Unraveling the Search Space of Abusive Language in Wikipedia with Dynamic Lexicon Acquisition

Wei-Fan Chen
Bauhaus-Universität Weimar
wei-fan.chen@uni-weimar.de

Khalid Al-Khatib
Bauhaus-Universität Weimar
khalid.alkhatib@uni-weimar.de

Matthias Hagen
Martin-Luther-Universität Halle-Wittenberg
matthias.hagen@informatik.uni-halle.de

Henning Wachsmuth
Paderborn University
henningw@upb.de

Benno Stein
Bauhaus-Universität Weimar
benno.stein@uni-weimar.de

Abstract

Many discussions on online platforms suffer from users offending others by using abusive terminology, threatening each other, or being sarcastic. Since an automatic detection of abusive language can support human moderators of online discussion platforms, detecting abusiveness has recently received increased attention. However, the existing approaches simply train one classifier for the whole variety of abusiveness. In contrast, our approach is to distinguish explicitly abusive cases from the more "shadowed" ones. By dynamically extending a lexicon of abusive terms (e.g., including new obfuscations of abusive terms), our approach can support a moderator with explicit unraveled explanations for why something was flagged as abusive: due to known explicitly abusive terms, due to newly detected (obfuscated) terms, or due to shadowed cases.

1 Introduction

The web has become the primary medium for people to share and discuss their opinions, stances, and knowledge. But not all people behave ethically on the respective online platforms: different types of abusive language have widely spread on the web. Systems that (semi-)automatically detect abusive language have gained quite some attention in the recent years. Such tools could support human moderators who try to protect online platforms from abusive language and to maintain high-quality user-generated content.

People use various ways to offend others. On one hand, they either *directly* offend the recipient of a text (direct recipient) or *indirectly* offend some other person, entity, or group (other recipient). On the other hand, abusive words and phrases may be used *explicitly* (e.g., "asshole!"), possibly in obfuscated form (e.g., "a$$h0le"), or abusiveness can also happen *implicitly* via sarcasm (e.g., "go back to school, whatever you learned didn't stick") or via new racist or abusive codes (e.g., on the platform *4chan*, "Google" is used as a slur for black people, "skittle" for Arabs, and "butterfly" for gays).[1]

Some recent studies have pointed to different types and to the importance of separating them, especially (Waseem et al., 2017). However, the distinction between the different offending dimensions has hardly been investigated for the development of abusive language classifiers (Schmidt and Wiegand, 2017). Accordingly, existing approaches consider the language of all abusive texts irrespective of their offending dimensions as one single search space. They simply train one machine learning model with different linguistic features on this space in order to classify unseen text as being abusive or not. Due to the diversity of language in offending dimensions, we expect such models to often result in limited effectiveness in practice. The reason is that, when learning to detect abusive texts following one way, for instance, the inclusion of training texts following other ways induces noise that diminishes the visibility of discriminative patterns.

As a solution, we propose to unravel the search space of abusive language via a three-stage classification approach. First, utilizing an abusive lexicon, we split the search space into two subspaces: texts with abusive words or phrases from the lexicon,

[1] https://mic.com/articles/155739

Proceedings of the 2nd Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, pages 76–82
Hong Kong, China, November 4, 2019 ©2019 Association for Computational Linguistics

