Supplementary Notes to *Exploiting Emojis for Abusive Language Detection*

September 30, 2020

1 Introduction

This document provides more detailed information regarding certain aspects of our research for which there was not sufficient space in the main paper. We focus on the following aspects:

- details regarding the choice of emojis (§2)
- details regarding word embeddings (§3)
- details regarding creating the vocabulary of negative polar expressions (§4)
- details regarding classifiers used for the induction of abusive lexicons (§5)
- details regarding classifiers used for cross-domain classification of abusive microposts (§6)
- details regarding noise reduction for inducing lexicons of non-English abusive words (§7)
- details regarding replicating the resource-intensive lexicon of abusive words from Wiegand et al. [2018a] for German and why it is not possible to produce a comparable resource for Portuguese (§8)
- motivation for the non-English datasets chosen in our crosslingual experiments (§9)
- details regarding crosslingual classification using BERT (§10)
- details regarding the creation of the gold standard for disambiguating potentially abusive words (§11)
- details regarding classifiers and features used for contextual disambiguation of potentially abusive words (§12)

2 Emojis on Twitter

We manually selected a set of emojis that have some abusive connotation and that we also observed with abusive language. We used Twitter as textual data since it is known to contain a significant amount of both emojis and abusive language. Despite the variety of different emojis[1], there is only a smaller fraction used regularly on Twitter. For instance, the dataset from Zampieri et al. [2019] includes less than 10% of the overall existing emojis (about 300 out of 3,000 emojis). Our final choice only includes emojis for which we could obtain a significant amount of tweets (i.e. several thousand tweets) after running a query containing these emojis using the Twitter-streaming API for a few days. A few emojis, such as 💣 (bomb) or ⚡ (high voltage), had to be excluded from our study, simply because they were too rare.

Since our observation was that some emojis are used interchangeably and some of the individual emojis suffered from data sparsity particularly on non-English data, we decided to group (near-)synonymous emojis. Table 1 displays the resulting grouping of emojis. Throughout our research described in the main paper, we always used emojis based on the grouping in that table.

3 Choice of Word Embeddings

For our projection-based approach to induce a lexicon of abusive words with the help of emojis, we needed a pre-trained set of word embeddings.

For English, we chose GloVe embeddings [Pennington et al., 2014] induced on Twitter. We chose these embeddings since they are the most widely used embeddings that have been induced from Twitter. The underlying text source is important. It should be a text type in which abusive words regularly occur. Otherwise, we do not obtain a reasonable vector representation to start with. Most other pre-trained embeddings have been induced from other text types, such as Wikipedia, which we consider less suitable. Abusive words are much more unlikely to be observed. Moreover, the text type should resemble the texts that are going to be classified which, in our research, are mostly tweets.

For our experiments on Portuguese and German, we considered slightly different types of embeddings:

First, we induced embeddings using word2vec [Mikolov et al., 2013] rather than GloVe. We deliberately chose a different induction algorithm in order to show that the induction method is not really critical for the projection-based approach to work. The word2vec-toolkit is also widely used and has the advantage to run fast even on larger corpora.

Second, we could not use pre-trained word embeddings that originate from Twitter because, for many languages, there is no such resource available and we opted for a setting representative for most languages. Inducing embeddings from a Twitter corpus from scratch is no option either since this requires a Twitter corpus. We are not aware of any suitable corpus of that kind that would be sufficiently large and also sufficiently unbiased.[2] We chose Web As

[1] `https://unicode.org/emoji/charts/full-emoji-list.html`

[2] One can easily extract tweets from Twitter for almost any known language. However, one typically has to extract tweets with the help of query words which heavily distorts the word distribution in the resulting set of tweets that are returned. Such a set of tweets cannot be

group	emoji	unicode	description
angry		U+1F620	angry face
		U+1F621	pouting face
		U+1F92c	face with symbols on mouth
fist		U+1F44A	oncoming fist
middle finger		U+1F44A	middle finger
monkey		U+1F412	monkey
		U+1F1435	monkey face
		U+1F449	hear-no-evil monkey
		U+1F648	see-no-evil monkey
		U+1F64A	speak-no-evil monkey
		U+1F698D	gorilla
pistol		U+1F52B	pistol
poo		U+1F4A9	pile of poo
skill		U+1F480	skull
		U+2620	skull and crossbones
vomit		U+1F922	nauseated face
		U+1F92E	face vomiting

Table 1: Grouping of Emojis.

