Originally published in: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, April 3, 2017, Valencia, Spain. - Stroudsburg: Association for Computational Linguistics, 2017. Pp. 1-10.

A Survey on Hate Speech Detection using Natural Language Processing

Anna Schmidt
Spoken Language Systems
Saarland University
D-66123 Saarbrücken, Germany
anna.schmidt@lsv.uni-saarland.de

Michael Wiegand
Spoken Language Systems
Saarland University
D-66123 Saarbrücken, Germany
michael.wiegand@lsv.uni-saarland.de

Abstract

This paper presents a survey on hate speech detection. Given the steadily growing body of social media content, the amount of online hate speech is also increasing. Due to the massive scale of the web, methods that automatically detect hate speech are required. Our survey describes key areas that have been explored to automatically recognize these types of utterances using natural language processing. We also discuss limits of those approaches.

1 Introduction

Hate speech is commonly defined as any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic (Nockleby, 2000). Examples are (1)-(3).[1]

(1) Go fucking kill yourself and die already useless ugly pile of shit scumbag.
(2) The Jew Faggot Behind The Financial Collapse
(3) Hope one of those bitches falls over and breaks her leg

Due to the massive rise of user-generated web content, in particular on social media networks, the amount of hate speech is also steadily increasing. Over the past years, interest in online hate speech detection and particularly the automatization of this task has continuously grown, along with the societal impact of the phenomenon. Natural language processing focusing specifically on this phenomenon is required since basic word filters do not provide a sufficient remedy: What is considered a hate speech message might be influenced by aspects such as the domain of an utterance, its discourse context, as well as context consisting of co-occurring media objects (e.g. images, videos, audio), the exact time of posting and world events at this moment, identity of author and targeted recipient.

This paper provides a short, comprehensive and structured overview of automatic hate speech detection, and outlines the existing approaches in a systematic manner, focusing on feature extraction in particular. It is mainly aimed at NLP researchers who are new to the field of hate speech detection and want to inform themselves about the state of the art.

2 Terminology

In this paper we use the term ***hate speech***. We decided in favour of using this term since it can be considered a broad umbrella term for numerous kinds of insulting user-created content addressed in the individual works we summarize in this paper. *Hate speech* is also the most frequently used expression for this phenomenon, and is even a legal term in several countries. Below we list other terms that are used in the NLP community. This should also help readers with finding further literature on that task.

In the earliest work on hate speech, Spertus (1997) refers to *abusive* messages, *hostile* messages or *flames*. More recently, many authors have shifted to employing the term *cyberbullying* (Xu et al., 2012; Hosseinmardi et al., 2015; Zhong et al., 2016; Van Hee et al., 2015; Dadvar et al., 2013; Dinakar et al., 2012). The actual term *hate speech* is used by Warner and Hirschberg (2012), Burnap and Williams (2015), Silva et al. (2016), Djuric et al. (2015), Gitari et al. (2015), Williams and Burnap (2015) and Kwok and Wang (2013). Further,

[1] The examples in this work are included to illustrate the severity of the hate speech problem. They are taken from actual web data and in no way reflect the opinion of the authors.

Sood et al. (2012a) work on detecting (personal) *insults*, *profanity* and user posts that are characterized by *malicious intent*, while Razavi et al. (2010) refer to *offensive language*. Xiang et al. (2012) focus on *vulgar* language and *profanity-related offensive content*. Xu et al. (2012)[2] further look into jokingly formulated *teasing* in messages that represent (possibly less severe) bullying episodes. Finally, Burnap and Williams (2014) specifically look into *othering language*, characterized by an us-them dichotomy in racist communication.

3 Features for Hate Speech Detection

As is often the case with classification-related tasks, one of the most interesting aspects distinguishing different approaches is which features are used. Hate speech detection is certainly no exception since what differentiates a hateful speech utterance from a harmless one is probably not attributable to a single class of influencing aspects. While the set of features examined in the different works greatly varies, the classification methods mainly focus on supervised learning (§6).

