[3] https://support.twitter.com/articles/15367-posting-a-tweet#

[4] http://hunspell.sourceforge.net/

[5] http://www.infochimps.com/collections/moby-project-word-lists

information. Since this wordlist performs best on the given dataset and thus reducing the normalization effort, we employ it as a basis for the investigation.

In a second step we manually count the profane words and words that refer to a person among the marked OOV words. We count 180 profane words and 323 words that relate to a person. Together they comprise 14.2% of the OOV words. Both word types are useful information for the classification of Online Harassment, which would be lost without a normalization step.

We implement a normalization module based on the method described by Mosquera et al. (2012). We selected this approach since it results in better normalization performance than current machine translation approaches (Mosquera et al. 2012). In a first step the module tries to match the OOV word against a slang and abbreviation dictionary which we built from noslang.com. If no match is found, the module tries to normalize the word by computing a simplified phonetic representation with the double metaphone algorithm. We then look for words that share the same phonetic representation within a prebuilt index based on the Moby project and the profane wordlist. To evaluate the module regarding our requirements we repeat the above-mentioned process. After the normalization step we count 28 remaining profane words and 25 remaining words that relate to a person. In addition we find 13 incorrect normalized profane words and 3 incorrect normalized words concerning a person. The normalization module successfully reduces the amount of these words to 1.95%. We further investigate the effect on the Online Harassment classification in the evaluation section.

### *Person identification*

Relations to persons can be stated explicitly by a name or implicitly by personal pronouns like "you". In case of Twitter, messages can also be broadcasted to several receivers addressed with their usernames. We identified four reference types that are summarized in Table 2 accompanied by an example of their usage in Online Harassment and neutral messages. The person identification module marks sequences of tokens that relate to a person and annotates them with the type of the reference. As a basis we use insights from the research field of Named Entity Recognition which includes the subtask of person identification (Jufarsky and Martin 2009). Named Entity Recognition treats a text as a sequence of tokens and searches for patterns that describe an entity including persons, locations and organizations. However, such tools are not suitable for our purposes. They are restricted to explicit (named) references and they cannot be extended to dynamically incorporate usernames. Thus, we implemented our own module that focuses on persons and can distinguish between the types summarized in Table 2.

**Table 2. Person references**

| | Online Harassment example | Neutral example |
|---|---|---|
| Implicit reference by personal pronoun | "Fuck you and your mom.." | "Im a super bitch today #watchout" |
| Implicit reference from the point of the authors view | "My new psych advisor is such an asshole" | "So my dumbass ended up dropping my phone in the locker room and now it has like 2 dents." |
| Explicit reference to a common name | "So gabby eat ass" | "@<anonymziedUser> Eric hahahahaa i fell like such a moron but i actually thought ya got kidnapped or somet" |
| Explicit reference to a user | "@<anonymziedUser> you asshole!" | "@<anonymziedUser> HAPPY BIRTHDAYYY!!!!!!!! ❤" |

**Table 2. Person references**

Implicit references can be detected by a list of personal pronouns. A disambiguation between references to others and to the author himself is important to avoid false positives. The person identification module marks tokens that contain a form of "I" as self-reference. While a form of "you" is an unambiguously reference to another person, a form of "we" represents a self-reference and a reference to one or more other persons.

An implicit relation from the point of the authors view can be detected by the possessive determiner "my" in combination with a list of nouns that relate to persons. Such relations are often used in the context of schools to refer to a certain teacher. As the example shows, they are not necessarily built upon a strict sequence consisting of "my" and a noun. Adjectives, conjunctions and additional nouns might be included to further specify the person. Thus, we apply an acceptor which is a special form of a finite state machine to detect such sequences with variable length. The acceptor consumes tokens until either the state "isPerson" or "error" is reached. The state "isPerson" is reached if a combination of "my" and a noun that describes a person is found. The acceptor ignores preceding clarifying this noun more in detail as mentioned above. Such an acceptor can be further configured to match a certain type of implicit referenced persons, i.e. teachers or classmates in context of schools.