Corpus [Baroni et al., 2009, Filho et al., 2018] as a source from which to induce embeddings, as we consider this text type to contain a sufficient amount of informal language and thus also abusive words.

Regarding the configuration of the word embeddings, we basically used the default configuration of word2vec. We chose 200 dimensions since we considered it the most widely used setting. Irrespective of the induction tool, the number of dimensions is typically chosen to lie between 100 and 300 dimensions.[3] It may well be that the performance of our projection-based approach is even stronger with a different configuration. However, we refrained from tuning parameters to our data since we wanted to avoid overfitting.

4 Creating a Vocabulary of Negative Polar Expressions

The basis of our induction-based approach is a large vocabulary of negative polar expressions. We describe this process in this section, despite the fact that, for English, we could re-use the set introduced by Wiegand et al. [2018a].[4] We felt this description is necessary, since one has to repeat this process for every new language one considers. In our case, we had to replicate this method on Portuguese and German data.

4.1 The General Procedure

The general procedure is to train a feature-based classifier, such as an SVM, on polar expressions of a sentiment lexicon. As features, Wiegand et al. [2018a] propose to use the word embeddings (§3) representing each respective polar expression. The class labels represent the polarity labels of the polar expressions assigned by the sentiment lexicon, i.e. either *positive*, *negative* or *neutral*. The resulting classifier is subsequently run on a very large set of words. Wiegand et al. [2018a] propose to extract this set from Wiktionary. The words predicted as negative are considered as the final vocabulary of negative polar expressions.

The resulting vocabulary of negative polar expressions has to fulfill a set of requirements so that the induction approach is really effective. On the one hand, the vocabulary has to be sufficiently large so that abusive words are a proper subset. This is why Wiegand et al. [2018a] chose as a text source Wiktionary since, unlike other general purpose lexical resources, it describes a substantial proportion of informal language (of which abusive language is a subset). On the other hand, since the lexicon we want to induce should be effective across different domains, the vocabulary should be fairly domain-independent and comprise unambiguously negative polar expressions. Therefore, all proper nouns from the vocabulary are excluded in a pre-processing step. It may well be that some proper nouns are used as abusive words in some context, e.g. *You are just like Hitler*. However, previous work established that in abusive language detection,

considered as a representative sample of Twitter. Therefore, the word embeddings induced on them would also be very biased.

[3] `https://stackoverflow.com/questions/26569299/word2vec-number-of-dimensions` or `https://fasttext.cc/docs/en/unsupervised-tutorial.html`

[4] The *expanded lexicon* of that work does not only comprise the abusive words learnt by the proposed method but also all negative polar expressions predicted to be non-abusive.

machine learning methods are susceptible to learn spurious correlations. For instance, Wiegand et al. [2019a] points out that on the dataset from Kumar et al. [2018], classifiers typically learn Arabic person names as abusive words. Moreover, the set of tweets from which we learn abusive words, will be created fairly ad-hoc in the sense that we do not make any attempts to balance the underlying word distribution. So, if we did include also proper nouns in our vocabulary, our lexicon induction approach would erroneously extract proper nouns currently co-occurring with our target emojis, e.g. *Trump*.

While most common sentiment lexicons exclusively focus on positive and negative polar expressions, the lexicon which was used by Wiegand et al. [2018a], i.e. the *Subjectivity Lexicon* [Wilson et al., 2005], also included *neutral* expressions. Since this third class seems to be vital in order to produce a reliable list of negative polar expressions, we also had to provide training data for neutral expressions when we replicated this classification approach on Portuguese and German data. As neutral expressions, we simply extracted high-frequency (content) words, i.e. nouns, verbs and adjectives, from corpora such as Web As Corpus, that are not in the list of positive and negative polar expressions of the given sentiment lexicon.

4.2 Creating the German Vocabulary of Negative Polar Expressions

The procedure proposed by Wiegand et al. [2018a] can be largely adapted to German. As a sentiment lexicon, we used the German version of the PolArt-lexicon [Klenner et al., 2009]. Access to the German part of Wiktionary was enabled by the JWKTL API [Zesch et al., 2008]. As a result, we obtained about 14,000 negative polar expressions. This set of negative polar expressions is considerably larger than its English counterpart (with only about 7,000 expressions) because of the high proportion of lexicalized compounds in German [Wiegand et al., 2016].