3.1 Simple Surface Features

For any text classification task, the most obvious information to utilize are surface-level features, such as bag of words. Indeed, unigrams and larger n-grams are included in the feature sets by a majority of authors (Chen et al., 2012; Xu et al., 2012; Warner and Hirschberg, 2012; Sood et al., 2012b; Burnap and Williams, 2015; Van Hee et al., 2015; Waseem and Hovy, 2016; Burnap and Williams, 2016; Hosseinmardi et al., 2015; Nobata et al., 2016). These features are often reported to be highly predictive. Still, in many works n-gram features are combined with a large selection of other features. For example, in their recent work, Nobata et al. (2016) report that while token and character n-gram features are the most predictive single features in their experiments, combining them with all additional features further improves performance.

Character-level n-gram features might provide a way to attenuate the spelling variation problem often faced when working with user generated comment text. For instance, the phrase *kil1 yrslef a$$hole*, which is regarded as an example of hate speech, will most likely pose problems to token-based approaches since the unusual spelling variations will result in very rare or even unknown tokens in the training data. Character-level approaches, on the other hand, are more likely to capture the similarity to the canonical spelling of these tokens. Mehdad and Tetreault (2016) systematically compare character n-gram features with token n-grams for hate speech detection, and find that character n-grams prove to be more predictive than token n-grams.

Apart from word- and character-based features, hate speech detection can also benefit from other surface features (Chen et al., 2012; Nobata et al., 2016), such as information on the frequency of URL mentions and punctuation, comment and token lengths, capitalization, words that cannot be found in English dictionaries, and the number of non-alpha numeric characters present in tokens.

3.2 Word Generalization

While bag-of-words features usually yield a good classification performance in hate speech detection, in order to work effectively these features require predictive words to appear in both training and test data. However, since hate speech detection is usually applied on small pieces of text (e.g. passages or even individual sentences), one may face a data sparsity problem. This is why several works address this issue by applying some form of *word generalization*. This can be achieved by carrying out word clustering and then using induced cluster IDs representing sets of words as additional (generalized) features. A standard algorithm for this is *Brown clustering* (Brown et al., 1992) which has been used as a feature in Warner and Hirschberg (2012). While Brown clustering produces hard clusters – that is, it assigns each individual word to one particular cluster – *Latent Dirichlet Allocation (LDA)* (Blei et al., 2003) produces for each word a topic distribution indicating to which degree a word belongs to each topic. Such information has similarly been used for hate speech detection (Xiang et al., 2012; Zhong et al., 2016).

More recently, distributed word representations (based on neural networks), also referred to as *word embeddings*, have been proposed for a similar purposes. For each word a vector representation is induced (Mikolov et al., 2013) from a large (unlabelled) text corpus. Such vector representations have the advantage that different, semanti-

[2]The data from this work are available under `http://research.cs.wisc.edu/bullying`

cally similar words may also end up having similar vectors. Such vectors may eventually be used as classification features, replacing binary features indicating the presence or frequency of particular words. Since in hate speech detection sentences or passages are classified rather than individual words, a vector representation of the *set* of word vectors representing the words of the text to be classified is sought. A simple way to accomplish this is by averaging the vectors of all words occurring in one passage or sentence. For detecting hate speech, this method is only reported to have limited effectiveness (Nobata et al., 2016), no matter whether general pretrained embeddings are used or the embeddings are induced from a domain-specific corpus. Alternatively, Djuric et al. (2015) propose to use embeddings that directly represent the text passages to be classified. These *paragraph embeddings* (Le and Mikolov, 2014), which are internally based on word embeddings, have been shown to be much more effective than the averaging of word embeddings (Nobata et al., 2016).

3.3 Sentiment Analysis

Hate speech and sentiment analysis are closely related, and it is safe to assume that usually negative sentiment pertains to a hate speech message. Because of this, several approaches acknowledge the relatedness of hate speech and sentiment analysis by incorporating the latter as an auxiliary classification. Dinakar et al. (2012), Sood et al. (2012b) and Gitari et al. (2015) follow a multi-step approach, in which a classifier dedicated to detect negative polarity is applied prior to the classifier specifically checking for evidence of hate speech. Further, Gitari et al. (2015) run an additional classifier that weeds out non-subjective sentences prior to the aforementioned polarity classification.

Apart from multi-step approaches, there are also single-step approaches that include some form of sentiment information as a feature. For example, in their supervised classifier, Van Hee et al. (2015) use as features the number of positive, negative, and neutral words (according to a sentiment lexicon) occurring in a given comment text.