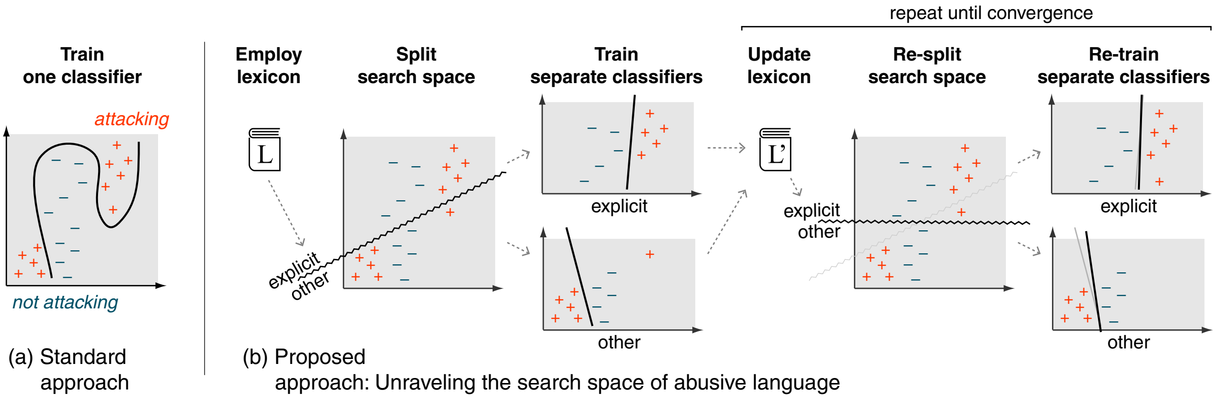

Figure 1: (a) Standard abusive language detection: Train a single classifier on all instances. (b) Proposed approach: Iteratively split the search space based on the offending dimension and train classifiers for each subspace.

and texts without such words. Second, we train a distinct classifier for each subspace. Third, using the predictions of the two classifiers, we perform an ablation test to discover new abusive terms from the subspaces. The found abusive words are added to the abusive lexicon that can serve as a dynamic source of explanations for a moderator that questions the detectors decision to flag a text as abusive. Figure 1 compares our approach to the "standard" single-search-space method.

To evaluate our approach to abusive language detection, we carried out several experiments using the *personal attacks corpus* of Wulczyn et al. (2017). The corpus consists of more than 100,000 comments from Wikipedia talk pages, each labeled as being a personal attack or not. In addition, the corpus includes manual labels for the target of attack, i.e., being the direct recipient or a third party.

The experimental results show that our search space unraveling slightly improves over state-of-the-art single-space classifiers with the additional bonus of a dynamic abusiveness lexicon that can help to explain the classifier's decisions.

The contribution of this paper is three-fold:

- We investigate how to unravel the search space of abusive language based on the underlying offending way.
- We develop computational approach that performs the unraveling in practice, and we evaluate it for the classification of Wikipedia talk page comments as being abusive or not.
- We dynamically develop a new lexicon for new abusive terms.

The developed resources are freely available on `https://webis.de`.

2 Related Work

The automatic detection of abusive language has been studied extensively in the last years. Proposed approaches target different types of abusive language, ranging from hate speech (Warner and Hirschberg, 2012) and cyberbullying (Nitta et al., 2013) to profanity (Sood et al., 2012) and personal attacks (Wulczyn et al., 2017).

Despite the importance of labeled data for abusive language detection, only few datasets are available so far for this task. Most of them come from large online platforms, such as Twitter (Waseem and Hovy, 2016), Yahoo (Nobata et al., 2016), and Wikipedia (Wulczyn et al., 2017). In terms of the number of labeled texts, the latter is the biggest, consisting of more than 100,000 Wikipedia talk page comments. We use this dataset for the evaluation of our approach.

Abusive (or offensive) language detection usually follows a supervised learning paradigm with either binary or multi-class classifiers. While existing abusiveness classifiers exploit a variety of lexical, syntactic, semantic, and knowledge-based features, one study showed character n-grams alone to be very good features (Mehdad and Tetreault, 2016). Until recently, the most effective overall approaches rely on neural network architectures such as CNN and RNN (Badjatiya et al., 2017; Pavlopoulos et al., 2017). On the personal attacks corpus, Pavlopoulos et al. (2017) have developed several very effective deep learning models with word embedding features. We employ the best-performing neural model, but we analyze the effect of adding our new approach (i.e., to unravel the abusiveness search space) that simultaneously helps to improve lexicon-based explainability.