Explicit references stated by names can be identified by their word type (noun) in combination with a list of common names. We use the one provided by the Moby project vocabulary. Such lists permit to determine if a recognized first name is typically female or male. This information could be used in fine grained rules as described by Dinakar et al. (2012). Twitter offers a comfortable way to detect references to users by a special token called Twitter Mention starting with the symbol "@" and followed by a username. The addressed user will receive the message in the Twitter network together with all the followers of the author. This kind of reference is the most direct way to address a person since Twitter user names are unambiguous.

## *Online Harassment Classifier*

The Online Harassment Classifier solves a binary classification task which separates Online Harassment documents from other documents. We propose a pattern-based approach which incorporates information from the preceding person identification step. To prevent overfitting we deduce just a small set of general applicable patterns from our dataset. However, since these links depend on the type of the profane phrase, we deduce a set of profane types first and specify the patterns that express the link in a second step.

We introduce an extended profane word lexicon first, which includes profane words or phrases, their POS-tags and a profane type. We use the wordlists provided by Noswearing.com (2014), Broadcasting Standards Authority (2013) and Hargrave (2000) as a reference. We deduced four profane types which are summarized in Table 3 by analyzing the evaluation dataset.

| **Table 3. Profane types** | |
|---|---|
| | Example |
| Profane noun | "@<anonymizedUser> cunt" |
| Profane property | "@<anonymizedUser> is dumb… #illuminati" |
| Profane verb | "@<anonymizedUser> SHUT YOUR FACE RIGHT NOW!!!!!-_-" |
| Profane imperative | "Your presence is making my life awkward… Die" |

**Table 3. Profane types**

Each type of profane word is used in different ways in a text sequence in general and in particular in the context of Online Harassment. These forms are tightly related to the patterns introduced in the next section. Except the profane imperative, the profane types can be determined by the POS tag.

Profane nouns are used in is-a-relations and exclamations representing the most common use case in the dataset. Due to the noisy language and incomplete sentence structure found in Tweets, such a relation might not be stated correctly in terms of grammar. However, the relation between a person and a noun is implicitly clear even without a form of "to be" as shown in the example. In contrast, a profane property requires a form of "to be" to establish a link between the word and a person.

Profane verbs are used to express actions consisting of a phrase that describes the action and a relation to a victim. Current wordlist-based approaches focus on single words restrained by their underlying bag-of-words model limiting their capability to detect phrases like in the example above. Often a verb is

ambiguous and is only considered harmful in certain phrases. We focus on profane phrases covered by our wordlist to prevent overfitting. However, an extension of the wordlist would improve the percentage of detected Online Harassment messages. Finally, we consider profane imperatives which are tightly related to verbs. They require another kind of links to persons to express a harmful statement. Thus, we introduce them as a separate profanity type. As shown in the example above the reference to the person is included in the preceding sentence statement while the imperative stands separately.

We identified the following patterns without the claim of completeness to express a connection between a profane type and a person by analyzing our dataset:

**Table 4. Proposed patterns**

| Pattern | Description | Example |
|---|---|---|
| n-direct reference before | Person reference is at most n (n=3) tokens off, adjective or determiner might be in between | "@<anonymizedUser> Y r U on Fast? U always wrong. plus u r an rude asshole." |
| n-is (a) | Person reference and a form of "to be" is at most n (n=3) tokens ahead, adjective, adverb or determiner might be in between | "@<anonymizedUser> is dumb… #illuminati" |
| n-direct reference after | Person reference is at most n (n=2) token behind, preposition might be in between | "fuck you Tim haha" |
| Subject predicate object | profane word is in between a self-reference and reference to someone else, a form of future tense might be in between | "I fucking hate you" |
| Unambiguous reference | there is a (potentially distant) person reference, but no neutral or self-references in the whole document | "People get unfollowed, don't take it personal. Chances are you were just a little too dumb, ugly, or complete fucking garbage." |
| n-locality of reference | there are person references and neutral or self-references, but the person reference is n (n=3) steps closer to the profane word | "@<anonymizedUser> @<anonymizedUser> you're being a whiny racist prick because you didn't get your way and im laughing at you … im bored now, thanks tho." |
| Separately standing exclamation | the profane word stands separately at the end or right after a sentence | "Your all compulsive liars. Everything Shit thing that happens somehow ends up my fault. and you wonder why I don't wanna be here. Cunts" |

**Table 4. Proposed patterns**

The configuration of the patterns contains several degrees of freedom. Profane phrases do not necessarily link a profane word and a person reference directly within a successive sequence. In the first example above the words "an" and "rude" are enclosed in a profane phrase. Thus, we introduce a distance (denoted with n) to permit certain words that are in between a profane word and a person reference. The specification of the distance and the set of allowed enclosed words specified by POS types is part of the configuration. We computed an optimal configuration for each pattern by evaluating them in an isolated manner. The resulting values are denoted in parentheses in Table 4. Further research is needed to confirm these settings on other datasets and exclude the possibility of overfitting.

