Inducing a Lexicon of Abusive Words – A Feature-Based Approach

Michael Wiegand*, **Josef Ruppenhofer**†, **Anna Schmidt***, **Clayton Greenberg***
*Spoken Language Systems, Saarland University, D-66123 Saarbrücken, Germany
†Institute for German Language, D-68161 Mannheim, Germany
michael.wiegand@lsv.uni-saarland.de
ruppenhofer@ids-mannheim.de
anna.schmidt@lsv.uni-saarland.de
clayton.greenberg@lsv.uni-saarland.de

Abstract

We address the detection of abusive words. The task is to identify such words among a set of negative polar expressions. We propose novel features employing information from both corpora and lexical resources. These features are calibrated on a small manually annotated base lexicon which we use to produce a large lexicon. We show that the word-level information we learn cannot be equally derived from a large dataset of annotated microposts. We demonstrate the effectiveness of our (domain-independent) lexicon in the cross-domain detection of abusive microposts.

1 Introduction

Abusive or offensive language is commonly defined as hurtful, derogatory or obscene utterances made by one person to another person.[1] Examples are (1)-(3). In the literature, closely related terms include *hate speech* (Waseem and Hovy, 2016) or *cyber bullying* (Zhong et al., 2016). While there may be nuanced differences in meaning[2], they are all compatible with the general definition above for abusive language.[3]

(1) stop editing this, you **dumbass**.
(2) Just want to slap the **stupid** out of these **bimbos**!!!
(3) Go lick a pig you arab muslim piece of **scum**.

Due to the rise of user-generated web content, in particular on social media networks, the amount of abusive language is also steadily growing. NLP methods are required to focus human review efforts towards the most relevant microposts.

In this paper, we address the task of detecting abusive words (e.g. *dumbass*, *bimbo*, *scum*). Our main assumption is that abusive words form a subset of negative polar expressions. The classification task is to **filter the abusive words from a given set of negative polar expressions**. We proceed as follows. On a base lexicon that is a small subset of negative polar expressions where the abusive words among them have been marked via crowdsourcing (§3), we calibrate a supervised classifier by examining various novel features (§4). A classifier trained on that base lexicon, which contains 551 abusive words, is then applied to a very large list of unlabeled negative polar expressions (from Wiktionary) to extract an expanded lexicon of 2989 abusive words (§5).

We extrinsically evaluate our new lexicon in the novel task of **cross-domain classification** of abusive documents (§6) where we use it as a *high-level* feature. In this work, we consider microposts as documents. While for in-domain classification, supervised classifiers trained on generic features, such as bag of words or word embeddings, usually score very well, on cross-domain classification they perform poorly since they latch on to domain-specific information. In subjectivity, polarity and emotion classification, high-level features based on predictive domain-independent word lists have been proposed to bridge the domain mismatch (Dias et al., 2009; Mohammad, 2012; Wiegand et al., 2013).

New abusive words constantly enter natural language. For example, according to Wiktionary[4] the word *gimboid*, which refers to an incompetent person, was coined in the British television series *Red Dwarf*, possibly from the word *gimp* and the suffix *-oid*. According to Urban Dictionary[5], the word *twunt*, which is a portmanteau of the swearwords *twat* and *cunt*, has been invented

[1] http://thelawdictionary.org/

[2] For example, several research efforts just focus on utterances addressed towards minorities.

[3] The examples in this work are included to illustrate the severity of abusive language. They are taken from actual web data and in no way reflect the opinion of the authors.

[4] https://en.wiktionary.org

[5] www.urbandictionary.com

Proceedings of NAACL-HLT 2018, pages 1046–1056
New Orleans, Louisiana, June 1 - 6, 2018.

class	adj		noun		verb		all	
	freq	%	freq	%	freq	%	freq	%
abusive	170	33.8	291	45.3	90	17.8	551	33.4
not abusive	332	66.2	352	54.7	415	82.2	1099	66.6

Table 1: The base lexicon: 1650 entries in total of which 551 are abusive.

class	all	intensity (§4.1.1)		views (§4.1.2)	
		weak	strong	actor	speaker
abusive	26.7	14.1	32.0	9.7	32.8
not abusive	73.3	85.9	68.0	90.3	67.2

all numbers only refer to the subset of the base lexicon (Table 1) taken from the Subjectivity Lexicon (i.e. 1500 entries)

Table 3: Percentage of abusive/not abusive instances among (binary) intensity and views.

unlabeled corpora (Table 2). The two larger corpora, the *Amazon Review Corpus – AMZ* (Jindal and Liu, 2008) and the *Web As Corpus – WAC* (Baroni et al., 2009), are used for inducing word embeddings (§4.2). AMZ and the smallest corpus, *rateitall.com – RIA*[8], are used for computing polar word intensity (§4.1.1) from star ratings.

