Detecting Derogatory Compounds – An Unsupervised Approach

Michael Wiegand[*0], **Maximilian Wolf**[*], **Josef Ruppenhofer**[†‡]

[*]Spoken Language Systems, Saarland University, Saarbrücken, Germany
[†]Leibniz ScienceCampus, Heidelberg/Mannheim, Germany
[‡]Institute for German Language, Mannheim, Germany
michael.wiegand@lsv.uni-saarland.de
maximilian.wolf@lsv.uni-saarland.de
ruppenhofer@ids-mannheim.de

Abstract

We examine the new task of detecting derogatory compounds (e.g. *curry muncher*). Derogatory compounds are much more difficult to detect than derogatory unigrams (e.g. *idiot*) since they are more sparsely represented in lexical resources previously found effective for this task (e.g. Wiktionary). We propose an unsupervised classification approach that incorporates linguistic properties of compounds. It mostly depends on a simple distributional representation. We compare our approach against previously established methods proposed for extracting derogatory unigrams.

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, they are all compatible with the general definition above.

(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.

A substantial amount of abusive utterances comprises derogatory words (e.g. *bimbo* or *scum*). Automatic extraction methods of such words are required since new derogatory words constantly enter language. Wiegand et al. (2018a) extracted a large list of such expressions and demonstrated its importance for text classification.

In this work, we focus on a subtype of derogatory terms, namely derogatory compounds (e.g. *booze hound*, *curry muncher*, *fault finder*). Distinguishing such multi-word expressions from non-derogatory ones (e.g. *fox hound*, *mile muncher*, *branch finder*) is more difficult than classifying unigrams since they are only sparsely represented in general-purpose lexical resources which have previously been found an effective source from which to learn abusive language, such as Wiktionary.[2] For example, while 97% of the derogatory unigrams of the gold standard lexicon in Wiegand et al. (2018a) are contained in Wiktionary, less than 17% of the derogatory compounds used as our gold standard in this work can be found.

Despite their sparsity in lexical resources derogatory compounds are a frequent phenomenon, particularly in German data, which is why we study this task on that language. On the German benchmark corpus for abusive language detection, the *GermEval corpus* (Wiegand et al., 2018b), we found that of the abusive microposts in the test set that include at least one derogatory expression, 39% contain a derogatory compound.

In our work, we focus on noun-noun compounds. Each compound (e.g. *curry muncher*) comprises two constituents, a modifier (i.e. *curry*) and a head (i.e. *muncher*). On the GermEval corpus, 77% of the derogatory compounds are noun-noun compounds. We only consider compounds whose constituents are not derogatory. 58% of the derogatory compounds on the GermEval corpus fall under this category. Given publicly available lists of derogatory unigrams, the detection of derogatory compounds containing derogatory constituents (e.g. *motherfucker*) is rather trivial.

[0] Present affiliation: Leibniz ScienceCampus, Heidelberg/Mannheim, Germany

[1] http://thelawdictionary.org/

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

Proceedings of NAACL-HLT 2019, pages 2076–2081
Minneapolis, Minnesota, June 2 - June 7, 2019. ©2019 Association for Computational Linguistics

There even exist abusive word generators employing such compounds.[3]

We present the first study to detect derogatory noun-noun compounds and propose an unsupervised classification approach based on distributional information that does not require any properly labeled training data. We demonstrate that linguistic features that have previously been found effective for the classification of derogatory unigrams are notably less effective for the detection of derogatory compounds. We created a new dataset of derogatory compounds which will be made **publicly available**.[4]

Our task is framed as a **binary classification problem**. Each given compound is to be classified out of context as either derogatory or not. For the sake of accessibility, we use English translations of our German compounds in this paper.

2 Related Work

Lexical knowledge for the detection of abusive language has only received little attention in previous work (Schmidt and Wiegand, 2017), the notable exceptions are Razavi et al. (2010) who present a manually-compiled lexicon, Gitari et al. (2015) who bootstrap *hate verbs* and Wiegand et al. (2018a) who induce a lexicon of derogatory words. In all these researches, however, derogatory compounds are not explicitly addressed.