4.3 Creating the Portuguese Vocabulary of Negative Polar Expressions

For Portuguese, we faced the problem that we could not derive the input for polarity classification from Wiktionary since the API we used, i.e. JWKTL [Zesch et al., 2008], does not enable access to the Portuguese part of that resource. The only alternative API for Wiktionary, i.e. WiktionaryParser[5], has a much more restricted functionality that does not enable quick access to the complete set of all entries in a particular language from Wiktionary. As an alternative to obtain a sufficiently large set of Portuguese words, we chose OpenWordNet–PT [de Paiva et al., 2012]. As a Portuguese sentiment lexicon, we chose the Portuguese lexicon within the *Sentiment Lexicons for 81 Languages*.[6]

[5] `https://github.com/Suyash458/WiktionaryParser`

[6] `www.kaggle.com/rtatman/sentiment-lexicons-for-81-languages`

5 Classifiers for Inducing Abusive Words

5.1 Ranking Words by Learning a Projection

For the projection-based approach to learn abusive words by ranking a set of negative polar expressions according to their occurrence in microposts containing specific emojis, we modified the code of the implementation of *Non-Linear Sub-Space Embedding (NLSE)* [Astudillo et al., 2015]. Our modification meant removing the hidden layer from NLSE and reducing the dimensionality of the projected embeddings to a scalar. All remaining components of that software were left to their default configuration.

5.2 Re-Ranking using Label Propagation

We made use of the Adsorption propagation algorithm as implemented in *junto* [Talukdar et al., 2008]. For all different sets of seeds we experimented with, we only consider the default configuration of that tool.

6 Classifiers for Cross-Domain Micropost Classification

For our experiments on English cross-domain microposts, we meticulously adhered to the configuration of classifiers used in Wiegand et al. [2018a]. We refer the reader to the supplementary notes of that paper[7] for more specific details.

6.1 Classifiers from Wiegand et al. [2018a]

In our evaluation, we also considered two classifiers from Wiegand et al. [2018a], which are the feature-based approach from Nobata et al. [2016] and the lexicon-based classifier using the induction approach proposed by Wiegand et al. [2018a]. Since we added a further dataset in our evaluation, namely the dataset from Zampieri et al. [2019], we had to replicate these classifiers for the new dataset. Regarding the feature-based approach, we follow the implementation used in Wiegand et al. [2018a] which is an approximation of Nobata et al. [2016]. The original implementation of Nobata et al. [2016] cannot be replicated since several resources are not publicly available. More details regarding the approximation can be found in the supplementary notes from Wiegand et al. [2018a].

6.2 BERT

In addition to the classifiers examined in Wiegand et al. [2018a], we additionally use BERT [Devlin et al., 2019] as a cross-domain supervised classifier. As a model, we used BERT-Large, Uncased: 24-layer, 1024-hidden, 16-heads, 340M parameters.[8] We deliberately used the *uncased* variant since all other classifiers in our work also employ a lowercase text representation. We felt that using the *cased* variant of BERT would not mean a fair comparison to the other classifiers used in this submission. For classification, we fine-tuned the BERT model on the

[7] `https://github.com/uds-lsv/lexicon-of-abusive-words/supplementaryNotes.pdf`
[8] `https://github.com/google-research/bert`

respective training data by adding another layer on top of the existing model. The model was trained with standard hyperparameter settings: batch size: 32; learning rate: 2e-5; number of epochs: 3.

We had difficulties with training BERT on small datasets where, due to the class imbalance of abusive and non-abusive tweets, BERT tends to produce a majority-class classifier. Therefore, we adjusted the training data based on the two smallest datasets (i.e. the datasets by Razavi et al. [2010] and Warner and Hirschberg [2012]) by up-sampling the number abusive microposts to the number of non-abusive microposts. This resulted in classifiers being stronger than a majority-class classifier. This modification of the training data was only carried out in order to produce a stronger baseline for the small datasets. Our aim for the classifiers based on BERT was to produce the strongest classifiers possible.