The Online Harassment classifier is configured by providing a matrix which assigns each profane type a set of Online Harassment patterns. Whenever a profane word is detected in a sequence, the classifier tries to match at least one of the associated patterns at the corresponding position. If a pattern matches, the sequence is classified as Online Harassment. The composition of the matrix adds more degrees of freedom to the configuration of the proposed method. Thus, there are additional possibilities to configure the

proposed method by selecting which patterns are linked to which profane type. We discuss this configuration and its impact on the classification performance in the next section.

## Method and Evaluation

We discussed the proposed pattern-based approach in the previous section. This section is intended to assess the effectiveness of our proposed artifacts.

### *Method*

To evaluate our proposed approach we compare it against a naive wordlist classifier based on a bag-of-words model. This baseline classifier examines a text for the presence of at least one profane word. If such a word is found, the message is classified as Online Harassment. To further investigate the role of person references we extend this approach by our proposed person identification module. The extended naive classifier only judges a message as Online Harassment if at least one profane word and a person reference are found. Additionally, the normalization module is evaluated separately for each classifier.

We evaluate each classifier with the main and school dataset since they do not require training data. As evaluation metrics we compute recall, precision and f1 values as proposed in (Jufarsky and Martin 2009; Hirschmann and Mani 2003). Precision measures the proportion between correctly classified messages and all messages classified as Online Harassment. Recall measures the proportion between correctly classified messages and the number of real Online Harassment messages. The f1 value is the harmonic mean between precision and recall. Accuracy is not considered because of the sparse nature of Online Harassment. The dataset contains 4.09% Online Harassment messages, which means that by simply classifying every message as neutral, an accuracy value of 95.91% can be achieved.

### *Evaluation*

We omit the evaluation results for the class of neutral messages since the proportion of such messages is substantially high. Thus, the evaluation metrics will yield good results without added value for our conclusions. Table 5 shows the results of the evaluation.

**Table 5. Evaluation of Online Harassment classifier**

| | Main dataset | | | School dataset | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 |
| Naive | 37.47% | 64.55% | 47.41% | 45.85% | 65.46% | 53.93% |
| Normalization | 35% | 70% | 46.67% | 43.63% | 70.62% | 53.94% |
| Naive with person recognition | 63.82% | 57.73% | 60.62% | 73.29% | 60.83% | 66.48% |
| Normalization | **58.75%** | **64.09%** | **61.3%** | **70.05%** | **67.53%** | **68.77%** |
| Pattern-based (balanced setting) | 77.72% | 68.18% | 72.64% | 80.14% | 58.25% | 67.46% |
| Normalization | **73.45%** | **71.82%** | **72.64%** | **79.01%** | **65.98%** | **71.91%** |
| Pattern-based (precision setting) | 94.52% | 31.36% | 47.1% | 91.67% | 22.68% | 36.36% |
| Normalization | **94.74%** | **32.73%** | **48.65%** | **91.67%** | **22.68%** | **36.36%** |

**Table 5. Evaluation of Online Harassment classifier**

The normalization module improves the recall values for all examined classifiers while the precision is only reduced marginally. This confirms our preliminary examinations and the findings of Sood et al. (2012) who point out the noisy character of Web 2.0 texts. The classifiers are not able to match profane phrases which are not in their canonical form thus yielding lower recall values. Contrary to the three classifiers listed on top, the normalization step improves the results for the pattern-based classifier (precision setting) both in recall and precision.

