Detecting Offensive Statements towards Foreigners in Social Media

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

Abstract

Recently, politicians and media companies identified an increasing number of offensive statements directed against foreigners and refugees in Europe. In Germany, for example, the political group "Pegida" drew international attention by frequently publishing offensive content concerning the religion of Islam. As a consequence, the German government and the social network Facebook cooperate to address this problem by creating a task force to manually detect offensive statements towards refugees and foreigners. In this work, we propose an approach to automatically detect such statements aiding personnel in this labor-intensive task. In contrast to existing work, we assess severity values to offensive statements and identify the referenced targets. This way, we are able to selectively detect hostility towards foreigners. To evaluate our approach, we develop a dataset containing offensive statements including their target. As a result, a substantial amount of offensive statements and a moderate amount of the referenced victims was detected correctly.

1. Introduction

The ongoing civil war in Iraq and Syria and its consequences are dominantly present in the media. The crisis led to the displacement of millions of refugees that are forced to search asylum in other countries. For example, Germany alone expected around one million asylum-seekers in 2015 [1]. These large numbers of refugees arriving in Europe caused controversial political discussions about the feasibility and consequences of their accommodation [2]. In this context, social media is growing in popularity to organize political discussions, exchange opinions and form groups of mutual interest [3,4]. In recent years, social media platforms have recorded a substantial increase in user numbers. Facebook, for example, has over 1 billion daily active users [5]. New content rapidly spreads in social media networks reaching a large amount of users [6] and enabling similar minded people to easily find and connect with each other [7].

Besides people having sympathy for the critical situation of the refugees, there are also people sharing a negative view. In extreme cases, they direct offensive statements towards refugees or foreigners in general expressing their fear and aggression [2]. In Germany, for example, the political group "Pegida" drew international attention by frequently publishing new content containing offensive statements towards foreigners, especially towards followers of Islam [4]. This form of offensive language is often referred to as cyberhate or hate speech, which is a general problem in social media [8,9].

Recently, German politicians recognized hostility towards foreigners in social media as a growing problem since it might facilitate public incitement against foreigners. Moreover, radical groups and political parties might take advantage of the recent situation spreading their ideology and eventually recruiting new supporters [9,10]. Social media platforms intensify this problem by the possibility to anonymously create content rapidly reaching a large number of users [6]. More importantly, content containing one-sided and radical viewpoints might be a problem in political opinion-formation, if users have only restricted access to credible opposing opinions [11,12]. This way, an important concept of democracy is violated: taking informed decisions in the context of competing opinions and ideas [13].

In a current project, the social network Facebook cooperates with the German government to address this problem by introducing an action plan. The plan contains a task force consisting of people from online communities, political parties and the German justice ministry to detect offensive statements towards refugees and foreigners [1]. However, due to the vast amount of messages in social media, the task of detecting hate speech is labor-intensive and time-consuming [3,14]. Additionally, there is only a limited amount of automated approaches that are able to detect hate speech directed against a certain target [15]. These approaches are not effective as hate speech towards foreigners is

URI: http://hdl.handle.net/10125/41423
ISBN: 978-0-9981331-0-2

often paraphrased and complex [9]. As a result, they are not capable of detecting the target of hate speech. This is, however, important to distinguish hate speech without certain targets from hate speech directed towards certain people or groups.

We extend current research on the detection of hate speech by the following contributions. First, we present an approach to detect hate speech towards foreigners in social media including the referenced target. Second, we develop an annotated dataset to assess the performance of our approach as there are no reference datasets yet. We provide access to this dataset as a benchmark for further research. Third, we discuss applications of our approach and strategies to tackle the problem of hate speech towards foreigners and refugees.

The rest of this paper is organized as follows: Section 2 contains the theoretical background of this study including a discussion of freedom of speech versus hate speech in the context of social media and an overview of exisitng work in hate speech detection including its related forms. In section 3, the development of the annotated datasets containing user comments from public Facebook pages is presented. In section 4, the proposed approach is introduced in detail. An evaluation based on the annotated datasets is presented in chapter 5. Section 6 discusses practical applications in social media platforms. Finally, section 7 summarizes the results and points out aspects for further research.

2. Theoretical background

2.1. Freedom of speech versus hate speech in social media

Freedom of expression, especially freedom of speech, is regarded as a fundamental individual right anchored in the Universal Declaration of Human Rights of the United Nations that is ratified by the majority of the countries in the world [16]. In the legally binding instrument of this declaration, the "International Covenant on Civil and Political Rights" (ICCPR), freedom of speech is defined as the right to "receive and impart information and ideas of all kinds" [17].

Article 19 (3) of the ICCPR defines restrictions to freedom of speech as it might conflict with "the rights or reputations of others" or "the protection of national security or public order […], or of public health or morals" [17]. The interpretation of the exceptions stated in article 19 (3) ICCPR as well as their implementation in national law is different from country to country [18]. China, for example, applies a very restrictive interpretation in terms of national security and system critic opinions [19]. In democracies, freedom of speech is regarded as a fundamental right and core concept [19,20]. In the United States, for example, freedom of speech is anchored in the first Amendment [19]. A liberal and self-regulating approach is applied based on the principle that ideas contest each other in a marketplace of competing ideas [13,19].