4 Feature Calibration

In the following, we describe the two types of features of our feature-based approach: novel linguistic features and generic word embeddings. They will be examined against some baselines on our base lexicon. As a classifier we use an SVM as implemented in SVM^{light} (Joachims, 1999). We chose that classifier since it is most commonly used for the detection of abusive language (Schmidt and Wiegand, 2017). *For all classifiers in this paper, the supplementary material*[6] *contains information regarding (hyper)parameter settings.*

4.1 Linguistic Features

4.1.1 Polar Intensity (INT)

Intuitively, abusive language should coincide with high polar intensity. We inspect 3 different types.

Binary Intensity (INT$_{bin}$). Our first feature is a simple binary intensity feature we obtain from the Subjectivity Lexicon. In that resource, each entry is categorized as either a *weak* polar expression (e.g. *dirty*) or a *strong* polar expression (e.g. *filthy*). Table 3 (left half), which shows the distribution of intensity on the intersection of our base lexicon and the Subjectivity Lexicon, confirms that abusive words are rarely weak polar expressions and more frequently strong polar expressions.

Fine-grained Intensity (INT$_{fine}$). We also investigate a more fine-grained feature which assigns a real-valued intensity score to polar expressions. It is computed by leveraging the star-rating assigned to the reviews comprising the AMZ corpus (Table 2), a large publicly available review corpus. A review is awarded between 1 and 5 stars where 1 is the most negative score. We infer the polar intensity of a word by the distribution of star-ratings associated with the reviews in which it occurs. We assume negative polar expressions with a very high polar intensity to occur significantly more often in reviews assigned few stars (i.e. 1 or 2). Ruppenhofer et al. (2014) established that the most effective method to derive such polar intensity is by ranking words by their *weighted mean of star ratings* (Rill et al., 2012). All words of our base lexicon are ranked according to that score. As a feature we use the rank of a word.

Intensity Directed towards Persons (INT$_{person}$). Not all negative polar expressions with a high intensity are equally likely to be abusive. The high intensity expressions should also be words typically directed towards persons. Most polar statements in AMZ, however, are directed towards a movie, book or some electronic product. In order to extract negative polar intensity directed towards persons, we replace the AMZ corpus with the RIA corpus (Table 2). RIA contains reviews on arbitrary entities rather than just commercial products as in the case of AMZ. Each review has a category label (e.g. *computer*, *person*, *travel*) that very easily allows us to extract from RIA just those reviews that concern persons.

Table 4 compares a typical 1-star review from AMZ with one from RIA. We consider the RIA-review an abusive comment. It contains many words predictive of abusive language (e.g. *self-absorbed*, *loser*, *arrogant* or *loud-mouthed*).

4.1.2 Sentiment Views (VIEW)

Wiegand et al. (2016b) define sentiment views as the perspective of the opinion holder of polar expressions. They distinguish between expressions conveying the view of the implicit speaker of the utterance typically referred to as *speaker views* (e.g. *cheating* in (4); *ugly* and *stinks* in (5)), and expressions conveying the view of event participants typically referred to as *actor views* (e.g. *disappointed* and *horrified* in (6); *protested* in (7)).

[8]This is a crawl from the review website www.rateitall.com.

of glosses are similar among abusive words, we treat glosses as a bag-of-words feature.

We also exploit information on **word usage**. Many abusive words are marked with tags such as *pejorative*, *derogatory* or *vulgar*. Both WordNet and Wiktionary contain such information. However, in Wiktionary more than 6 times as many of our entries include a tag compared to WordNet.

In order to incorporate a semantic representation more general than individual words, we employ **supersenses**. Supersenses are only contained in WordNet. They represent a set of 45 classes into which entries are categorized. They have been found effective for sentiment analysis (Flekova and Gurevych, 2016). Some categories correlate with abusive words. For example, 76% of the words of our base lexicon that belong to the supersense *person* (e.g. *loser*, *idiot*) are abusive words.