The baseline classifier performs poorly on the datasets in terms of the achieved f1 value. However, while the precision value is very low, the baseline classifier achieves moderate recall values. Classifiers yielding substantial recall values are able to detect a large fraction of the Online Harassment messages. However, if the precision is low simultaneously, they also detect a large amount of false positives. The content of the profane wordlist influences this relationship and thus the performance of the classifier. Large wordlists lead to high recall values but many false positives as confirmed by the findings of Kontostathis et al. (2013). Their complete wordlist achieves high recall (78%) but low precision (44%) values. Subsets of this wordlist, which are optimized for their dataset, achieve high precision (84%) but low recall (37%) values. All investigated variations only achieve moderate f1 values between 28% and 57% (Kontostathis et al. 2013). Similar results are found by Sood et al. (2012). They accomplish precision values between 49% and 63% and recall values between 20% and 41%. Machine learning methods achieve slightly better classification results regarding the f1 values, which vary between 47% and 63% on their individual evaluation datasets (Kontostathis et al. 2013; Dinakar et al. 2012; Sood et al. 2011; Yin et al. 2009).

The person identification module improves substantially the precision value of the naive bag-of-word model while decreasing the recall marginally. This effect is more pronounced in the school dataset and might be caused by a larger amount of direct and unambiguous insults. With this naive setting alone, the classifier is able to achieve good results compared to existing work. Thus, the link between a profane word and a person reference is a relevant feature in Online Harassment detection. However, one supporting factor could be the limited length of a twitter message, which makes person references often unambiguous.

The pattern-based classifier requires the person identification module. Configured with the balanced setting, the classifier performs best with respect to the f1 value in both datasets. Similar performance measurements in both datasets indicate that the patterns are not specifically fitted to the main dataset. We achieve an improvement of 15% compared to existing wordlist-based approaches respectively 9% compared to machine learning approaches. The pattern-based approach allows affecting the recall and precision values without modifying the wordlist by adjusting its configuration. It can be configured by associating a subset of the available patterns to the types of profane words. We selected a combination of patterns for a balanced and a precision setting to demonstrate this effect. The precision setting achieves a precision value greater than 90% which makes it suitable for automatically blocking potential Online Harassment messages. However, the examination of the full space of configuration possibilities is computationally very expensive. Each configuration can be measured and captured in a diagram in terms of their resulting recall and precision values. More research is needed to compute all the Pareto efficient combinations in that manner.

## Practical applications within Social Networks

Aponte and Richards (2013) analyze different forms of Online Harassment and Cyber Bullying and suggest solutions for practical applications with the objective to prevent psychological harm. Such messages should be blocked or marked. This can be achieved by using a method as described in this work.

Our proposed method can be used to extend Social Networks enabling them to automatically detect and block Online Harassment messages in a proactive manner as stated by Patchin and Hinduja (2006). Such systems are able to analyze messages in real time causing negligible delay within the Social Network. However, since no human control instance verifies whether the message is blocked correctly, false positives can arise. If such messages are blocked the author might be frustrated at the Social Network. Thus, these approaches rely on a classifier with a high precision value. None of the approaches introduced in previous publications is capable of achieving substantial precision values. In contrast, our proposed approach can be configured to achieve precision values greater than 90% with the cost of low recall values around 20% to 30%. Anyhow, despite the low recall values a substantial fraction of Online Harassment messages could be blocked due to the vast amount of messages within Social Networks.

Systems that automatically block Online Harassment messages address the following problems. First, an Online Harassment message causes psychological damage once a victim reads it, regardless of the time it is visible afterwards. Hence, if the message is blocked before, no psychological damage can occur. In addition, Tokunaga (2010) suggests ignoring the author of an Online Harassment message as a coping strategy. Consequently, an immediately blocked message seems as it was ignored by the victim from the

point of the offenders view. Second, Online Harassment messages can be distributed easily within or between Social Networks by spreading them in a viral manner (Li 2007). Thus, it might be difficult for human control instances to cope with the amount of messages spread. However, automated systems can deal with such an amount of messages since they are scalable. Third, Online Harassment messages can be sent anonymously (Li 2007). Even if a victim blocks potentially offending accounts within a Social Network, he can only protect himself against anonymous accounts with a proactive system.