In this work, we follow the interpretation of freedom of speech from the European Union. In contrast to the United States, the European Union is more restrictive, especially with respect to hate speech [20]. In line with current research [9,14], the Council of Europe's Committee of Ministers notes that no universally accepted definition of hate speech exists [21]. As an orientation for European case law, they state that hate speech "covers all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance." [21]. Violations concerning the publication of hate speech might lead to legal consequences primarily for the author of offensive content [20]. As a recent decision from the European Court of Human Rights shows, the social media platform might be held responsible as well [22].

As a consequence, a conflict between the protection of the victims, the social media platform and the fundamental right of freedom of speech exists. The primary focus of this work is to propose an approach to detect hate speech and their referenced targets that might be used in different ways to comply with national rights. In section 6 we discuss these ways in form of practical applications and their potential consequences on freedom of speech.

As stated above, freedom of speech is an important element of democracy fostering political discussions of competing opinions and ideas [13]. However, using hate speech in political discourse to prevail extremist viewpoints might deter other users wishing to participate in a civil discussion [3]. Consequently, users expecting civil discussions often favor moderation to restrict uncivilized behavior by removing messages that do not conform with community norms [3]. In the context of social media, moderation is a labor-intensive task that causes financial costs [3,14]. In addition, coping with uncivilized behavior causes emotional costs both for moderators and participants with a civil but potentially opposing opinion. Based on the theory of Hochschild's "emotion work", such users need to perform "deep acting" to adjust their inner emotions to match the expectations on emotions required in a civil discussion [23]. While this theory originates from face-to-face communication [23], other researches apply it to the digital context. Menking and Erickson [24], for example, found that women avoid engaging in the Wikipedia as it requires them to perform "deep acting" to cope with harassment.

Another obstacle for discourse are echo chambers, a phenomenon first described by Key [25] in the political context. In social media networks, they might facilitate homogenous viewpoints by superseding opposing viewpoints [11]. Within such a network, users create mutual connections, for example by friendship or follower relations as well as by forming groups. The content displayed to a user often depends on these relations, for example, Facebook's EdgeRank filters by analyzing such relations [26]. Content published by friends or connected groups is more likely to be displayed than content from other users. More importantly, the resulting content often contains one-sided viewpoints as friends typically share similar interests and opinions [11]. In the context of political opinion-formation, echo chambers might be a problem if users are exposed to homogenous opinions favoring an extreme political viewpoint while having restricted access to credible opposing opinions [11,12].

Detecting and automatically resolving these obstacles characterized by hate speech might help administrators to moderate discussion eventually fostering civil discourse. Furthermore, the problem of echo chambers containing mostly hate speech and homogenous viewpoints might be addressed as stated in section 6.

2.2. Approaches to detect hate speech

Hate speech, cyberhate and offensive language are umbrella terms often used in the context of social media to denote offending content in general [9,14]. Hostility towards foreigners is, in particular, characterized by a referenced victim similar to the related form of online harassment. Tokunaga [27] defines online harassment as the process of sending messages over electronic media to cause psychological harm to a victim [27]. Thus, we consider existing approaches in the research fields of hate speech as well as online harassment detection. As we are interested in applying an approach to exclusively detect hate speech towards foreigners including the referenced target, we discuss their strengths and weaknesses in this regard. The related approaches are subsumed in table 1.

Table 1. Existing approaches

	[14]	[28]	[29]	[30]	[31]	[15]
Hate speech	X	X	-	-	X	-
Online Harassment	-	-	X	X	-	X
Referenced victim	X	-	-	X	-	X
Victim identification	-	-	-	-	-	-

The majority of the publications apply either lexicon [28,31] or machine learning approaches [14,29,30]. Lexicon approaches entirely rely on a lexicon containing offensive words typically used in hate speech. In their basic form, they classify a text as hate speech, if it contains at least one offensive word. A major advantage of these approaches is their simplicity and independence of training data as well as easy adoption in other languages by providing adequate lexica by experts. However, their practical applicability is limited, especially in the context of online harassment detection as they achieve only reasonable to moderate classification performance [28]. As a consequence, they are often used to preselect potential offending messages to perform subsequent analyses [31].

Machine learning approaches, in contrast, rely on training data to automatically learn rules to classify hate speech messages. As these rules are derived from statistical relationships, they require numerical inputs in form of features. These features are derived by experts from characteristics of hate speech messages and include, for example, the presence of offending words defined in a lexicon [14,30] and the presence of words typically referring to persons [30]. Compared to lexicon approaches, the classification performance is only slightly better [28,29,31]. Additionally, the collection of an adequate amount of training data is cumbersome due to the lack of annotated datasets [28,31].