3 Data

We built a **gold standard of derogatory compounds** to train and test classifiers. We inspected a range of websites containing derogatory word lists.[5] Since ambiguity is a massive problem in these lists[6] which makes them hardly usable for abusive language detection (Wiegand et al., 2018a), we manually extracted noun-noun compounds which we considered **unambiguously** derogatory. In order to produce non-derogatory compounds, we randomly sampled from the COW16 corpus (Schäfer, 2015) for each derogatory compound (e.g. *booze hound*) other compounds sharing the same head (e.g. *fox hound*, *stag hound*). Since among those putative non-derogatory instances, there could well be further derogatory compounds, we manually annotated them as well. We limited the set of compounds sharing the same head, which we henceforth call **head group**, to 20 compounds. Thus, we hope to avoid any biases towards particular heads.

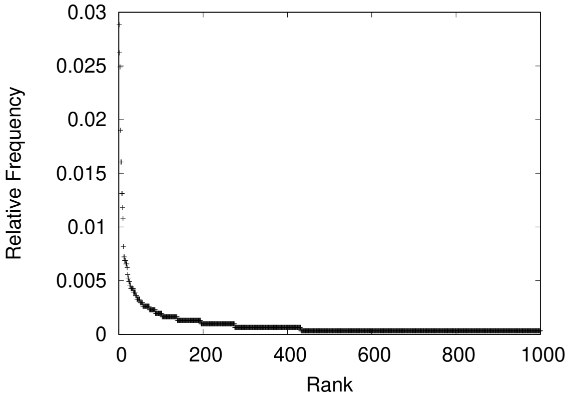

Figure 1: Head distribution of derogatory compounds.

Property	Freq
total compounds	3500
derogatory compounds	382
head groups (*each group contains 20 compounds*)	175
average no. of derogatory compounds in head group	2.2

Table 1: Some statistics of the gold standard.

We also looked at the natural distribution of heads on derogatory compounds. As a proxy we considered the union of all derogatory compounds found on the above websites.[5] Figure 1 plots the frequency rank of the heads against the relative frequency of a particular head. The plot suggests that the heads follow a power-law distribution (Zipf, 1965). As a consequence, one cannot assume that this task could be solved by looking up heads in a finite lexicon with words that often form derogatory compounds in combination with different modifiers.

On a sample of 600 compounds, we measured a substantial agreement of Cohen's $\kappa = 0.61$ (Landis and Koch, 1977) between 2 annotators. Our final dataset (Table 1) comprises 3,500 compounds with only 11% being derogatory.

We also created a **gold standard of derogatory unigram words** in order to examine in how far derogatory compounds can be detected by a classifier trained on derogatory unigrams. For this lexicon, we manually translated the *base lexicon* from Wiegand et al. (2018a) to German.

[3] http://sweary.com

[4] https://github.com/uds-lsv/offensive-compounds

[5] www.hyperhero.com/de/insults.htm
www.schimpfwoerter.de
www.seechat.de/warmduscher.htm

[6] These lists contain many compounds commonly used in a non-offensive manner, e.g. *Colatrinker (coke drinker)*.

4 Method

Our method does not require any labeled training data. In the first step (§4.1), we apply high-precision diagnostics for the detection of derogatory compounds. The output are rankings in which derogatory compounds should be ranked highest. In the second step (§4.2), we combine and re-rank the output of those diagnostics. Our method largely relies on a distributional representation of our compounds. We induced embeddings of our compounds using Word2Vec (Mikolov et al., 2013) on the COW16 corpus, which with its 30B tokens is one of the largest German corpora. Since we exclusively work on German data and German compounds occur as closed compounds, e.g. *Milchbube (milk sop)* or *Schnapsdrossel (booze hound)*, we can employ standard tokenization[7] for inducing embeddings for our compounds.

4.1 Individual High-Precision Diagnostics

Negative Polarity (NEG). Derogatory words form a subset of negative polar expressions. Due to their sparsity, however, derogatory compounds are rarely part of any sentiment lexicon (containing polar expressions). We, therefore, rank all our compounds according to their cosine similarity to a centroid embedding-vector computed from all negative polar expressions from the German PolArt sentiment lexicon (Klenner et al., 2009).