Systems that mark Online Harassment messages rely on a classifier with a high recall value. Our proposed method can be used to add this functionality to Social Networks or to external programs like parental control systems. Such systems reduce the effort for personnel to act as human control instance. They only need to consider marked messages and decide if it really is Online Harassment and whether further actions have to be taken. Thus, it is desired to cover a high fraction of potential Online Harassment messages while keeping the amount of false positives low. The proposed pattern-based classifier achieves high recall and precision values which makes it suitable for such tasks. Additionally, the system can be further improved when it is combined with a blended mechanism. Low precision values can be compensated by human interaction allowing the system to receive feedback regarding the classification and eventually learn from it.

The following problems are addressed by systems that mark Online Harassment messages. First, the vast amount of messages in Social Networks causes substantial effort for personnel to act as human control instance. However, several authors suggest introducing Social Network policies at schools (Sonhera et al. 2012; Li 2007; Patchin and Hinduja 2006; Campbell 2005). Consequently, to ensure these policies the messages among students need to be monitored by personnel (Sonhera et al. 2012). Our proposed approach can support this task by preselecting potential harassing messages. Second, a substantial fraction of Online Harassment victims does not inform their parents or other adults about these incidents (Tokunaga 2010; Li 2007). While some of these victims avoid involving an adult because they want to cope with the situation themselves or fear restrictions regarding their access to Social Networks, others are just overwhelmed by the situation (Tokunaga 2010). Parental control systems allow parents to be aware of these incidents by monitoring the accounts of their children so they can decide if an intervention is necessary. Third, the barrier to communicate in an offending way is lower in Social Networks compared to face-to-face communication (Tokunaga 2010). Furthermore, offenders might not be aware of the harm they cause with their messages or they might overreact caused by a recent event. The system could notify the authors before publishing potential Online Harassment messages. Psychological harm can be prevented, if the author reconsiders the publication.

Besides the applications mentioned above, the investigated methods can be used to improve Cyber Bullying detection. Cyber Bullying is based on Online Harassment messages that are sent repeatedly to the same victim by the same author. The person recognition module can be extended to detect such relations. Even multiple Social Networks could be analyzed to track cyber bullies using different communication channels as proposed by Dadvar and de Jong (2012). Furthermore, retrieving training data for Online Harassment classification is a challenging task due to the sparseness of such messages. Systems with high recall can help to preselect Online Harassment candidates for a subsequent labeling step. The data can also be used within other research fields like psychology as described by Xu et al. (2012).

Finally, aspects regarding the freedom of speech need to be considered when using either of the systems. While monitoring public channels belonging to schools seems adequate, monitoring of private accounts especially without their knowledge might interfere with the right of privacy. Particularly parental control systems require the approval of the account being monitored. Instead of enforcing the integration of such systems, parents could try to achieve a consensus with their children about using such systems as a safety mechanism allowing the parents to intervene if necessary. In addition, the difference between Online Harassment and harsh criticism might be subtle. Systems that block messages containing criticism regarding these entities restrict the freedom of speech. In this case the prevention of psychological harm and the preservation of freedom of speech are conflicting goals. However, it needs to be distinguished between individuals along with children in particular and public figures, companies, governments and other referenced entities. Consequently, the protection of individuals might deserve a more vigorous consideration than the preservation of unrestricted message exchange, especially if a blocked message can be rephrased. Our proposed approach is capable of achieving this goal by performing a fine grained

person identification excluding entities not related to individuals. Furthermore, the person identification can be used to protect only certain accounts like those from children.

## Limitations

The proposed approach cannot measure whether a message really causes psychological harm to a person. Additionally, no quantification of the severity of a profane phrase is applied. Further work could assign weights to the profane phrases within the dictionary by common consensus about which phrase is considered more or less harmful. In this work we assume that the severity of the damage cannot be quantified appropriately since it is based on the individual's perception and its current context (i.e. current mood, relation to sender, visibility of the message) (Tokunaga 2010; Patchin and Hinduja 2006).

Since we assume that only messages containing direct references to persons are considered Online Harassment, we cannot correctly classify messages that refer to a group of size n. However, even for human annotators it is hard to decide whether the group is small enough, so the message can be judged as inflicting psychological harm to the individuals.

The normalization module relies on language data containing common slang and abbreviations. Some abbreviations are context dependent and can lead to misclassification, i.e. "af" could mean "as fuck" or "autofocus" in the context of cameras. If several word candidates exist, a context dependent decision is necessary.