4.1.6 FrameNet (FN)

FrameNet (Baker et al., 1998) is a semantic resource which provides over 1200 semantic frames that comprise words with similar semantic behaviour. We use the frame-memberships of a word as features, expecting that abusive and non-abusive words occur in separate frames.

4.2 Generic Features: Word Embeddings

We induce word embeddings from the two largest corpora, i.e. AMZ and WAC (Table 2) using *Word2Vec* (Mikolov et al., 2013) in default configuration (i.e. 200 dimensions; cbow). The best performance was obtained by concatenating for each word the vectors induced from the two corpora.[9]

4.3 Baselines to Feature-based Approach

In addition to a majority-class classifier we consider the following baselines:

Weak Supervision (WSUP). With this baseline we want to build a lightweight classifier that does not require proper labeled training data. It is inspired by previous induction approaches for sentiment lexicons, such as Hatzivassiloglou and McKeown (1997) or Velikovich et al. (2010) which heuristically label some seed instances and then apply graph-based propagation to label the remaining words of a dataset. On the basis of word embeddings (§4.2), we build a word-similarity graph, where the nodes represent our negative polar expressions and each edge denotes the semantic similarity between two arbitrary words. We compute it by the cosine of their word-embedding vectors. The output of PAT from Twitter (§4.1.4) is considered as positive class seed instances. We chose PAT since it is an effective feature that does not depend on a lexical resource. As negative class seeds, we use the most frequent words in the WAC corpus (Table 2). Our rationale is that high-frequency words are unlikely to be abusive. We chose WAC instead of Twitter since the evidence of PAT (Table 5) suggested less abusive language in that corpus. This word-similarity graph is illustrated in Figure 1. In order to propagate the labels to the unlabeled words from the seeds, we use the Adsorption algorithm (Talukdar et al., 2008).

[9]We also ran experiments with pretrained embeddings from *GoogleNews* but they did not improve classification.

Using Labeled Microposts (MICR). With our last baseline we examine in how far we can detect abusive words by only using information from labeled microposts rather than labeled words. These experiments are driven by the fact that labeled microposts already exist. We consider two methods using the largest dataset comprising manually labeled microposts, *Wulczyn* (Table 8). The class labels of the microposts and our base lexicon (§3) are the same. Our aim is to produce a ranking of words where the high ranks represent words more likely to be abusive. Since we want to produce a strong baseline, we consider the best possible cut-off rank (*see supplementary material*[6]). Every word higher than this rank is considered abusive and all other words not abusive.

The first method **MICR:pmi** ranks the words of our base lexicon by their Pointwise Mutual Information with the class label *abusive* that is assigned to microposts. To be even more competitive, we introduce a second method **MICR:proj** that learns a projection of embeddings. MICR:proj has the advantage over MICR:pmi that it does not only rank words observed in the labeled microposts but all words represented by embeddings. Since our embeddings (§4.2) are induced on the combination of AMZ and WAC corpora, which together are about 360 times the size of the *Wulczyn* dataset, MICR:proj is likely to cover more abusive words. Let $\mathbf{M} = [\mathbf{w_1}, \ldots, \mathbf{w_n}]$ denote a labeled micropost of n words. Each column $\mathbf{w} \in \{0, 1\}^v$ of $\mathbf{M}$ represents a word in a one-hot form. Our aim is learning a one-dimensional projection $\mathbf{S} \cdot \mathbf{E}$ where $\mathbf{E} \in \mathbb{R}^{e \times v}$ represents our unsupervised embeddings of dimensionality e over the vocabulary size v (§4.2) and $\mathbf{S} \in \mathbb{R}^{1 \times e}$ represents the learnt