Further attempts to isolate the subset of hate speech from the set of negative polar utterances rest on the observation that hate speech also displays a *high degree* of negative polarity (Sood et al., 2012b; Burnap et al., 2013). To that end, polarity classifiers are employed which in addition to specifying the type of polarity (i.e. *positive* and *negative*) also predict the polar intensity of an utterance. A publicly available polarity classifier which produces such an output is *SentiStrength* (Thelwall et al., 2010). It is used for hate speech detection by Burnap et al. (2013).

3.4 Lexical Resources

Trying to make use of the general assumption that hateful messages contain specific negative words (such as slurs, insults, etc.), many authors utilize the presence of such words as a feature. To obtain this type of information lexical resources are required that contain such predictive expressions.

A popular source for such word lists is the web. There are several publicly available lists that consist of *general* hate-related terms.[3] Apart from works that employ such lists (Xiang et al., 2012; Burnap and Williams, 2015; Nobata et al., 2016), there are also approaches, such as Burnap and Williams (2016) which focus on lists that are *specialized* towards a particular subtype of hate speech, such as ethnic slurs[4], LGBT slang terms[5], or words with a negative connotation towards handicapped people.[6]

Apart from publicly-available word lists from the web other approaches incorporate lexicons that have been specially compiled for the task at hand. Spertus (1997) employs a lexicon comprising so-called *good verbs* and *good adjectives*. Razavi et al. (2010) manually compiled an *Insulting and Abusing Language Dictionary* containing both words and phrases with different degrees of manifestation of flame varieties. This dictionary also assigns weights to each lexical entry which represents the degree of the potential impact level for hate speech detection. The weights are obtained by *adaptive learning* using the training partition of the data set used in that work. Gitari et al. (2015) build a resource comprising *hate verbs* which are verbs that condone or encourage acts of violence. Despite their general effectiveness, rel-

[3] www.noswearing.com/dictionary, www.rsdb.org, www.hatebase.org

[4] https://en.wikipedia.org/wiki/List_of_ethnic_slurs

[5] https://en.wikipedia.org/wiki/List_of_LGBT_slang_terms

[6] https://en.wikipedia.org/wiki/List_of_disability-related_terms_with_negative_connotations

atively little is known about the creation process and the theoretical concepts that underlie the lexical resources that have been specially compiled for hate speech detection.

Most approaches employ lexical features either as some baseline or in addition to other features. In contrast to other features, particularly bag of words (§3.1) or embeddings (§3.2), they are usually insufficient as a stand-alone feature (Nobata et al., 2016). Contextual factors play an important role. For example, Hosseinmardi et al. (2015) find that 48% of media sessions in their data collection were not deemed hate speech by a majority of annotators, even though they reportedly contained a high percentage of profanity words.

3.5 Linguistic Features

Linguistic aspects also play an important role for hate speech detection. Linguistic features are either employed in a more generic fashion or are specifically tailored to the task.

Xu et al. (2012) explore the combination of ngram features with POS-information-enriched tokens. However, adding POS information does not significantly improve classifier performance.

Taking into account deeper syntactic information as a feature, Chen et al. (2012) employ typed dependency relationships. Such relationships have the potential benefit that non-consecutive words bearing a (potentially long-distance) relationship can be captured in one feature. For instance, in (4) a dependency tuple `nsubj(pigs, Jews)` will denote the relation between the offensive term *pigs* and the hate-target *Jews*.

(4) Jews are lower class pigs.

Obviously, knowing that those two words are syntactically related makes the underlying statement more likely to convey hate speech than those keywords occurring in a sentence without any syntactic relation. Dependency relationships are also employed in the feature set from Gitari et al. (2015), Burnap and Williams (2015), Burnap and Williams (2016) and Nobata et al. (2016). Burnap and Williams (2015) and Burnap and Williams (2016) report significant performance improvements based on this feature; the other papers do not conduct ablation studies from which one could conclude the effectiveness of this particular feature. There is also a difference in the sets of dependency relationships representing a sentence which are used. Burnap and Williams (2015) apply some statistical feature selection (*Bayesian Logistic Regression*), Chen et al. (2012) and Gitari et al. (2015) manually select the relations (e.g. by enforcing that one argument of the relation is an offensive term) while Nobata et al. (2016) do not carry out any further selection. Unfortunately, there does not exist any evaluation comparing these feature variations. Zhong et al. (2016) do not use the presence of explicit dependency relations occurring in a sentence as a feature but employ an *offensiveness level score*. This score is based on the frequency of co-occurrences of offensive terms and user identifiers in the same dependency relation.