All of the above-mentioned approaches rely on bag-of-words models representing a text as a vector of words. As a consequence of these simple models, the order of the words and thus their context is lost. However, the context of the offending passage is important to detect links between offending words and the targeted victims. These approaches are neither capable of detecting such links nor of detecting the passage with the referenced victim. As a consequence, they only achieve moderate classification results in online harassment classification as this form is characterized by containing a link to a victim [14,15].

Chen et al. [14] introduce a refined machine learning approach to address these shortcomings. They note that strong offensive words often occur in unambiguous hate speech messages while weak offensive words are only considered offensive when they are directed against a person. As a consequence, they apply a lexicon distinguishing between strong and weak offensive words. They compute the dependency graph of a given text to analyze its grammatical relations eventually detecting links between offending words and persons. In contrast to bag-of-words models, the dependency graph is a complex text model representing sentences of a text as sets of grammatical relations [14]. The ability to process such relations is the main advantage of the underlying text model. However, the model is designed for short texts that are treated as a single sentence to capture their whole context possibly resulting in incorrect grammatical relations [14]. Moreover, the approach requires a dependency parser for each language and dismisses the detected victim references as they are not required for further processing.

Xu et al. [30] apply a sequence label task in addition to a machine learning approach to identify online harassment cases including involved roles. First, the machine learning approach is used to detect online harassment. In a second step, role labeling is applied to assign the author of the message, the victim and additional roles. They achieve reasonable results for the identification of the offender. However, the performance for assigning the other roles mentioned within the text, especially the victim, is moderate [30]. Furthermore, an additional training data set is required to perform the sequence label task [30].

More recently, Bretschneider et al. [15] proposed a pattern-based approach to detect offending passages in text messages including the referenced victim. Instead of a bag-of-words model, they apply a sequence model that preserves the order of the words. In contrast to the dependency graph in [14], the sequence model is not restricted in length and easier to compute [15]. Compared to the other approaches that exclusively detect online harassment, they achieve substantially improved classification results by employing patterns that represent typical ways to link offending passages to persons [15]. Similar to the grammatical relations in [14], these patterns need to be defined by experts.

Even though the approaches presented in [15] and [30] are capable of detecting referenced victims, none of the existing approaches further process them to actually identify the victim. Moreover, while online harassment messages are directed towards a person, xenophobic or racist content is typically directed towards groups of people, nationalities or races. Currently, there is only limited amount of work available that addresses the detection of xenophobic or racist content in social media including the referenced victims. The sheer detection of passages referencing a victim is not sufficient to unambiguously identify the target. Often, the offender refers to people by using indirect references that need to be resolved first [15]. In this work, we extend existing approaches to detect text passages containing hostility towards foreigners and identify the referenced target.

3. Construction of the dataset

We constructed three datasets by accessing publicly available Facebook pages, to evaluate our proposed approach and to acquire training data. We crawled Facebook posts including the comments published in response to them. The two popular Facebook pages "Pegida" (dataset 1) and "Ich bin Patriot, aber kein Nazi" ("I'm a patriot, not a nazi") (dataset 2) were selected as they are known for their critical view regarding foreigners and refugees [4] and thus presumably contain offensive statements. In addition, we select the page "Kriminelle Ausländer raus" ("Criminal foreigners get out") (dataset 3) as a training dataset since it is known for xenophobe and racist comments. We crawled the latest 50 posts including their comments beginning from February 2016. We only included 20 posts for dataset 1 to acquire a comparable amount of comments for dataset 1 and dataset 2. Two human experts annotated the datasets marking offensive statements, their severity and the intended target. To the best of our knowledge, there are not yet any reference datasets containing this information.

Each offending passage is marked and assessed with a severity value. Statements that are perceived by the experts as slightly offensive to offensive are denoted with a severity value of 1 and explicit to substantial offensive statements with a value of 2. The severity value is applied in different evaluation scenarios and practical applications described in the method section and section 6 respectively. Additionally, we leverage this information in the training dataset to derive severity values for the offending words in our lexicon.

We employ Cohen's Kappa to measure the inter-rater agreement for offensive statement annotation. The assessed severity value is used as class label. Since the class distribution between offending and neutral messages is substantially skewed in favor of neutral messages, the resulting kappa value would overestimate the agreement. Consequently, we compute a kappa value only considering offending messages marked by at least one annotator. The results indicate a substantial agreement and are denoted in table 2 along with other descriptive metrics of the datasets.

Table 2. Constructed datasets

Dataset	**1**	**2**	**3**
#comments	2649	2641	546
#cases (severity = 1)	99	112	50
#cases (severity = 2)	137	112	130
Cohens Kappa	0.78	0.68	0.73
Target Foreigner	**24.38%**	**37.95%**	**76.67%**
Target Government	33.88%	33.04%	3.89%
Target Press	17.36%	8.04%	2.22%
Target Community	3.72%	4.91%	6.67%
Target Other	16.12%	14.29%	8.89%
Target Unknown	5.37%	1.79%	1.67%

Furthermore, the annotators identified the referenced target. We focus on offending statements directed towards foreigners and refugees and find evidence that a substantial amount of these statements is indeed directed towards foreigners, especially in dataset 3. However, the coding process revealed that frequently other related entities are referenced, for example the German government. As a consequence, we derive 6 target groups frequently referenced in the datasets: foreigners and refugees, the government represented by political parties and politicians, the community of the Facebook group, the press and media, other identifiable targets and unknown targets. Unknown targets arise if the human annotators are not able to resolve the reference.