An approach somewhat comparable to ours has been proposed by Dinakar et al. (2011) to detect cyberbullying on YouTube: different classifiers trained for different cyberbullying topics (e.g., sexuality, intelligence, and culture). The best results come from combining the individual classifiers, while a single multi-class classifier (mixing the different topics) was less effective.

Our approach is also related to co-training (Blum and Mitchell, 1998) and iterative feature selection/discovery (Liu et al., 2003; Xiang et al., 2012). In co-training, a labeled training set is extended by iteratively adding trustful instances from an unlabeled set based on the predictions of the classifier. Similarly, our approach extends its abusiveness lexicon iteratively. The iterative feature selection/discovery aims at finding new discriminating features to train the classifiers. This is in line with the third stage of our approach where new abusive terms are learned based on the predictions of the classifiers. The dynamically-updated lexicon can then serve as a good source for explaining many classifier decisions on the in-lexicon cases.

3 Data

In this section, we detail the data that we employ for the implementation and evaluation of our approach. Specifically, we describe the Wikipedia personal attack corpus (Wulczyn et al., 2017) and the abusive language lexicon of Wiegand et al. (2018).

3.1 Wikipedia Personal Attack Corpus

Wikipedia is one of the online platforms suffering from abusive language, especially from personal attacks (Shachaf and Hara, 2010). In particular, each Wikipedia article is associated to a so called *talk page*, where users are solicited to write comments in order to discuss and improve the quality of the article's content. While the large majority of comments is valuable, some users attack others with texts comprising hate speech and harassment, among others.

Our analysis and evaluation are based on the personal attack corpus (Wulczyn et al., 2017) that includes 115,864 comments extracted from Wikipedia talk page comments. Each comment has been labeled by at least ten crowdsourced annotators as an 'attack' (i.e., being abusive) or 'not-attack' (i.e., non-abusive) with an inter-annotator agreement of 0.45 in terms of Krippendorff's α. The label of each comment was aggregated based on the distribution of the labels and the majority vote (about 12% are attacks). The corpus comes with a 60-20-20 split into training, validation, and test set (see Table 1 for corpus statistics).

	Train	Validation	Test
Attack	8,079	2,755	2,880
Not-attack	61,447	20,405	20,298
All	69,526	23,160	23,178

Table 1: Statistics of the personal attacks corpus.

3.2 Abusive Language Lexicon

To carry out our approach, we employ the lexicon of Wiegand et al. (2018). This lexicon has been built through an in-depth examination of negative polar expressions. To this end, a set of candidate abusive words has been collected from the negative polar expressions from the 'subjectivity lexicon' of (Wilson et al., 2005) as well as the frequently listed abusive words in the lexicons surveyed by Schmidt and Wiegand (2017). The expressions in this set have been manually labeled into abusive and non-abusive using a crowdsourcing setting. Based on the resulting labels, a new supervised classifier that distinguishes between abusive and non-abusive expressions has been developed. This classifier, then, has been applied to a large number of negative polar expressions derived from Wiktionary, in order to label them into abusive and non-abusive.

Accordingly, two versions of the lexicon have been created: (1) *the base lexicon* which comprises the manually labeled expressions, and (2) *the expanded lexicon* which includes the automatically labeled expressions in accordance with the predictions of the developed classifier. The first lexicon contains 1650 words and expressions in which 551 of them are abusive, while the second contains 8478 words and expressions with 2989 abusive ones.

The results of using the lexicon for detecting the abusive language in micro-posts demonstrate high effectiveness, particularly in cross-domain settings.

4 Approach

Our approach unravels the search space based on the hypothesis that the differences of abusive texts with and without explicit abusive words are reflected in varying, possibly opposite feature distributions on the lexical, syntactic, semantic, or pragmatic level. In an iterative ablation test step, more domain-specific abusive words are detected.

4.1 Unraveling the Search Space

In contrast to standard approaches training abusiveness classifiers on all examples at once, we propose to apply a three-stage approach.