In her work on the *Smokey* system, Spertus (1997) devises a set of linguistic features tailored to the task of hate speech detection. The syntactic features include the detection of *imperative* statements (e.g. *Get lost!*, *Get a life!*) and the co-occurrence of the pronoun *you* modified by noun phrases (as in *you bozos*). The *Smokey* system also incorporates some semantic features to prevent false positives. On the one hand, so-called *praise rules* are employed, which use regular expressions involving pre-defined *good words*. Since that work categorizes webpages, the praise rules try to detect co-occurrences of good words and expressions referring to the website to be classified. On the other hand, Spertus (1997) also employs *politeness rules* represented by certain polite words or phrases (e.g. *no thanks*, *would you* or *please*). Nobata et al. (2016) use a similar feature.

3.6 Knowledge-Based Features

Hate speech detection is a task that cannot be solved by simply looking at keywords. Even if one tries to model larger textual units, as researchers attempt to do by means of linguistic features (§3.5), it remains difficult to decide whether some utterance represents hate speech or not. For instance, (5) may not be regarded as some form of hate speech when only read in isolation.

(5) Put on a wig and lipstick and be who you really are.

However, when the context information is given that this utterance has been directed towards a boy on a social media site for adolescents[7], one could infer that this is a remark to malign the sexuality or gender identity of the boy being addressed (Dinakar et al., 2012). (5) displays stereotypes most

[7]The example utterance from above is from Formspring.

commonly attributed to females (i.e. *putting on a wig and lipstick*). If these characteristics are attributed to a male in a heteronormative context, the intention may have been to insult the addressee.

The above example shows that whether a message is hateful or benign can be highly dependent on world knowledge, and it is therefore intuitive that the detection of a phenomenon as complex as hate speech might benefit from including information on aspects not directly related to language. Dinakar et al. (2012) present an approach employing automatic reasoning over world knowledge focusing on anti-LGBT hate speech. The basis of their model is the general-purpose ontology *ConceptNet* (Liu and Singh, 2004), which encodes concepts that are connected by relations to form assertions, such as *"a skirt is a form of female attire"*. *ConceptNet* is augmented by a set of stereotypes (manually) extracted from the social media network *Formspring*.[8] An example for such a stereotype assertion is *"lipstick is used by girls"*. The augmented knowledge base is referred to as *BullySpace*.[9] This knowledge base allows computing the similarity of concepts of common knowledge with concepts expressed in user comments.[10] After extracting concepts present in a given user comment, the similarity between the extracted concepts and a set of four *canonical concepts* is computed. Canonical concepts are the four reference concepts *positive* and *negative valence* and the two genders, *male* and *female*. The resulting similarity scores between extracted and canonical concepts indicate whether a message might constitute a hate speech instance. A hate speech instance has a high similarity to the canonical concept *negative valence* and the canonical concept representing the gender opposed to the actual gender of the user being addressed in the message post. For example, for the sentence given above, a high similarity to *negative valence* and *female* would correctly indicate that the utterance is meant as hate speech.

Obviously, the approach proposed by Dinakar et al. (2012) only works for a very confined subtype of hate speech (i.e. anti-LGBT bullying). Even though the framework would also allow for other types of hate speech, it would require domain-specific assertions to be included first. This would require a lot of manual coding. It is presumably this shortcoming that explains why, to our knowledge, this is the only work that tries to detect hate speech with the help of a knowledge base.

3.7 Meta-Information

World knowledge gained from knowledge bases is not the only information available to refine inconclusive classification. Meta-information (i.e. information *about* an utterance) is also a valuable source to hate speech detection. Since the text commonly used as data for this task almost exclusively comes from social media platforms, a variety of such meta-information is usually offered and can be easily accessed via the APIs those platforms provide.

Having some background information about the user of a post may be very predictive. A user who is known to write hate speech messages may do so again. A user who is not known to write such messages is unlikely to do so in future. Xiang et al. (2012) effectively employ this heuristic in inferring further hate speech messages. Dadvar et al. (2013) use as a feature the number of profane words in the message history of a user. Knowing the gender of the user may also help (Dadvar et al., 2012; Waseem and Hovy, 2016). Men are much more likely to post hate speech messages than women.