features used in SVM	Prec	Rec	F1
MAJORITY	33.3	50.0	40.0
INT_{fine}	62.0	57.0	$59.4^{\dagger}$
INT_{bin}	61.7	60.4	61.0^{*}
INT_{person}	70.8	55.4	62.1^{*}
INT_{fine}+INT_{bin}+INT_{person}	70.8	60.7	$65.3^{*\dagger}$
NRC	60.2	60.1	60.2
VIEW	65.6	62.8	$64.2^{\dagger}$
INT_{bin}+NRC+VIEW	66.9	68.8	$67.9^{*\dagger}$
PAT_{noun}	79.9	58.4	67.4
PAT_{noun}+PAT_{adj}	76.4	63.2	69.1
WN_{usage}	82.6	52.6	64.3
FN	66.3	66.4	66.4
WK_{usage}	76.7	61.0	$67.9^{*\dagger}$
WK_{gloss}	74.8	64.9	$69.5^{*\dagger}$
WN_{super}	78.7	64.9	$71.1^{*\dagger}$
WN_{gloss}	75.9	67.4	71.4^{*}
WN_{usage}+WN_{super}+WN_{gloss}	76.7	68.0	72.0^{*}
WK_{usage}+WK_{gloss}	79.5	67.0	72.7^{*}
all WN + *all* WK	80.0	68.7	73.9^{*}
all WN + *all* WK + FN	80.3	69.5	74.5^{*}
all from above	**81.6**	**73.8**	$\mathbf{77.5}^{*\dagger}$

statistical significance testing (paired t-test at $p < 0.05$): *: better than previous line but 1; †: better than previous line

Table 7: Performance of the different linguistic features on base lexicon (Table 1).

are identified by applying to the vocabulary of Wiktionary an SVM trained on the words from the Subjectivity Lexicon with their respective polarities. As features, we use word embeddings (§4.2). In order to produce the feature-based lexicon of abusive words another SVM is trained on our base lexicon (Table 1) using the best feature set from Table 6. With 2989 abusive words, our expanded lexicon is 5 times as large as the base lexicon.

In order to measure the impact of our proposed features on the quality of the resulting lexicon, we devised an alternative expansion which just employs word embeddings. For this, we used **SentProp**, the most effective induction method from the *SocialSent* package (Hamilton et al., 2016).[11]

6 Cross-domain Classification

6.1 Motivation and Set Up

We now apply our expanded lexicon (§5) to the classification of abusive microposts, i.e. we classify entire comments rather than words out of context. Table 8 shows the datasets of labeled microposts that we use. The difference between these datasets is the source from which they originate. Consequently, different topics are represented in the different datasets. Still, we find similar types

dataset	size†	abusive	source
(Warner and Hirschberg, 2012)	3438	14.3%	*diverse*
(Waseem and Hovy, 2016)	16165	35.3%	Twitter
(Razavi et al., 2010)	1525	31.9%	UseNet
(Wulczyn et al., 2017)	115643	11.6%	Wikipedia

†: total number of microposts in the dataset

Table 8: Datasets comprising labeled microposts.

of abusive language (e.g. *racism*, *sexism*). For example, both (10)-(11) from *Waseem* and (12) from *Wulczyn* are sexist comments[12] but (10)-(11) discuss the role of women in sports while (12) addresses women's hygiene in Slavic countries.

(10) *from Waseem dataset:* maybe that's where they should focus? Less **cunts** on *football.*

(11) *from Waseem dataset:* I would rather brush my teeth with sandpaper then watch *football* with a girl!!

(12) *from Wulczyn dataset:* slavic women don't like to wash ... Their **pussy** stinks.

Since our aim is to produce the best possible cross-domain classifier, **all classifiers are trained on one dataset and tested on another**. This is a real-life scenario. Often when a classifier for abusive microposts is needed, sufficient labeled data is only available for other text domains.

Having different topics in training and test data makes cross-domain classification difficult. For example, since a large proportion of sexist comments in *Waseem* relate to sports, traditional supervised classifiers (using bag of words or word embeddings) will learn correlations between words of that domain with the class labels. For instance, the domain-specific word *football* occurs frequently in *Waseem* (i.e. 90 occurrences) with a strong correlation towards abusive language (precision: 95%). Other words, such as *sports* and *commentator*, display a similar behaviour. A supervised classifier will assign a high weight to such words. While such domain-specific words may aid in-domain classification and enable a correct classification of microposts, such as (11), we will show that it has a detrimental effect on cross-domain classification. We claim that the predictive words that abusive comments share across different domains are abusive words, just of the sort that our expanded lexicon contains, e.g. *cunts* in (10) and *pussy* in (12).