1) Splitting the Search Space Using an abusive lexicon, we split the training and validation sets into two subspaces of texts containing explicit abusive terms and other texts (see Figure 1(b)).

2) Training Two Abusiveness Classifiers On each training set of the two resulting subspaces (*explicit / other*), a distinct classifier is trained to predict the 'not-attack' probability.

3) Collecting New Abusive Terms Each of the two classifiers is run on 100 random attack and 100 random not-attack texts from the respective validation set ('attack' / 'not-attack' according to ground-truth majority vote). In an ablation test, each word from these selected texts is iteratively removed and the probability of the text to be 'not-attack' is compared to the prediction with that word. The words are then ordered by their "abusiveness" (i.e., words are ranked higher the more their removal raises the 'not-attack' score). Ideally, obfuscated abusive words and sarcastic expressions will be ranked high. The top-k "new" abusive words for each subset (*explicit / other*) and each ground-truth label ('attack' / 'not-attack') are added to the lexicon ($\leq 4k$ words at most per iteration, k being set to 20 after pilot experiments).

4.2 Iterative Unraveling

At the end of an iteration (i.e., splitting the datasets, training two classifiers, and collecting new abusive words), the effectiveness of the classifiers is tested on the validation set. When there is no improvement for three iterations, the process stops.

4.3 Abusiveness Classification

Given an unknown text (e.g., in the test set), we check whether it contains an explicit abusive word from the developed lexicon, and select the appropriate classifier accordingly.

5 Experiments and Results

We compare our approach to the state of the art on the personal attack corpus, following the original suggestion of using the 2-class area under the ROC curve (AUC) and Spearman rank correlation as the evaluation metrics (AUC computed between derived 'attack' probabilities and the corpus majority vote while Spearman considers the fraction of corpus votes agreeing with a prediction).

5.1 Experimental Setup

To represent the state of the art, we employ the best-performing model on the personal attack corpus proposed by Pavlopoulos et al. (2017): an RNN model where the basic cell is a GRU. An embedding layer transforms an input word sequence into a word embedding sequence. Then, the model learns a hidden state from the word embeddings. The hidden state is employed to predict the probability of 'not-attack' using a linear regression layer.

We use 300-dimensional word embeddings (Pennington et al., 2014) pre-trained on the Common Crawl with 840 billion tokens and a vocabulary size of 2.2 million. Out-of-vocabulary words are mapped to one random vector. We use Glorot (Glorot and Bengio, 2010) to initialize the model, with mean-square error as loss function, Adam for optimization (Kingma and Ba, 2014), a learning rate of 0.001, and a batch size of 128.

The initial abusive lexicon used for splitting the search space is the complete set of words in the *base lexicon* of Wiegand et al. (2018) containing 1650 negative polar expressions. This lexicon performed better in our pilot experiments compared to the weakly labeled set of expressions in the *expanded lexicon*.

5.2 Results

On the personal attacks corpus, we compare our approach to the effectiveness reported by Wulczyn et al. (2017) and Pavlopoulos et al. (2017), and to our re-implementation of the RNN model of Pavlopoulos et al. (2017) that forms the basis of our approach (some implementation details missing in the original paper).

As can be seen in Table 2, our approach is slightly better than the re-implementation in terms of AUC and Spearman in both splits and the whole test set. Our approach is on a par with the previous best approach reported (slight AUC improvement to 97.80, but slightly lower Spearman score). The fact that the concatenation of *explicit* and *other* yields a higher AUC than any subspace is a result of the substantially lower predicted probabilities of attack on the *other* set as well as of the highly imbalanced distribution of 'attack' in the two sets.