For the other classifiers in our cross-domain experiments, we maintained the original class distribution of the datasets both as training and test data since this is the setting from Wiegand et al. [2018a]. In general, we wanted to replicate that setting as closely as possible.

7 Noise Reduction for Inducing Lexicons of Non-English Words

We observed that due to the data sparsity of non-English data, our emoji-based ranking of abusive words for Portuguese and German was less accurate than the ranking we obtained for English. In order to remove some noise, we applied the noise-reduction method proposed by Wiegand et al. [2019b] based on PageRank [Agirre and Soroa, 2009]. It produced some slight performance increases. However, we observed it is not essential when end-to-end output is evaluated. That is why, we omitted the description of noise reduction from the main paper. For the sake of full transparency, however, we briefly describe the method here:

PageRank operates on a word-similarity graph, where the nodes are words to be ranked and the edges encode cosine-similarities of their embeddings. It produces a ranking of nodes where the highest ranked nodes are the most highly connected ones. These nodes correspond to distributionally similar words, i.e. hopefully abusive words. (We assume abusive words to be in the majority on the high ranks and more distributionally similar than the false positives.) 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 different damping factors to the individual nodes. As prior information, we assign a uniform damping factor (α) to the nodes representing the words returned by the higher ranks of our projection (we chose top 1000) while all other nodes receive a value of 0.[9]

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

8 Replicating the Resource-based Abusive Lexicon from Wiegand et al. [2018a] on Other Languages

The induced lexicon from Wiegand et al. [2018a] currently represents the most effective lexicon of abusive words for cross-domain classification of microposts on English data. Therefore, we also wanted to replicate this resource on the other languages we considered in our work, i.e. German and Portuguese. While it was possible to do so for German, it was not for Portuguese since a significant number of resources are not (publicly) available in that language. In §8.1, we describe how we replicated the approach on German data. In §8.2, we explain what type of resources posed obstacles for replicating this lexicon on Portuguese data.

8.1 Replication on German

The induction approach by Wiegand et al. [2018a] comprises two steps. In the first step, a base lexicon has to be produced by manual annotation (§8.1.1). The manual lexicon comprises negative polar expressions labeled as either abusive or non-abusive. In the second step, a supervised classifier is trained on the *base lexicon*. This classifier is run on a large set of negative polar expressions. The negative polar expressions predicted as abusive form the final lexicon of abusive words. This lexicon is also referred to as *expanded lexicon* (§8.1.2).

In the following, we detail the two steps we reconstructed on German language data.

8.1.1 Base Lexicon

The base lexicon had already been produced by previous research [Wiegand et al., 2019a] by manually translating the English base lexicon to German. Manual translation was preferred over automatic translation since the German words should preserve the degree of abusiveness of the original English words. (This cannot be achieved by automatic translation.) Unfortunately, a substantial number of English abusive words could not be translated into German (about 60 from 1650) as, due to cultural differences, some English words simply have no German counterpart. For instance, the word *spic* refers to a Spanish-speaking person from Central or South America or the Caribbean, especially a Mexican. Such persons represent a minority in North America that are frequently verbally offended. Since in Germany, Spanish-speaking persons do not represent a typical minority, there is no appropriate German translation for the word *spic*.

8.1.2 Expanded Lexicon

For the expanded lexicon, we used as input the large German lexicon of negative polar expressions (§4). On this lexicon, we ran an SVM classifier trained on features proposed by Wiegand et al. [2018a]. How these features were adapted to German is described in the following subsection.

8.1.3 How were the linguistic features of the supervised classifier replicated?

Polar Intensity. Wiegand et al. [2018a] proposed 3 different approaches to determine the polar intensity of words. They are:

1. A lexicon look-up using a sentiment lexicon that contains binary intensity information.

2. A derivation of polar intensity scores from the distribution of star ratings of reviews using a standard review corpus.

3. A derivation of polar intensity scores from the distribution of star ratings of reviews using a special review corpus that exclusively contains reviews that address persons. (The authors propose the usage of a crawl from the rateitall-website.[10])

Unfortunately, for German data, we are not aware of any sentiment lexicon containing polar intensity information which meant that Approach 1 could not be replicated. We refrained from automatically translating this resource since our intuition is that translations of polar expressions not necessarily preserve the same level of polar intensity in the target language.