Our **proposed classifier** for labeling *microposts* is an SVM trained on features derived from our expanded lexicon (§5). We do not use a binary feature encoding the presence of abusive words. Instead, we rank all abusive words of our lexicon

[11] Since SentProp produces a ranking rather than a classification, we consider 2989 as a cut-off value to separate the instances into 2 classes. This corresponds to the size of abusive words predicted by our feature-based lexicon (Table 9).

[12] (12) is also a racist comment.

datasets					SVM						
						baseline lexicons			newly created lexicons		
test	training	majority	FastText	RNN	Yahoo	Hatebase	Derogat.	Ottawa	base	SentProp	feature-b.
Razavi	Warner	40.50	50.59	53.76	53.40	40.50	40.50	60.95	61.08	64.20	**66.13**
	Waseem	40.50	51.64	53.39	51.66	44.29	51.35	63.13	69.69	63.12	**74.15**
	Wulczyn	40.50	71.74	71.59	**75.10**	40.50	40.50	40.50	40.50	68.50	74.83
	Average	40.50	57.99	59.58	60.05	41.76	44.12	54.86	57.09	66.27	**71.70**
Warner	Razavi	46.14	57.73	48.99	55.42	46.14	57.49	59.81	63.57	**67.57**	64.98
	Waseem	46.14	61.45	57.63	56.54	63.52	57.49	**64.67**	63.57	62.75	64.64
	Wulczyn	46.14	58.35	57.36	60.19	46.14	46.14	46.14	46.14	**65.34**	63.35
	Average	46.14	59.18	54.66	57.38	51.93	53.71	56.87	57.76	**65.22**	64.32
Waseem	Razavi	40.62	60.91	54.67	57.83	40.62	52.66	52.95	57.33	**64.56**	63.32
	Warner	40.62	58.28	58.85	**60.65**	40.62	40.62	40.62	54.93	51.98	58.66
	Wulczyn	40.62	56.33	54.13	51.76	40.62	40.62	40.62	40.62	50.27	**62.90**
	Average	40.62	58.51	55.88	56.75	40.62	44.63	44.73	50.96	55.60	**61.63**
Wulczyn	Razavi	46.88	64.65	64.43	70.70	46.88	50.97	57.70	69.56	67.69	**73.71**
	Warner	46.88	56.21	56.13	52.73	46.88	46.88	55.93	59.55	66.38	**70.06**
	Waseem	46.88	52.66	57.33	51.23	43.51	50.97	60.08	69.56	66.38	**72.39**
	Average	46.88	57.84	59.30	58.22	45.76	49.61	57.90	66.22	63.52	**72.05**

Table 11: Different classifiers on **cross-domain** classification of microposts; best result in **bold**; (*eval.: F1-score*).

which roughly preserves that ratio, includes about 800 adjectives in total. Since abusive adjectives often co-occur with abusive nouns (§4.1.4), they may compensate for abusive nouns that are missing from the lexicon. Such unknown nouns often occur when authors of microposts try to obfuscate their abusive language, e.g. *sneaky assh0le*, *slimy b*st*rd*. Interestingly, the modifying adjectives are not obfuscated, probably because they are considered slightly less offensive in tone.

Given that among the newly created lexicons our feature-based expanded lexicon performs best, we conclude that the expansion is effective (since we improve over the base lexicon), and the features are more effective than a generic induction approach (i.e. *SentProp*).

6.3 Explicitly vs. Implicitly Abusive Microposts

The results in Table 11 also show that the cross-domain performance of our proposed feature-based lexicon is lower on the two datasets *Warner* and *Waseem*. We observed that while on the other two datasets almost all abusive microposts can be considered *explicitly abusive* posts, i.e. they contain abusive words, a large proportion of microposts labeled abusive in *Warner* and *Waseem* are *implicitly abusive* (Waseem et al., 2017), i.e. the abuse is conveyed by other means, such as sarcasm or metaphorical language (11). We asked raters from Prolific Academic to identify explicitly abusive microposts by marking abusive words in those posts. The annotators were <u>not</u> given access to any lexicon of abusive words. We then conducted cross-domain classification on those subsets where the abusive instances were only those rated as explicit. The results are displayed in Table 12. The table shows that our feature-based lexicon is much better on this subset, while the most sophisticated supervised classifier (*Yahoo*) still performs worse. From that we conclude that only *explicitly* abusive microposts can be reliably detected in cross-domain classification.