Compound Occurrence vs. Constituent Occurrence (COMCON). Derogatory compounds can be creative word constructions (e.g. *booze hound*, *oxygen thief*, *keyboard warrior*). Consequently, their constituents are often not semantically related. For instance, in *booze hound*, *booze* bears no common semantic relation to *hound*. Therefore, the corpus frequency of a derogatory compound should be much higher than its constituents co-occurring in a sentence (i.e. with other words occurring in between). Such co-occurrences should be coincidental.

We capture this by the following formula (frequencies are computed on the COW16 corpus):

$$COMCON = \frac{\text{\# compound mentions in corpus}}{\text{\# constituents co-occurring in sentence}} \tag{1}$$

In prose, COMCON ranks all compounds by the ratio of observed compound mentions and constituent co-occurrences in a sentence. For derogatory compounds, there should be a high frequency of compound mentions but only a low frequency of the constituents co-occurring in a sentence. Therefore, COMCON will have a high score. While there is a similarly high frequency of compound mentions for non-derogatory compounds, there is also a high frequency of the constituents co-occurring in a sentence since these constituents are usually semantically related (e.g. *landowner* or *circus clown*). This should result in COMCON producing comparably lower scores.

Derogatory Compound Must Be Person (PERSON). We rank our compounds with regard to how likely they represent a person since many non-derogatory compounds represent either objects or animals (e.g. *booze hound* vs. *sight hound*, *fox hound*, *stag hound*). We first compute a centroid vector representing persons. Then, we rank compounds by their similarity to that vector.

As a proxy for persons, we took embeddings of words representing professions, e.g. *banker*, *lawyer*, *salesman*. We also experimented with personal pronouns as a proxy for persons. However, we found them unsuitable since they are also often used as referring expressions to other entities, such as animals. Professions, on the other hand, can only refer to humans. The list of professions we used was created ad-hoc. It should be reproducible in any arbitrary language. *The full list is included in the supplementary notes.*[4]

Outlier Compound(s) in Head Group (OUT). In most head groups, derogatory compounds represent a clear minority with only 1 or 2 compounds. The derogatory compounds are also often semantically different from the non-derogatory compounds (*keyboard warrior* vs. *rajput warrior*, *ninja warrior*, *samurai warrior*). This is particularly true if the non-derogatory compounds are very homogeneous. From that observation we derive a diagnostic in which we determine the semantic outlier(s) for each head group. First, we compute for each compound the average pairwise similarity to all other compounds within its head group. The resulting score of a compound (converted to a dissimilarity score by taking its inverse) is then multiplied by a weight representing the homogeneity of all compounds within that head group.[8] (*Pseudocode is provided in the supplementary notes.*[4]) This is done since for head

[7] Any alphanumeric string separated by spaces is considered a token.

[8] The homogeneity weight is the average pairwise similarity of all compounds belonging to the same head group.

groups that are heterogeneous (e.g. ***legacy hunter***, *job hunter*, *autograph hunter*), there are less obvious outliers.

4.2 Combination and Reranking

Combination (COMB). Negative polarity is a pre-requisite for being derogatory (Sood et al., 2012; Dinakar et al., 2012; Gitari et al., 2015). Therefore, we base our combination on the ranking of NEG. From that ranking we remove all those compounds which have not co-occurred at the high ranks of at least one of the other diagnostics (COMCON, OUT, PERSON).[9] Compounds that are highly ranked by several diagnostics should more likely represent derogatory compounds.

Re-Ranking by PageRank (PRANK). We observed that among the top ranks of COMB, the derogatory compounds are semantically similar (e.g. *dwarf tosser*, *mischief maker*, *slimeball*) while the non-derogatory compounds are semantically different from each other (e.g. *biker club*, *spirit bear*). Therefore, we run personalized PageRank (Agirre and Soroa, 2009) to further improve the ranking by enforcing the compounds on the high ranks to be distributionally similar. We build a word-similarity graph where our compounds are nodes and edges encode cosine-similarities of their embeddings. PageRank then produces a ranking of nodes where the highest ranked nodes are the ones most highly connected. In personalized PageRank prior information is added. A biased graph is constructed in which attention is drawn towards particular regions of interest. This is achieved by assigning re-entrance weights to the individual nodes. As prior information, we set the nodes representing the compounds returned by COMB with a uniform re-entrance weight (α)[10] while all other nodes receive a weight of 0.