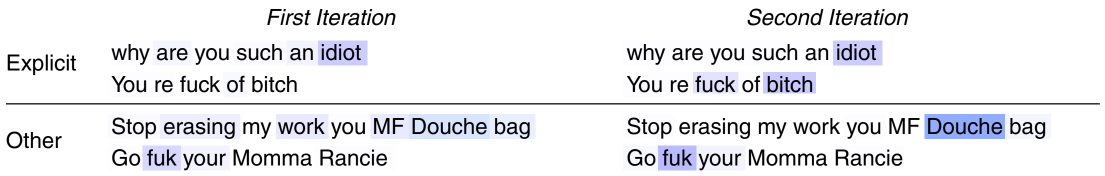

Figure 2: The abusiveness of words in texts with explicit abusive terms (above the line) and without abusive terms (below the line) in the first two iterations. Darker color indicates a higher abusiveness.

Approach	**AUC**	**Spearman**
Our proposed approach		
- all cases	**97.80**	70.26
- explicit	97.69	78.06
- other	97.05	55.37
Reimplementation		
- all cases	97.17	67.98
- explicit	97.08	75.45
- other	96.38	52.06
Pavlopoulos et al. (2017)	97.71	**72.79**
Wulczyn et al. (2017)	96.59	68.17

Table 2: Effectiveness on the test set of the personal attacks corpus (AUC and Spearman coefficients): our proposed approach, the previous state of the art (Pavlopoulos et al., 2017), our reimplementation of it, and the "standard" approach by Wulczyn et al. (2017).

Table 3 shows the AUC values and Spearman coefficients for the first five iterations of our approach on the unraveled validation and test set. The approach stops at the fifth iteration since the highest AUC performance (our target evaluation measure) on *all* and the *explicit* subspace of the validation set was obtained in the second iteration (three failed improvement attempts). The highest AUC for the *other* subspace is achieved in the first iteration, though. The Spearman values increase after each iteration, except again for the *other* subspace where the first iteration works best.

The expansion rates of the abusive lexicon are shown in Table 4. Fewer and fewer terms are added in later iterations since it becomes increasingly less likely for the ablation test to discover important new abusive words. Additionally, we asked two experts to also check the newly added words; they confirmed that more and more abusive terms are added (inter-annotator agreement of 0.59).

Our approach iteratively identifies new "highly abusive" words and moves the respective texts from the *other* subspace to the *explicit* subspace. Since the abusive terms are important clues for the classification, this will force the model for the *other* subspace to utilize new features. As a result, the texts without explicit abusive terms become more "difficult", such that the effectiveness in the *other* subspace decreases over time.

Measure	**1**	**2**	**3**	**4**	**5**
AUC - Valid. all	97.17	**97.46**	97.40	97.34	97.33
AUC - Valid. explicit	96.94	**97.40**	97.21	97.25	97.14
AUC - Valid. other	**97.63**	96.58	96.36	95.46	95.32
AUC - Test all	97.58	**97.80**	97.74	97.68	97.69
AUC - Test explicit	97.25	**97.69**	97.51	97.55	97.55
AUC - Test other	**97.29**	97.05	96.94	94.14	96.15
Spearman - Valid. all	69.19	70.26	70.40	70.25	**70.41**
Spearman - Valid. explicit	76.67	77.43	78.05	**78.47**	78.46
Spearman - Valid. other	**56.88**	54.62	51.64	49.21	47.73
Spearman - Test all	69.73	71.07	71.26	70.87	**71.26**
Spearman - Test explicit	77.38	78.06	78.47	**78.79**	78.59
Spearman - Test other	**57.10**	55.37	53.37	50.50	50.14

Table 3: Effectiveness (AUC values and Spearman coefficients) of our approach's first five iterations.

	1	**2**	**3**	**4**	**5**
Size	1650	1725	1780	1829	1875
Increment		+75	+55	+49	+46
Partially abusive		+20	+30	+24	+18
Abusive		+14	+13	+18	+21
Non-abusive		+41	+12	+ 7	+ 7

Table 4: Increment and of the abusive lexicon in the first five iterations of our approach. The rows *partially abusive*, *abusive*, and *non-abusive* indicate the numbers of abusive words agreed by *one of*, *both*, *none of* the experts in the newly added words respectively.