A consensus annotation is computed by merging the annotations from both annotators. Severity values are combined by computing the average and rounding down. For example, a severity value pair of 1 and 2 results in a consensus severity value of 1. For the assessed targets, we only consider targets marked by both annotators. If there is no consensus, we classify the target as unknown. We anonymized the dataset by employing a hash function on each username for the purpose of the publication. We provide access to the datasets under the URL www.ub-web.de/research/.

4. Proposed method

4.1. System architecture

In this work, we propose the system architecture depicted in figure 1. The architecture is based on elements employed without modification as described in [15], which are denoted in the dotted line box.

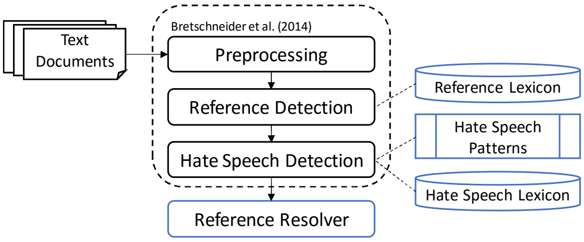

Figure 1. System architecture

Our decision to select this particular approach is primarily justified by the requirement to detect and identify the referenced victims. Only the approaches described in [15] and [30] are capable of accessing the passage including the referenced victim. However, the classification results achieved in [30] are moderate, while the results from [15] are more promising. Compared to the dependency graph in [14], the underlying sequence model in [15] is suitable for longer texts and does not require a dependency parser for the German language.

In a first step of the resulting architecture, the text documents are preprocessed by decomposing the unstructured text into tokens. In addition, these tokens are normalized removing common abbreviations and slang. In contrast to bag-of-words models, a sequence model is applied preserving the order and thus the context of the words. In a second step, the reference identification module marks tokens in the sequence referring to entities of interest for this study. These entities are, for example, foreign nationalities, political groups and the government.

After these preprocessing steps, the hate speech detection module searches for offending words in the sequence. Once such a word is found, the hate speech patterns are applied searching relations between the offending word and a reference to a victim. If a pattern matches, the text is classified as hate speech directed towards a victim. Finally, we identify the victims referenced in these passages by performing a reference resolution. As a consequence of this architecture containing consecutive tasks, the reference resolver can only process cases that are correctly detected in the previous step. Thus, we are interested in detecting a preferably complete amount of offending statements without the cost of too many classification mistakes in the form of false positives. To achieve this goal, we follow the proposals presented in [15] and [14]. In line with Chen et al. [14], we distinguish two forms of offensive statements: severe offending statements not necessarily containing a referenced victim and offending statements directed against a target. While only focusing on the latter has the advantage of a low false positive rate, it also comes with the disadvantage of a lower detection rate [15].

Finally, the original method described in [15] is designed for text documents in English. As our dataset contains text documents in German, we modify the approach accordingly by creating a reference lexicon, a hate speech lexicon and hate speech patterns as described in the subsequent sections. These modifications are required for each language.

4.2. Reference detection

The reference detection is a preprocessing step that marks references to entities of interest that are further processed in subsequent steps. We distinguish between static and dynamic references that are both stored in a dynamic lexicon. Static references are expressed by common words found in appropriate lexica and are further classified into direct and indirect references. Experts need to define this part of the lexicon for each language manually.

Direct references refer among others to nations or religious groups. For example, the sentence "sieht wien scheiß kosovoalbaner aus” (“looks like a damn Kosovo-Albanian”) taken from the dataset contains the direct reference “kosovoalbaner” referring to the ethnical group of Kosovo-Albanians. In contrast, indirect references are often used as a shorthand for direct references or to paraphrase a reference to a victim that is apparent in the context. As an example, the sentence “Dieses Ratten Pack bringt nur unruhen” (“This rat rabble only brings unrest”) contains an indirect reference consisting of an article in combination with a word typically referring to a group of people (“pack”). In this case, the reference points at refugees in general and can be resolved by analyzing the corresponding Facebook post, which contains a short story about refugees. We employ the German dictonary “Duden” as a lexical resource to define such static references, especially by using the synonym functionality.

Finally, dynamic references are based on special terms and names that relate to the current political context and characteristics of the social media platform. Usernames, for example, are often unique identifiers in social media platforms to refer to each other. Publicly known names, for example the current German chancellor “Angela Merkel”, are often subject of political discussions. Political groups, for example “Pegida” arise and dissolve over time. To account for such dynamic terms, we build a dynamic database by employing expert knowledge. In further work, such information might be derived automatically, for example from knowledge databases like DBpedia.