Label Propagation (LP). While previous diagnostics were designed to isolate a few derogatory compounds with a high precision, LP aims for increasing recall. We define some high-precision seeds for the two categories of our task and then propagate the labels to the unlabeled compounds by using label propagation (Talukdar et al., 2008). The algorithm operates on the same word-similarity graph that we used for PRANK. We define highly ranked compounds from PRANK as derogatory seeds and lowly ranked compounds as non-derogatory seeds. Unlike the previous diagnostics, the output of LP is a binary categorization rather than a ranking. In order to make this output comparable to the other diagnostics, we converted the output of LP to a ranking. This is achieved by ranking the compounds predicted as derogatory according to the confidence score provided by the classifier.

5 Experiments

Table 2 shows the precision at rank n (P@n) of different rankings as measured on our compound gold standard. For LP, we consider the top 50 compounds from PRANK as derogatory seeds and the bottom 500 as non-derogatory seeds.[11] As a baseline we add a randomized ranking (RAND).

PRANK produces a very high precision on the high ranks, outperforming the individual rankings and COMB. We also tested a modification, PRANK$_{NEG}$, which applies personalized PageRank on the output of NEG, which is the strongest individual ranking. Since PRANK outperforms PRANK$_{NEG}$, we conclude that the high precision of PRANK also depends on the combination of the individual rankings. LP manages to notably raise scores on the lower ranks (e.g. P@300) which proves the advantage of LP over PRANK.

Table 3 compares our proposed method (LP) against supervised classifiers. We evaluate the entire classification output (with F1-measure) rather than a ranking. The classifiers are trained on our unigram or compound gold standard (§3). For the latter case, we conducted 10-fold crossvalidation. 500 of the 3500 compounds were reserved as a development set on which we tuned hyperparameters of the supervised classifiers. (*The supplementary notes*[4] *contain more details.*) As features we consider word embeddings and the linguistic features from Wiegand et al. (2018a). They are based on knowledge that is expensive to produce, such as sentiment views, polar intensity, or information from Wiktionary.[12]

Table 3 shows that learning from the compound gold standard is more effective than learning from the existing unigram gold standard. Given the

[9] We took top 350 from all these rankings which resembles the number of derogatory compounds on our dataset.

[10] Following Manning et al. (2008), we set $\alpha = 0.1$.

[11] The ratio of derogatory and non-derogatory compounds should vaguely reflect the class distribution.

[12] The method WSUP from Wiegand et al. (2018a) was not considered because of its poor performance on compounds.

P@n	RAND	COMCON	PERSON	OUT	NEG	COMB	PRANK$_{NEG}$	PRANK	LP
P@25	12.0	19.2	50.0	60.0	72.0	80.0	96.0	**100.0**	**100.0**
P@50	14.0	20.0	46.0	44.0	62.0	74.0	88.0	**94.0**	**94.0**
P@100	10.0	26.0	40.0	38.0	58.0	68.0	68.0	**82.0**	77.0
P@200	10.0	30.5	31.5	26.5	48.5	54.5	42.0	59.0	**68.5**
P@300	12.3	27.3	27.7	21.7	39.3	44.0	29.3	44.3	**60.3**

Table 2: Comparison of different rankings, evaluated by precision at rank n (P@n). (*PRANK$_{NEG}$ is the ranking applied solely on the output of NEG; PRANK is the ranking applied on the output of COMB.*)

Training	Classifier	Features	F1
unigram	SVM	embeddings+linguistic	63.3
compound	SVM	linguistic	66.1
	SVM	embeddings (*off-the-shelf*)	68.5
	SVM	embeddings	72.4
	SVM	embeddings+linguistic	74.7
none	LP	(embeddings)	73.5

Table 3: Comparison of different classifiers.