Beyond these, several other kinds of meta-information are common, such as the number of posts by a user, the number of replies to a post, the average of the total number of replies per follower or the geographical origin, but most of these have not been found effective for classification (Zhong et al., 2016; Waseem and Hovy, 2016). Moreover, there are certain kinds of meta-information for which conflicting results have been reported. For instance, Hosseinmardi et al. (2015) report a correlation between the number of associated comments to a post and hate speech while Zhong et al. (2016) report the opposite. (Both papers use Instagram as a source.) Many reasons may be responsible for that. Zhong et al. (2016) speculate that the general lack in effectiveness of the meta-information they examined may be due to the fact they consider celebrity accounts. Accounts from regular users, on the other hand, may display quite a different behaviour. From that we conclude that

[8] The augmentation is achieved by applying the joint inference technique *blending* after both *ConceptNet* and the assertions have been transformed into a so-called *AnalogySpace*.

[9] *BullySpace* contains 200 LGBT-specific assertions.

[10] Concepts are represented as vectors, so the similarity can be easily computed by measures such as cosine-similarity.

meta-information may be helpful but it depends on the exact type of information one employs and also the source from which the data originate.

3.8 Multimodal Information

Modern social media do not only consist of text but also include images, video and audio content. Such non-textual content is also regularly commented on, and therefore becomes part of the discourse of a hate speech utterance. This context outside a written user comment can be used as a predictive feature.

As for knowledge-based features, not too many contributions exist that exploit this type of information. This is slightly surprising, since among hateful user posts illustrated by websites documenting representative cases of severe cyber hate[11], visual context plays a major role.

Hosseinmardi et al. (2015) employ features based on image labels, shared media content, and labelled image categories. Zhong et al. (2016) make use of pixel level image features and report that a combination of those visual features and features derived from captions gives best performance. They also employ these features for predicting which images are *bully-prone*. These are images that are likely to attract hate speech comments, and are referred to as *bullying triggers*.

4 Persons Involved in Bullying Episodes and Their Roles

Apart from detecting hateful messages, a group of works focuses on persons involved in hate speech episodes and their roles. Xu et al. (2012) look at the entire bullying event (or *bullying trace*), automatically assigning roles to actors involved in the event as well as the message author. They differentiate between the roles *bully*, *victim*, *assistant*, *defender*, *bystander*, *reinforcer*, *reporter* and *accuser* for tweet authors and for person mentions within the tweet. Aside from classifying insulting messages, Sood et al. (2012b) also automatically predict whether such messages are directed at an author of a previous comment or at a third party. Silva et al. (2016) provide an analysis of the main hate target groups on the two social media platforms Twitter and Whisper. The authors conclude that both platforms exhibit the same top 6 hate target groups: People are mostly bullied for their ethnicity, behaviour, physical characteristics, sexual orientation, class or gender. Chau and Xu (2007) present a study of a selected set of 28 anti-Black *hate groups* in blogs on the Xanga site. Using a semi-automated approach, they find demographical and topological characteristics of these groups. Using web-link and -content analysis, Zhou et al. (2005) examine the structure of US domestic extremist groups.

5 Anticipating Alarming Societal Changes

Apart from detecting individual, isolated hateful comments and classifying the types of users involved, the overall *proportion* of extreme negative posts over a certain time-span also allows for interesting avenues of research. Insights into changes in public or personal mood can be gained. Information on notable *increases* in the number of hateful posts within a short time span might indicate suspicious developments in a community. Such information could be utilized to circumvent incidents such as racial violence, terrorist attacks, or other crimes before they happen, thus providing steps in the direction of *anticipatory governance*.

One work concerned with crime prediction is Wang et al. (2012). This work focuses on forecasting hit-and-run crimes from Twitter data by effectively employing semantic role labelling and event-based topic extraction (with *LDA*). Burnap et al. (2013) examine the automatic detection of *tension* in social media. They establish that it can be reliably detected and visualized over time using sentiment analysis and lexical resources encoding topic-specific actors, accusations and abusive terms. Williams and Burnap (2015) temporally relate online hate speech with offline terrorist events. They find that the first hours following a terrorist event are the critical time span in which online hate speech may likely occur.