For each reference we additionally store the corresponding group as defined in the previous section. For example, the chancellor “Angela Merkel” belongs to the government group. In further work, an ontology might be applied instead.

4.3. Offensive statement detection

As our dataset contains text documents in German, we need to modify the approach from Bretschneider et al. [15] accordingly by employing a German offending word lexicon [32] and creating new hate speech patterns tailored for the German language. Our resulting patterns are listed in table 3.

Table 3. Constructed hate speech patterns

Pattern	Example
Reference before	“**Dieses Ratten Pack** bringt nur Unruhen” („This rat rabble only brings unrest“)
Is-a-expression	“**Fluechtlinge sind Parasiten**!” („Refugees are parasites“)
Reference after	“**Scheiß Pegida**” („Shit Pegida“)
Isolated expression	“Achtkantig rausschmeißen**, die Penner**” („Throw these hobos out on their ears“)
Compound	“Raus mit dem **Antifapack**” („Out with this anti-facist-rabble”)
Explicit sentence	„Eben echte **Arschlöcher**“ („Simply real assholes“)
Physical violence	„Schwanz **abhacken**“ („Cut the dick off“)

As described in the data section, we use a separate dataset to develop the patterns to prevent overfitting. In line with [15], we derive general speech patterns expressing several ways to relate offending words to entities. We were able to adapt four of the seven harassment patterns to the German language with minor modifications accounting for possible intermediate tokens between the offending words and the detected reference. As an example, the “is-a-expression” pattern is depicted in figure 2, relating an offending word to an entity reference by a form of “to be”.

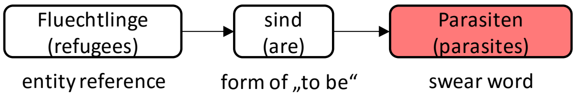

Figure 2. Is-a-expression pattern

In addition, we introduce the compound pattern. The German language allows to use compound words, for example by composing two nouns into a single word. The compound pattern relies on a preprocessing step that splits such compound words into its atomic components. If a combination of offending words and reference is found, the pattern will match. The physical violence pattern searches for combinations of words expressing physical violence towards human beings or parts of the human body. In line with [14], we additionally introduce the explicit sentence pattern. This pattern matches, if a sentence contains severe offensive words typically exclusively referring to persons, regardless of detected references. We distinguish severe offensive words from others by assessing a property to them in our lexicon. As the annotators marked severe offensive statements, we are able to identify the corresponding offensive words by analyzing our training dataset.

4.4. Reference resolver

The reference resolver identifies victims that are addressed in offensive statements detected in the previous step and maps them to one of the groups that are described in section 3. We propose four strategies to resolve such references.

First, if a pattern matches that already contains a direct reference, we directly process this reference and retrieve the corresponding group from the lexicon. Second, if the reference is indirect, we search for a direct reference in the context of the matching offending passage. If a direct and unambiguous reference can be found, it is resolved accordingly. For the case of detected ambiguous direct references, the closest one is chosen.

Typically, an article or post is the subject of discussion in the context of social media. Users eventually refer to this subject by publishing comments containing indirect references. As a third strategy, we try to analyze the content of the corresponding article searching for direct references using our reference detection module. If such references are found, we resolve them accordingly. Finally, we analyze all comments that are responses to the current article and compute the number of occurrences of direct references ordered by the corresponding group. In this strategy, we assume that comments containing only indirect references typically refer to the same subject that most of the other comments also refer to.

5. Method and evaluation

5.1. Method

We implemented our approach to detect offending statements towards foreigners in two consecutive steps. First, we performed a binary classification task identifying offensive statements. To assess the performance of this task, we computed the evaluation metrics precision (p), recall (r) and f1 as recommended in [33]. Second, we performed a binary classification task assigning each detected offensive statement to the classes offensive (severity = 1) or severely offensive (severity = 2). As the classifier can only process cases that are correctly detected in the first step (true positives), we computed the evaluation metrics without accounting for errors in the first step as we are interested in the performance considering the aspect of practical applicability of the system as discussed in section 6. Finally, we performed a multi class classification task assigning the identified victims to the classes we described earlier in section 3. In particular, we are interested in the performance of the approach to detect offensive statements directed towards foreigners. In analogy to the severity classification, we computed evaluation metrics without considering errors in the previous step. Finally, we implemented a baseline classifier to compare our evaluation results. The baseline classifier consists of a machine learning approach based on a bag-of-words model as described in [14]. We used the software “Rapid Miner” to evaluate different machine learning algorithms. In contrast to [14], we achieved the best results using a naïve bayes classifier without any modifications in Rapid Miner.

5.2. Evaluation

The evaluation results for the offending statement classification are listed in table 4. Both approaches, the baseline classifier and our pattern-based approach, achieved moderate to good results in terms of f1. However, while the baseline classifier achieved higher recall values, the pattern-based approach achieved substantially better precision values.