Table 5 shows the newly found words in each of the first iterations. For every iteration, we show words labeled as 'abusive' (two experts both agree they are abusive), 'partial abusive' (one of the experts agreed they are abusive) and 'non-abusive' (none of two experts both agrees they are abusive). For each label and each iteration, we select three words which have the highest 'abusiveness' (see the definition of 'abusiveness' in section 4.1). We found that our approach can find unusual abusive words (such as 'faggots') and also obfuscated/misspelled abusive words (such as 'fvck').

Figure 2 illustrates some texts with the abusive-

Iteration	Abusive	Partially abusive	Non-abusive
2	jerk fuckheads douchebag	masturbating freak clowns	headline heck nightmare
3	fucking fvck bastard	rudely dunce pederast	hometown lifetime imature
4	bithces sissy fuk	filthy lame harrassing	policemans foot die
5	niggers faggots fuckers	nazi hypocritical imposter	pint boss pay

Table 5: The newly added abusive words in the first iterations. By 'abusive', we refer to the words that both experts label as abusive. By 'partially abusive', we refer to the words that only one of the experts labels as abusive, and by 'non-abusive', we refer to the words that both experts label as non-abusive.

ness of each word in the first and second iteration. The classifier for the *explicit* subspace learns to emphasize the explicit abusive words (e.g., the more important "fuck" or "bitch" and the less important "are" or "an" in the second iteration) while the classifier for the *other* subspace identifies "new" abusive terms (e.g., "Douche" or "fuk") to be added to the lexicon.

6 Conclusion

Abusive language has become a ubiquitous problem on online platforms. Previous work aimed to train detectors on a single search space of potentially abusive texts. In contrast, we suggest to divide the search space into texts containing explicit abusive words (according to a dynamic lexicon) and texts that do not contain such terms. For each subspace, a different classifier is trained.

In an online scenario of consistently running our approach on new comments (some users may report offensive ones, etc.) to support human moderators on online platforms, newly "emerging" obfuscated offensive terms will quickly be spotted and are not "lost" in the dominating space of explicit abusiveness. The iterative extension of the lexicon also helps to increase effectiveness in our experiments showing our approach to be on a par with the previous state of the art on the personal attacks corpus.

Besides matching the previous state-of-the-art "black box" classification performance, our new approach with its dynamic lexicon comes with the benefit of an improved explainability that a human moderator may appreciate for the in-lexicon cases. For the human-in-the-loop platform moderation scenario, we plan a user study also including a functionality to manually add or blacklist terms from the lexicon in each iteration.

References

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 759–760. International World Wide Web Conferences Steering Committee.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Tao Liu, Shengping Liu, Zheng Chen, and Wei-Ying Ma. 2003. An evaluation on feature selection for text clustering. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 488–495.

Yashar Mehdad and Joel R. Tetreault. 2016. Do characters abuse more than words? In *SIGDIAL Conference*, pages 299–303.

Taisei Nitta, Fumito Masui, Michal Ptaszynski, Yasutomo Kimura, Rafal Rzepka, and Kenji Araki. 2013. Detecting cyberbullying entries on informal school websites based on category relevance maximization. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 579–586. Asian Federation of Natural Language Processing.

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*, WWW '16, pages 145–153. International World Wide Web Conferences Steering Committee.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deeper Attention to Abusive User Content Moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*, pages 1125–1135.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10. Association for Computational Linguistics.

Pnina Shachaf and Noriko Hara. 2010. Beyond vandalism: Wikipedia trolls. *Journal of Information Science*, 36(3):357–370.

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

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. 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. Association for Computational Linguistics.

Zeerak Waseem, Thomas, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. *CoRR*, abs/1705.09899.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words–a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 1046–1056.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354. Association for Computational Linguistics.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. *WWW '17 Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399.

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. ACM.