6 Classification Methods

The methods utilized for hate speech detection in terms of classifiers are predominantly supervised learning approaches. As classifiers mostly *Support Vector Machines* are used. Among the more recent methods, deep learning with *Recurrent Neural Network Language Models* has been employed in Mehdad and Tetreault (2016). There

[11] One example documenting disturbing cases of gender-based hate on facebook is `www.womenactionmedia.org/examples-of-gender-based-hate-speech-on- facebook/`

exist no comparative studies which would allow making judgement on the most effective learning method.

The different works also differ in the choice of classification procedure: Standard one-step classification approaches exist along with multi-step classification approaches. The latter approaches employ individual classifiers that solve subproblems, such as establishing negative polarity (§3.3).

Furthermore, some works employ semi-supervised approaches, particularly bootstrapping, which can be utilized for different purposes in the context of hate speech detection. On the one hand, it can be used to obtain additional training data, as it is for example done in Xiang et al. (2012). In this work, first a set of Twitter *users* is divided into *good* and *bad users*, based on the number of offensive terms present in their posts. Then *all* existing tweets of those bad users are selected and added to the training set as hate speech instances.

In addition, bootstrapping can also be utilized to build lexical resources used as part of the detection process. Gitari et al. (2015) apply this method to populate their hate verb lexicon, starting with a small seed verb list, and iteratively expanding it based on WordNet relations, adding all synonyms and hypernyms of those seed verbs.

7 Data and Annotation

To be able to perform experiments on hate speech detection, access to labelled corpora is essential. Since there is no commonly accepted benchmark corpus for the task, authors usually collect and label their own data. The data sources that are used include: Twitter (Xiang et al., 2012; Xu et al., 2012; Burnap et al., 2013; Burnap et al., 2014; Burnap and Williams, 2015; Silva et al., 2016), Instagram (Hosseinmardi et al., 2015; Zhong et al., 2016), Yahoo! (Nobata et al., 2016; Djuric et al., 2015; Warner and Hirschberg, 2012), YouTube (Dinakar et al., 2012), ask.fm (Van Hee et al., 2015), Formspring (Dinakar et al., 2012), Usenet (Razavi et al., 2010), Whisper[12] (Silva et al., 2016), and Xanga[13] (Chau and Xu, 2007). Since these sites have been created for different purposes, they may have special characteristics, and may therefore display different subtypes of hate speech. For instance, on a platform specially created for adolescents, one should expect quite different types of hate speech than on a service that is used by a cross-section of the general public since the resulting different demographics will have an impact on the topics discussed and the language used. These implications should be considered when interpreting the results of research conducted on a particular social media platform.

In general, the size of collected corpora varies considerably in works on hate speech detection, ranging from around 100 labelled comments used in the knowledge-based work by Dinakar et al. (2012) to several thousand comments used in other works, such as Van Hee et al. (2015) or Djuric et al. (2015). Apart from the classification approach taken, another reason for these size differences lies in the simple fact that annotating hate speech is an extremely time consuming endeavour: There are much fewer hateful than benign comments present in randomly sampled data, and therefore a large number of comments have to be annotated to find a considerable number of hate speech instances. This skewed distribution makes it generally difficult and costly to build a corpus that is balanced with respect to hateful and harmless comments. The size of a data set should always be taken into consideration when assessing the effectiveness of certain features or (learning) methods applied on it. Their effectiveness – or lack thereof – may be the result of a particular data size. For instance, features that tackle word generalization (§3.2) are extremely important when dealing with small data sets while on very large data sets they become less important since data sparsity is a less of an issue. We are not aware of any study examining the relation between the size of labeled training data and features/classifiers for hate speech detection.

In order to increase the share of hate speech messages while keeping the size of data instances to be annotated at a reasonable level, Waseem and Hovy (2016)[14] propose to pre-select the text instances to be annotated by querying a site for topics which are likely to contain a higher degree of hate speech (e.g. *Islam terror*). While this increases the proportion of hate speech posts on resulting data sets, it focuses the resulting data set to specific topics and certain subtypes of hate speech (e.g. hate speech targeting Muslims).

In order to annotate a data set manually, either expert annotators are used or crowdsourcing ser-

[12] `http://whisper.sh`

[13] `http://xanga.com`

[14] The data from this work are available under `http://github.com/zeerakw/hatespeech`

vices, such as Amazon Mechanical Turk (AMT), are employed. Crowdsourcing has obvious economical and organizational advantages, especially for a task as time-consuming as the one at hand, but annotation quality might suffer from employing non-expert annotators. Nobata et al. (2016) compare crowdsourced annotations performed using AMT with annotations created by expert annotators and find large differences in agreement.