Table 4. Offending statement classification results (in %)

	Dataset 1			Dataset 2		
	p	r	f1	p	r	f1
Baseline	53.57	76.27	62.94	50.65	71.43	59.27
Pattern-based	75.26	61.86	67.91	73.89	53.46	62.03

The baseline classifier seems to causes false positives by misjudging cases that contain direct or indirect references not belonging to offensive statements. As the approach is based on a bag-of-words model, the context of offensive statements cannot be analyzed directly. In contrast, the pattern-based approach yields less false positives resulting in better precision values. Better precision values reduce the effort for personnel as fewer false positives are detected that need to be corrected in a subsequent step. Additionally, substantial precision values are more suitable for fully automated classification.

Furthermore, the pattern-based approach is able to assess severity values to detected offensive statements. We further investigated the classification performance by distinguishing between the classes offensive (severity 1) and severely offensive (severity 2). The results for each form are denoted in table 5.

Table 5. Severity classification results (in %)

	Dataset 1			Dataset 2		
Severity	p	r	f1	p	r	f1
1	49.51	83.33	62.11	42.17	74.47	53.85
2	69.64	84.78	76.47	70.24	81.94	75.64

While the results for cases with a severity value of 1 are moderate, we were able to achieve good results in terms of f1 value for the detection of severe offending statements. The precision values indicate, that the system is reasonably accurate in detecting such statements and might be used accordingly in practical applications as we will discuss in the next section.

Table 6. Target classification results (in %)

	Dataset 1			**Dataset 2**		
	p	r	f1	p	r	f1
Foreigner	**51.79**	**65.91**	**58**	**59.26**	**33.56**	**44.44**
Government	**76.32**	**58**	**65.91**	**74.07**	**51.28**	**60.61**
Community	12.5	20	15.39	55.56	83.33	66.67
Press	81.82	77.14	79.41	80	100	88.89

Finally, the evaluation results for the reference identification are subsumed in table 6. The performance measurement for the multi class problem yields contrary results. The results show that a moderate amount of the offensive statements directed towards foreigners was detected correctly, which is the main focus of our study. Frequently, offending statements towards foreigners come along with statements towards the government. In these cases, the classifier seems to misjudge foreigner references for government references and vice versa resulting in moderate overall performance for both of these classes. Additionally, substantial results for press and media class were achieved. These targets are often referenced directly and thus no indirect reference resolution is needed.

6. Practical applications

6.1. Automatic blocking of hate speech

Our approach can be used as a basis for systems that are able to automatically block offending comments. In a proactive manner, the system prevents offending content from its publication. This way, other users are not influenced by the content of the message in a way that facilitates incitement towards foreigners or political parties. Moreover, emotional costs are avoided as they do not have to cope with such content. In contrast to moderators, automated systems are capable of processing a vast amount of messages, which is important in the context of social media platforms as messages can rapidly spread in a viral manner [6]. Furthermore, users with the intention to facilitate incitement might create multiple accounts to bypass suspensions from the social media platform. A proactive system prevents the publication of offending content independent of the account and its message history.

The evaluation revealed that the presented approach is suitable for this kind of practical application with limitations. In automated processing no human control instance that examines the results is involved and thus the cost of false positives need to be considered. A high precision value results in fewer occurrences of false positives and thus reducing these costs. However, precision values around 70 percent result in a fair amount of false positives. Such falsely blocked messages might frustrate users as their message is deleted without proper reason. However, the presented approach allows to assess severity values to indicate the offensiveness of a message. By assuming that substantial offensive content is more likely to violate existing policies or laws, the system can automatically block or delete such messages selectively. As the results in the evaluation section show, the approach can distinguish between offensive and severely offensive statements with substantial precision.

Furthermore, blocking comments is opposed to the right of freedom of speech. Thus, a goal conflict exists between preserving freedom of speech and protecting the victims, authors and the social media platform against potential legal consequences caused by hate speech. It needs to be considered that the decision whether or not a concrete statement from a user violates a certain law is subject to courts and cannot be judged by an automated system.

6.2. Marking comments

The proposed approach can be used in a semi-automated way by automatically marking comments potentially containing offensive statements to present them to a moderator in a subsequent step. Moderators can examine the selected messages and decide, if further actions need to be taken. As a consequence, the effort is reduced compared to manually examining the vast amount of messages in total. Furthermore, communities that are characterized by a substantial amount of published hate speech can be detected as intended by the task force of the German government and Facebook [1]. Marked comments might also be displayed to the author himself before their publication. This way, the author might reconsider the formulation of the message. An offensive comment that is a result of hastily reactions or is, despite its formulation, not intended to be offensive, might be prevented. Finally, community managers might use the system as a third party tool to analyze the comments in response to their published posts. The community manager can then detect problematic comments independently of administrators and eventually remove them.

Considering the moderate to good overall classification performance, the approach is useful for such a task. Compared to automatic blocking or deletion, marking potentially offensive comments shifts the responsibility for the final decision to the human control instance. As a human being is able to take more informed decisions considering multiple aspects on a case-to-case basis, freedom of speech might be preserved more accurately.