Our approach enables the classifier to match profane phrases instead of matching just single words. However, profane phrases consisting of several words can be stated in various ways. A full enumeration of all possible combinations is laborious and results in large lexica. Our lexicon contains only a small number of such phrases and thus could be extended to further improve the percentage of detected Online Harassment messages. Finally, sarcasm is a general problem in text mining and especially in Sentiment Analysis (Tsytsaru and Palpanas 2012; Pang and Lee 2008). Online Harassment can be veiled by sarcastic formulations or metaphors. The proposed method cannot detect such figures of speech.

## Conclusion

Online Harassment is the process of sending messages over electronic media to cause psychological harm to a victim (Tokunaga 2010). In this paper, we have presented a pattern-based approach to detect Online Harassment in Social Networks and discussed its practical applications based on the suggestions of Aponte and Richards (2013).

Such systems should be able to block or mark Online Harassment messages. The pattern-based approach is suitable to realize these use cases by adapting its configuration. Due to the vast amount of messages within a Social Network and the sparse nature of Online Harassment messages, a manual classification is laborious. A balanced configuration of our proposed approach is able to mark potential Online Harassment messages. It achieves f1 values of around 72% which exceeds existing wordlist-based and machine learning approaches by 15% respectively 9%. It further helps to reduce the amount of work for a human control instance which can draw a decision afterwards and might initiate further actions. However, since such actions are reactive in their nature, harm still occurs to the victim if he reads the message. A high precision setting can help to prevent such harm by blocking messages that are very likely Online Harassment. Our approach achieves precision values greater than 90% which outperforms existing approaches by 30%. A high precision value reduces the number of false positives and makes the classifier more suitable for practical applications in Social Networks.

Despite the associated low recall value, a large amount of Online Harassment messages can be blocked among the vast total amount of messages within Social Networks. Previous research focuses on classifiers which are based on bag-of-words models. These approaches primarily analyze text documents regarding the presence of profane words. We use a sequence-based model that preserves the order of words in a document. Since Online Harassment targets at a person we further introduce a person identification module which marks words or phrases referring to persons within this sequence. Our proposed pattern-based approach incorporates information of this step to find links between a detected profane phrase and the addressed person. Such links are expressed by typical patterns we deduced from our dataset. This way

we are able to improve substantially the classification performance regarding the combined measure of precision and recall.

Because of the lack of datasets we provide two sets of manually annotated messages of the Social Broadcast Network Twitter. The labeling process is accomplished by three annotators since it is even for humans considerably hard to decide whether a message is classified as Online Harassment. The datasets are available at this URL[6] and can be used to evaluate other approaches. The analysis of the main dataset reveals that the language used within the messages is noisy. Noisy language contains spelling mistakes, abbreviations and slang. Thus, we extend our approach by a normalization module which transforms noisy text into its canonical form. We found that the module improves the performance of any of the investigated classifiers regarding their achieved recall values.

Future work could determine optimal configuration settings for the pattern-based approach to further improve the classification results. Several Pareto efficient combinations regarding the achieved recall and precision values are possible. The trade-off between recall and precision influences the practical applications of the classifier. Moreover, our deduced patterns need to be confirmed by further research. Furthermore, our proposed method can be extended to detect Cyber Bullying. Cyber Bullying detection is user centered and requires the identification of the bully and the victim across several messages. The proposed person identification module already provides a piece of this information. We also excluded Re-Tweets from our investigation. By incorporating these messages the original author and the users that support him by spreading the message could be identified.

## References

Aponte, D. F. G. and Richards, D. 2013. "Managing Cyber-bullying in Online Educational Virtual Worlds," in *Proceedings of The 9th Australasian Conference on Interactive Entertainment: Matters of Life and Death*, Edinburgh, Melbourne, Australia, pp. 18:1–18:9.

BBC News. 2014. *Cyberbullying suicide: Italy shocked by Amnesia Ask.fm case*. http://www.bbc.com/news/world-europe-26151425. Last accessed 22/04/2014.

Broadcasting Standards Authority. 2013. *What not to swear: The acceptability of words in broadcasting*. http://bsa.govt.nz/images/assets/Research/Acceptability_of_Words_2013_WEB.pdf. Last accessed 22/04/2013.