In addition to the issues mentioned above that, to some extent, challenge the comparability of the research conducted on various data sets, the fact that no commonly accepted definition of hate speech exists further exacerbates this situation.

Previous works remain fairly vague when it comes to the annotation guidelines their annotators were given for their work. Ross et al. (2016) point out that this is particularly a problem for hate speech detection. Despite providing annotators with a definition of hate speech, in their work the annotators still fail to produce an annotation at an acceptable level of reliability.

8 Challenges

As the previous section suggests, the community would considerably benefit from a benchmark data set for the hate speech detection task underlying a commonly accepted definition of the task.

With the exception of Dutch (Van Hee et al., 2015) and German (Ross et al., 2016), we are not aware of any significant research being done on hate speech detection other than on English language data. We think that particularly a multilingual perspective to hate speech may be worthwhile. Unlike other tasks in NLP, hate speech may have strong cultural implications, that is, depending on one's particular cultural background, an utterance may be perceived as offensive or not. It remains to be seen in how far established approaches to hate speech detection examined on English are equally effective on other languages.

Although in the previous sections we also described approaches that try to incorporate the context of hate speech by employing some specific knowledge-based features (§3.6), meta-information (§3.7) or multi-modal information (§3.8), we still feel that there has been comparatively little work looking into these types of features. In the following, we illustrate the necessity of incorporating such context knowledge with the help of three difficult instances of hate speech. For all these cases, it is unclear whether the methods we described in this survey would correctly recognize these remarks as hate speech.

In (6) a woman is ridiculed for her voice. There is no explicit evaluation of her voice but it is an obvious inference from being compared with *Kermit the frog*. In (7), a Muslim is accused of bestiality. Again, there is no explicit accusation. The speaker of that utterance relies on his addressee to be aware of stereotyped prejudices against Islam. Finally, in (8), the speaker of that utterance wants to offend some girls by suggesting they are unattractive. Again, there is no explicit mention of being unattractive but challenging someone else's opposite view can be interpreted in this way.

(6) Kermit the frog called and he wants his voice back.
(7) Your goat is calling.
(8) Who was responsible for convincing these girls they were so pretty?

These examples are admittedly difficult cases and we are not aware of one individual method which would cope with all of these examples. It remains to be seen, whether in the future new computational approaches can actually solve these problems or whether hate speech is a research problem similar to sarcasm where only certain subtypes have been shown to be automatically detected with the help of NLP (Riloff et al., 2013).

9 Conclusion

In this paper, we presented a survey on the automatic detection of hate speech. This task is usually framed as a supervised learning problem. Fairly generic features, such as bag of words or embeddings, systematically yield reasonable classification performance. Character-level approaches work better than token-level approaches. Lexical resources, such as list of slurs, may help classification, but usually only in combination with other types of features. Various complex features using more linguistic knowledge, such as dependency-parse information, or features modelling specific linguistic constructs, such as imperatives or politeness, have also been shown to be effective. Information derived from text may not be the only cue suggesting the presence of hate speech. It may be complemented by meta-information or information from other modalities (e.g. images attached to messages). Making judgements about the general effectiveness of many of the complex features is

difficult since, in most cases, they are only evaluated on individual data sets, most of which are not publicly available and often only address a subtype of hate speech, such as bullying of particular ethnic minorities. For better comparability of different features and methods, we argue for a benchmark data set for hate speech detection.

Acknowledgements

We would like David M. Howcroft for proofreading this paper. The authors were partially supported by the German Research Foundation (DFG) under grant WI 4204/2-1 and the Cluster of Excellence Multimodal Computing and Interaction of the German Excellence Initiative.

References

David M. Blei, Andrew Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

P. Burnap and M. Williams. 2014. Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making. In *Internet, Policy and Politics Conference*, Oxford, United Kingdom.

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

Pete Burnap and Matthew L. Williams. 2016. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science*, 5(1):1–15.

Pete Burnap, Omer F. Rana, Nick Avis, Matthew Williams, William Housley, Adam Edwards, Jeffrey Morgan, and Luke Sloan. 2013. Detecting tension in online communities with computational twitter analysis. *Technological Forecasting and Social Change*, pages 96–108, May.