6.3. Breaking echo chambers

To tackle the problem of homogenous viewpoints in echo chambers, they first need to be detected, especially those characterized by polarized and homogenous right-wing opinions concerning foreigners or refugees. Our presented approach is suitable for this task, as it is able to detect the referenced victims. If a substantial amount of the offensive statements detected in a community (or in our case Facebook page) is directed towards foreigners, it is likely that the community is characterized by such an echo chamber. The evaluation results reveal, that the foreigner group can be identified precisely and thus, such a detection is possible. Due to the large amount of messages, the chance of detection is improved further.

After the detection of such echo chambers, the beliefs of the users might be challenged by presenting them controversial and well-researched information [11]. The EdgeRank in Facebook, for example, could be adjusted to selectively inject such content. This way, freedom of speech is not violated and each user can decide on his own whether to consider the presented content in its opinion-formation process or not. The presented approach is not able to select appropriate information and selectively inject it into social media. However, prior research addressed this problem in the context of news [34] as well as political discourse in blogs [35]. Such methods might be applied to select appropriate information sources.

7. Conclusion

Recently, offending statements towards refugees and foreigners in social media drew attention to the broader public and are recognized by politicians and social media companies as a growing problem [4,1]. In this work, we proposed, implemented and evaluated an approach to automatically detect offensive statements directed towards foreigners to aid social media platforms in the labor-intensive task of moderation.

We modified the pattern-based approach from Bretschneider et al. [15] to support the German language and to detect and resolve referenced victims, especially foreigners and refugees as well as the government. This step is required as users often refer to their targets indirectly, for example, by paraphrasing or referring to content of the corresponding article or post. Finally, the approach assesses severity values to indicate slightly to offensive statements and severe offensive statements.

To evaluate our approach, we developed an annotated dataset with two human experts providing access to it as a benchmark for further research under the URL www.ub-web.de/research/. The annotations contain offending passages, the referenced victim and a severity value. As evaluation metrics were applied precision, recall and f1-measure. Compared to a machine learning baseline classifier our pattern-based approach yields substantial precision values (75.26% and 73.89%) and moderate overall classification performance in terms of f1 value (67.91% and 62.03%).

We discussed three practical applications: automated blocking and marking of offensive content as well as detecting echo chambers. The achieved precision values allow automated processing of offensive content with limitations as there is a fair amount of remaining false positives. The approach could be used selectively by distinguishing between severely offending content that might be automatically blocked and other offending statements that might be presented to moderators in a semi-automated manner. As we are able to identify the referenced victims, the approach can be used to detect echo chambers containing homogenous xenophobic or racist viewpoints. To aid users kept in such echo chambers, controversial and well-researched information might be presented to them [11]. This way, the existing, potentially polarized, beliefs of social media users are challenged and the political opinion-formation-process is based on more diverse information [13]. Applying such an approach has ethical implications that need to be carefully considered. A major concern is the conflict between preserving freedom of speech and protecting others from hate speech possibly conflicting with their individual rights [17]. Furthermore, if the system is used in an automated manner the responsibility of judging the behavior of users entirely relies on a machine.

Further research is desired on several aspects. First, we did not consider characteristics of the sender of hate speech as the approach can be applied in anonymous contexts. However, such characteristics might improve the classification performance. Second, the approach is not capable of detecting paraphrased offending statements, for example in the form of gender based harassment. To identify such cases, semantic approaches might be applied as an extension. Moreover, to apply the method to different languages, a general framework or guideline could be created to aid this process in a structured way. Finally, the system is not capable of incorporating cross-cultural differences in the perception of offending content. To capture such differences, several configurations containing different hate speech patterns could be analyzed.

10. References

[1] URL: http://www.bbc.com/news/world-europe-34256960, last accessed 05/30/2016.

[2] URL: http://wapo.st/1LMY05q, last accessed 05/30/2016.

[3] K. Wise, B. Hamman, and K. Thorson, “Moderation, Response Rate, and Message Interactivity”, *Journal of Computer-Mediated Communication*, vol. 12, no. 1, pp. 24-41, 2006.

[4] URL: http://www.bbc.co.uk/newsbeat/article/30694252 /why-are-thousands-of-germans-protesting-and-who-are-pegida, last accessed 05/30/2016.

[5] URL: http://newsroom.fb.com/company-info/, last accessed 06/02/2016.

[6] R.A. King, P. Racherla, and V.D. Bush, “What We Know and Don't Know About Online Word-of-Mouth”, *Journal of Interactive Marketing*, vol. 28, no. 3, pp. 167-183, 2014.

[7] E. Gilbert, and K. Karahalios, “Predicting Tie Strength with Social Media”, in *SIGCHI Conference on Human Factors in Computing Systems*, Boston, USA, pp. 211-220, 2009.

[8] L.M. Jones, K.J. Mitchell, and D. Finkelhor, “Online harassment in context: Trends from three Youth Internet Safety Surveys (2000, 2005, 2010)”, *Psychology of Violence*, vol. 3, no. 1, pp. 53-69, 2013.