Campbell, M. A. 2005. "Cyber Bullying: An Old Problem in a New Guise?," *Australian Journal of Guidance and Counselling* (15:1), pp. 68-76.

Dadvar, M. and de Jong, F. 2012. "Cyberbullying Detection: A Step Toward a Safer Internet Yard," in *Proceedings of the 21st International Conference Companion on World Wide Web*, Lyon, France, pp. 121–126.

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., and Picard, R. 2012. "Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying," *ACM Transactions on Interactive Intelligent Systems (TiiS)* (2:3), pp. 18:1-18:30.

Dinakar, K., Reichart, R., and Lieberman, H. 2011. "Modeling the Detection of Textual Cyberbullying," in *Proceedings of the International Conference on Weblog and Social Media (Social Mobile Web Workshop)*.

Easley, D., and Kleinberg, J. 2010. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*, New York, NY, USA: Cambridge University Press.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database,* Cambridge, MA: MIT Press.

Hargrave, A. M. 2000. *Delete Expletives?: Research Undertaken Jointly by the Advertising Standards Authority, British Broadcasting Corporation, Broadcasting Standards Commission and the Independent Television Commission,* London, UK: Advertising Standards Authority.

Herrmann, T., Mohammed, M., Niehues, J., Waibel, A. 2011. "The Karlsruhe Institute of Technology Translation Systems for the WMT 2011," in *Proceedings of the 6th Workshop on Statistical Machine Translation*, Edinburgh, Scotland, UK, pp. 379–385.

Hirschman, L., and Mani, I. 2003. "Evaluation," in *The Oxford Handbook of Computational Linguistics*, Ruslan Mitkov (ed.), Oxford University Press, Oxford, pp. 414-429.

[6] http://www.ub-web.de/research/index.html

Jurafsky, D., and Martin, J. H. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition,* Upper Saddle River, NJ: Prentice Hall PTR.

Kontostathis, A., Reynolds, K., Garron, A., and Edwards, L. 2013. "Detecting Cyberbullying: Query Terms and Techniques," in *Proceedings of the 5th Annual ACM Web Science Conference*, Paris, France, pp. 195–204.

Li, Q. 2007. "New Bottle but Old Wine: A Research of Cyberbullying in Schools," *Computers in Human Behaviour* (23:4), pp. 1777-1791.

Mosquera, A., Lloret, E., and Moreda, P. 2012. "Towards Facilitating the Accessibility of Web 2.0 Texts through Text Normalisation," in *Proceedings of the LREC workshop: Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, Istanbul, Turkey, pp. 9–14.

Noswearing.com. 2014. *Bad Word List & Swear Filter*. http://www.noswearing.com. Last accessed 22/04/2014.

Pang, B., and Lee, L. 2008. "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval* (2:1-2), pp. 1-135.

Patchin, J. W., and Hinduja, S. 2006. "Bullies Move Beyond the Schoolyard: A Preliminary Look at Cyberbullying," *Youth Violence and Juvenile Justice* (4:2), pp. 148-169.

Sonhera, N., Kritzinger, E., and Loock, M. 2012. "A proposed cyber threat incident handling framework for schools in South Africa," in *SAICSIT Conf.*, Centurion, South Africa, pp. 374-383.

Sood, S. O., Churchill, E. F., and Antin, J. 2012. "Automatic Identification of Personal Insults on Social News Sites," *Journal of the American Society for Information Science and Technology* (63:2), pp. 270-285.

Tokunaga, R. S. 2010. "Review: Following You Home from School: A Critical Review and Synthesis of Research on Cyberbullying Victimization," *Computers in Human Behaviour* (26:3), pp. 277-287.

Tsytsarau, M., and Palpanas, T. 2012. "Survey on mining subjective data on the web," *Data Mining and Knowledge Discovery* (24:3), pp. 478-514.

Xu, J., Jun, K.-S., Zhu, X., and Bellmore, A. 2012. "Learning from Bullying Traces in Social Media," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montrèal, Canada, pp. 656-666.

Yin, D., Xue, Z., Hong, L., Davison, B., Kontostathis, A., and Edwards, L. 2009. "Detection of Harassment on Web 2.0.," in *Proceedings of the Content Analysis in the WEB 2.0 (CAW2.0) Workshop at WWW2009*, Madrid, Spain.