7 Conclusion

We examined the task of inducing a lexicon of abusive words. We presented novel features including surface patterns, sentiment views, polar intensity and general purpose lexical resources, particularly Wiktionary. The information we thus acquire cannot be learnt all that effectively from labeled microposts, not even with a projection-based classifier. While a lexicon of abusive words can only aid the detection of explicit abuse, its effectiveness was demonstrated on the novel task of cross-domain detection of abusive microposts, where our domain-independent lexicon outperforms previous supervised classifiers which suffer from overfitting to domain-specific features.

Acknowledgements

The authors would like to thank Thomas Kleinbauer, Katja Markert and Ines Rehbein for feedback on earlier drafts of this paper. We are also grateful to William Warner and Diana Inkpen for granting us access to their data on abusive language detection. Special thanks go to Stefan Kazalski for crawling the *rateitall*-website. We also give thanks to John Pavlopoulos for helping us reconstructing the configurations of his RNN. The authors were partially supported by the German Research Foundation (DFG) under grants RU 1873/2-1 and WI 4204/2-1.

Amir Hossein Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive Language Detection Using Multi-level Classification. In *Proceedings of the Canadian Conference on Artificial Intelligence*. Ottawa, Canada, pages 16–27.

Sven Rill, Johannes Drescher, Dirk Reinel, Joerg Scheidt, Oliver Schuetz, Florian Wogenstein, and Daniel Simon. 2012. A Generic Approach to Generate Opinion Lists of Phrases for Opinion Mining Applications. In *Proceedings of the KDD-Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM)*. Beijing, China.

David. E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Parallel distributed processing: explorations in the microstructure of cognition. In *Learning internal representations by error propagation*, MIT Press Cambridge, pages 318–362.

Josef Ruppenhofer, Michael Wiegand, and Jasper Brandes. 2014. Comparing methods for deriving intensity scores for adjectives. In *Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL)*. Gothenburg, Sweden, pages 117–122.

Niloofar Safi Samghabadi, Suraj Maharjan, Alan Sprague, Raquel Diaz-Sprague, and Thamar Solorio. 2017. Detecting Nastiness in Social Media. In *Proceedings of the ACL-Workshop on Abusive Language Online*. Vancouver, Canada.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the EACL-Workshop on Natural Language Processing for Social Media (SocialNLP)*. Valencia, Spain, pages 1–10.

Partha Pratim Talukdar, Joseph Reisinger, Marius Pasca, Deepak Ravichandran, Rahul Bhagat, and Fernando Pereira. 2008. Weakly-Supervised Acquisition of Labeled Class Instances using Graph Random Walks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Honolulu, HI, USA, pages 582–590.

Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan McDonald. 2010. The Viability of Web-derived Polarity Lexicons. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*. Los Angeles, CA, USA, pages 777–785.

William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *Proceedings of the Workshop on Language in Social Media (LSM)*. Montréal, Canada, pages 19–26.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the ACL-Workshop on Abusive Language Online*. Vancouver, BC, Canada, pages 78–84.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL – Student Research Workshop)*. San Diego, CA, USA, pages 88–93.

Michael Wiegand, Christine Bocionek, and Josef Ruppenhofer. 2016a. Opinion Holder and Target Extraction on Opinion Compounds – A Linguistic Approach. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*. San Diego, CA, USA, pages 800–810.

Michael Wiegand, Manfred Klenner, and Dietrich Klakow. 2013. Bootstrapping polarity classifiers with rule-based classification. *Language Resources and Evaluation* 47(4):1049–1088.

Michael Wiegand, Marc Schulder, and Josef Ruppenhofer. 2016b. Separating Actor-View from Speaker-View Opinion Expressions using Linguistic Features. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*. San Diego, CA, USA, pages 778–788.

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/EMNLP)*. Vancouver, BC, Canada, pages 347–354.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the International Conference on World Wide Web (WWW)*. Perth, Australia, pages 1391–1399.

Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting Offensive Tweets via Topical Discovery over a Large Scale Twitter Corpus. In *Proceedings of the Conference on Information and Knowledge Management (CIKM)*. Maui, HI, USA, pages 1980–1984.

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 *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. New York City, NY, USA, pages 3952–3958.