[9] M.L. Williams, and P. Burnap, “Cyberhate on Social Media in the aftermath of Woolwich”, *British Journal of Criminology*, vol. 56, no. 2, pp. 211-238, 2016.

[10] J. Glaser, J. Dixit, and D.P. Green, “Studying Hate Crime with the Internet: What Makes Racists Advocate Racial Violence?”, *Journal of Social Issues*, vol. 58, pp. 177-193, 2002.

[11] A. Gruzd, and J. Roy, “Investigating Political Polarization on Twitter: A Canadian Perspective”, *Policy & Internet*, vol. 6, no. 1, pp. 28-45, 2014.

[12] M.D. Conover, J. Ratkiewicz, M. Francisco, B. Goncalves, A. Flammini, and F. Menczer, “Political Polarization on Twitter”, in *International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain, 2011.

[13] C.R. Sunstein, “The Law of Group Polarization”, *Journal of Political Philosophy*, vol. 10, no. 2, pp. 175-195, 2002.

[14] Y. Chen, Y. Zhou, Y. Zhu, and H. Xu, “Detecting Offensive Language in Social Media to Protect Adolescent Online Safety”, in *International Conference on Privacy, Security, Risk and Trust*, Amsterdam, Netherlands, pp. 71-80, 2012.

[15] U. Bretschneider, T. Wöhner, and R. Peters, “Detecting Online Harassment in Social Networks”, in *International Conference on Information Systems*, Auckland, New Zealand, 2014.

[16] URL: http://www.un.org/en/universal-declaration-human-rights/, last accessed 22/08/2016.

[17] URL: http://www.ohchr.org/en/professionalinterest/ pages/ccpr.aspx, last accessed 22/08/2016.

[18] H. Keller, and M. Sigron, “State Security v Freedom of Expression”, *Human Rights Law Review*, vol. 10, no. 1, pp. 151-168, 2010.

[19] S. W. Kim, and A. Douai, “Google vs. China’s ‘Great Firewall’”, *Technology in Society*, vol. 34, no. 2, pp. 174-181, 2012.

[20] M. Oetheimer, “Protecting Freedom of Expression”, *Cardozo Journal of International & Comparative Law*, vol. 17, no. 3, pp. 427-443, 2009.

[21] A. Weber, “Manual on hate speech”, Council of Europe Publishing, Strasbourg Cedex, France, 2009.

[22] URL: http://hudoc.echr.coe.int/eng?i=001-155105, last accessed 24/05/2016.

[23] A. R. Hochschild, “Emotion Work, Feeling Rules, and Social Structure”, *American Journal of Sociology*, vol. 85, no. 3, pp. 551-575, 1979.

[24] A. Menking, and I. Erickson, “The Heart Work of Wikipedia: Gendered, Emotional Labor in the World’s Largest Online Encyclopedia”, in *ACM Conference on Human Factors in Computing Systems*, Seoul, Korea, pp. 207-210, 2015.

[25] V. O. Key, “The Responsible Electorate: Rationality in Presidential Voting”, Belknap Press, Cambridge, USA, 1966.

[26] URL: https://www.facebook.com/help/ 327131014036297/, last accessed 06/02/2016.

[27] R.S. Tokunaga, “Following you home from school: A critical review and synthesis of research on cyberbullying victimization”, *Computers in Human Behavior*, vol. 26, no. 3, pp. 277-287, 2010.

[28] S.O. Sood, E.F. Churchill, and J. Antin, “Automatic identification of personal insults on social news sites”, *Journal of the American Society for Information Science and Technology*, vol. 63, no. 2, pp. 270-285, 2012.

[29] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, “Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying”, *ACM Transactions on Interactive Intelligent Systems*, vol. 2, no. 2, pp. 1-30, 2012.

[30] J.M. Xu, K.S. Jun, X. Zhu, and A. Bellmore, “Learning from Bullying Traces in Social Media”, in *Conference of the NAACL: HLT*, Stroudsburg, USA, pp. 656-666, 2012.

[31] A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards, “Detecting cyberbullying”, in *the 5th Annual ACM Web Science Conference*, Paris, France, pp. 195-204, 2013.

[33] URL: http://www.hyperhero.com/de/insults.htm, last accessed 06/02/2016.

[34] M. Sokolova, and G. Lapalme, “A systematic analysis of performance measures for classification tasks”, *Information Processing and Management*, vol. 45, no. 4, pp. 427-437, 2009.

[35] S. Park, S. Kang, S. Chung, and J. Song, “NewsCube: delivering multiple aspects of news to mitigate media bias”, in *SIGCHI Conference on Human Factors in Computing Systems*, New York, USA, pp. 443-452, 2009.

[36] A. Oh, H. Lee, and Y. Kim, “User Evaluation of a System for Classifying and Displaying Political Viewpoints of Weblogs”, in *AAAI International Conference on Weblogs and Social Media*, San Jose, USA, 2009.