Pete Burnap, Matthew L. Williams, Luke Sloan, Omer Rana, William Housley, Adam Edwards, Vincent Knight, Rob Procter, and Alex Voss. 2014. Tweeting the terror: modelling the social media reaction to the woolwich terrorist attack. *Social Network Analysis and Mining*, 4(1):1–14.

Michael Chau and Jennifer Xu. 2007. Mining communities and their relationships in blogs: A study of online hate groups. *International Journal of Human-Computer Studies*, 65(1):57–70.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)*, pages 71–80, Amsterdam, Netherlands, September. IEEE.

Maral Dadvar, Franciska MG de Jong, RJF Ordelman, and RB Trieschnigg. 2012. Improved cyberbullying detection using gender information. *DIR 2012*, pages 22–25.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving Cyberbullying Detection with User Context. In *Proceedings of the European Conference in Information Retrieval (ECIR)*, pages 693–696, Moscow, Russia.

Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Trans. Interact. Intell. Syst.*, 2(3):18:1–18:30, September.

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

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Detection of cyberbullying incidents on the instagram social network. *CoRR*, abs/1503.03909.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In Marie desJardins and Michael L. Littman, editors, *AAAI*, pages 1621–1622, Bellevue, Washington, USA. AAAI Press.

Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the International Conference on Machine Learning (JMLR)*, pages 1188–1196, Beijing, China.

Hugo Liu and Push Singh. 2004. ConceptNet: A Practical Commonsense Reasoning Toolkit. *BT Technology Journal*, 22:211–226.

Yashar Mehdad and Joel Tetreault. 2016. Do characters abuse more than words? In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303, Los Angeles, CA, USA.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at the International Conference on Learning Representations (ICLR)*, Scottsdale, AZ, USA.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153, Geneva, Switzerland.

John T. Nockleby. 2000. Hate Speech. In Leonard W. Levy, Kenneth L. Karst, and Dennis J. Mahoney, editors, *Encyclopedia of the American Constitution*, pages 1277–1279. Macmillan, 2nd edition.

Amir H. Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Proceedings of the 23rd Canadian Conference on Advances in Artificial Intelligence*, AI'10, pages 16–27, Berlin, Heidelberg.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as Contrast between a Positive Sentiment and Negative Situation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 704–714, Seattle, WA, USA.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proceedings of the Workshop on Natural Language Processing for Computer-Mediated Communication (NLP4CMC)*, pages 6–9, Bochum, Germany.

Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *Proceedings of the Tenth International Conference on Web and Social Media*, pages 687–690, Cologne, Germany.

Sara Sood, Judd Antin, and Elizabeth Churchill. 2012a. Profanity use in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1481–1490, Austin, TX, USA. ACM.

Sara Owsley Sood, Elizabeth F. Churchill, and Judd Antin. 2012b. Automatic identification of personal insults on social news sites. *J. Am. Soc. Inf. Sci. Technol.*, 63(2):270–285, February.

Ellen Spertus. 1997. Smokey: Automatic recognition of hostile messages. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence*, AAAI'97/IAAI'97, pages 1058–1065, Providence, RI, USA. AAAI Press.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, and Di Cai. 2010. Sentiment Strength Detection in Short Informal Text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Proceedings of Recent Advances in Natural Language Processing, Proceedings*, pages 672–680, Hissar, Bulgaria.

Xiaofeng Wang, Matthew S Gerber, and Donald E Brown. 2012. Automatic crime prediction using events extracted from twitter posts. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 231–238.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, LSM '12, pages 19–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, USA, June. Association for Computational Linguistics.

Matthew Leighton Williams and Pete Burnap. 2015. Cyberhate on social media in the aftermath of woolwich: A case study in computational criminology and big data. *British Journal of Criminology*, pages 211–238.

Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1980–1984, Maui, HI, USA. ACM.

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

Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J. Miller, and Cornelia Caragea. 2016. Content-driven detection of cyberbullying on the instagram social network. In *IJCAI*, pages 3952–3958, New York City, NY, USA. IJCAI/AAAI Press.

Yilu Zhou, Edna Reid, Jialun Qin, Hsinchun Chen, and Guanpi Lai. 2005. US Domestic Extremist Groups on the Web: Link and Content Analysis. *IEEE intelligent systems*, 20(5):44–51.