Classifier	SVM (embeddings+linguistic)				LSTM
Unit	head	modifier	compound	combined	characters
F1	57.0	60.2	**74.7**	69.0	54.5

Table 4: Comparison of compositional approaches.

strong performance of embeddings, we also examined the performance of (publicly available) off-the-shelf embeddings[13] and found that the high classification scores can be mainly ascribed to the large corpus on which we induced our embeddings (i.e. COW16).

Our unsupervised approach (LP) is almost on a par with the most complex SVM. This is particularly appealing since we produced that classifier without manually labeled training data and those manually-created resources required for the linguistic features.

Compound embeddings are the most predictive information for our task, but even from the large COW16 corpus, we only obtained embeddings for 60% of our compounds.[14] In Table 4, we evaluate compositional information, which can also be used for compounds that lack an embedding. We apply an SVM with the best previous feature set (of which embeddings are the main contributor) on the constituents of the compounds. Moreover, we train an LSTM on the sequence of characters of the compound. Table 4 shows that information drawn from units other than the compound itself is less effective. The feature combination of head, modifier and compound is not effective either.

Instead of applying embeddings on constituents and concatenating them, we also examine a sophisticated compositional model (*Wmask*) based on a masking process that takes into account the variation of a constituent depending on whether it is a head or a modifier (Dima, 2015). Table 5 shows the performance of the two best previous classifiers where compounds lacking an embedding are represented by an embedding approximated by Wmask (rather than a dummy vector). The table shows that the two classifiers can be improved by adding the approximated embeddings.

Classifier	SVM (embed.+ling.)		LP	
Embeddings	plain	+approx.	plain	+approx.
F1	74.7	75.7	73.5	74.9*

Table 5: Compound embedding augmentation (**: statistically better than the plain classifier using a paired t-test at $p < 0.05$*).

6 Conclusion

We examined the new task of detecting derogatory compounds and proposed an unsupervised approach incorporating linguistic properties of compounds that mostly depend on a distributional representation. Our method outperforms linguistic features previously shown to be effective for the detection of derogatory unigrams and it is on a par with a far more expensive state-of-the-art supervised approach. Features defined on the constituents of a compound and training a classifier on derogatory unigrams are far less effective.

Acknowledgements

The authors would like to thank Ines Rehbein for feedback on earlier drafts of this paper. We are also grateful to Corina Dima for helping us to run her semantic composition toolkit *wordcomp* on our data. The authors were partially supported by the German Research Foundation (DFG) under grants RU 1873/2-1 and WI 4204/2-1.

[13] We took the publicly available embeddings induced on Twitter-data from: www.spinningbytes.com/resources/wordembeddings/

[14] For the remaining compounds, we used dummy vectors.

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL)*, pages 33–41, Athens, Greece.

Corina Dima. 2015. Reverse-engineering Language: A Study on the Semantic Compositionality of German Compounds. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1637–1642, Lisbon, Portugal.

Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems*, 2(3):18:1–18:30.

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):2015–230.

Manfred Klenner, Angela Fahrni, and Stefanos Petrakis. 2009. PolArt: A Robust Tool for Sentiment Analysis. In *Proceedings of the Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 235–238, Odense, Denmark.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Christopher D. Manning, Hinrich Schütze, and Prabhakar Raghavan. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

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.

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*, pages 16–27, Ottawa, Canada.

Roland Schäfer. 2015. Processing and querying large web corpora with the COW14 architecture. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC)*, pages 28–34, Lancaster, United Kingdom.

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)*, pages 1–10, Valencia, Spain.

Sara Owsley Sood, Elizabeth F. Churchill, and Judd Antin. 2012. Automatic identification of personal insults on social news sites. *Journal of the Association for Information Science and Technology*, 63(2):270–285.

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)*, pages 582–590, Honolulu, HI, USA.

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)*, pages 88–93, San Diego, CA, USA.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018a. Inducing a Lexicon of Abusive Words – A Feature-Based Approach. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 1046–1056, New Orleans, LA, USA.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018b. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of the GermEval Workshop*, pages 1–10, Vienna, Austria.

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)*, pages 3952–3958, New York City, NY, USA.

George K. Zipf. 1965. *The Psycho-Biology of Language*. MIT Press.