There are only very few review corpora in German that also contain star rating information. We are not aware of any review corpus which allows a reliable isolation to reviews addressing persons. Therefore, we could only replicate a method to compute the polar intensity according to Approach 2. We use the corpus by Prettenhofer and Stein [2010] for this purpose.

Sentiment Views. There does not exist a complete German sentiment lexicon with sentiment views. However, there exist subsets manually annotated from the German PolArt sentiment lexicon [Klenner et al., 2009]. Wiegand and Ruppenhofer [2015] manually annotated all sentiment verbs from that lexicon while Wiegand et al. [2016] provided a manual annotation of atomic German sentiment nouns from the same resource. We merged these two subsets to one lexicon and extracted the binary sentiment-view feature with the help of this combined resource.

Emotion Categories. We used the German version of the NRC lexicon [Mohammad and Turney, 2013] (Version 0.92). This version comprises the same emotion categories as the English lexicon and the same vocabulary (but translated from the original English lexicon into German).

WordNet and Wiktionary. For the features derived from WordNet [Miller et al., 1990], we used the German version of WordNet called *GermaNet* [Hamp and Feldweg, 1997]. The design of GermaNet largely follows the English original resource. Wiktionary also encompasses languages other than English including German. Therefore, the WordNet and Wiktionary features proposed by Wiegand et al. [2018a] can be replicated in a straightforward manner on the corresponding German-language resources.

Surface Pattern. In principle, the surface pattern for English proposed by Wiegand et al. [2018a] (1) can be replicated to German (3).

(1) *English pattern*: called me a(n) <`noun`>

[10] `www.rateitall.com`

(2) *English pattern match example*: He called me a **tosser**.
(3) *German pattern*: hat mich ein(e|en) <noun> genannt
(4) *German pattern match example*: Er hat mich einen **Vollidioten** genannt.

However, the pattern is far sparser than the English one. If one runs the German pattern as a query on Twitter and extracts all matching tweets coming in a time period of 14 days, like Wiegand et al. [2018a] proposed, only 25 different words are extracted and all except one are only observed once. Such infrequent words are not sufficiently reliable. From that one can conclude that the surface pattern could also work in German in principle. However, the period for streaming tweets from Twitter matching the query pattern would need to be significantly extended (i.e. several months or even a year), which was beyond the scope of our research.

As a consequence, we had to exclude this feature from the feature set to extract German abusive words. It also meant that the weakly-supervised method from Wiegand et al. [2018a] (WSUP) could not be re-implemented either as it requires as input abusive words that are obtained with the help of the surface pattern.

FrameNet. Although there exists a German equivalent to the English FrameNet [Baker et al., 1998], called Salsa-corpus [Burchardt et al., 2006], we refrained from using it for this work, since the German version is considerably smaller and mostly focuses on verbs (a few hundred words). We could hardly identify any offensive words among the lexical units that had been annotated for the German FrameNet. This does not come as a surprise since the most frequent part of speech in that resource, i.e. verbs, is known to yield only a very small fraction of abusive words [Wiegand et al., 2018a].

8.2 Why we could not induce a lexicon of abusive words for Portuguese analogously to the English approach

In the following, we list the significant obstacles we faced while trying to produce a lexicon of abusive words following the approach by Wiegand et al. [2018a] on Portuguese data:

8.2.1 Base Lexicon

The starting point for the approach by Wiegand et al. [2018a] is a manually created base lexicon in which a sample of negative polar expressions are annotated as either abusive or non-abusive. While for German, there exist such a resource from previous work [Wiegand et al., 2019a], there is no such resource for Portuguese. A quick automatic translation is not an option here. As our experiments using state-of-the-art machine translation clearly indicated (see Table 8 in the main paper), automatic translation of abusive words results in poor performance. The main reasons are ambiguity (an unambiguously abusive word in the source language may be ambiguous in the target language; ambiguously abusive words have been shown to be detrimental in micropost classification [Wiegand et al., 2018a]) and that for many abusive words, there exist no direct translations in the target language (see discussion in §8.1.1).

8.2.2 Sentiment Views

Unlike German, we are not aware of any Portuguese sentiment lexicons that include annotation with regard to sentiment views.

8.2.3 Polar Intensity

Unlike German, we are not aware of any publicly available resources for Portuguese that explicitly encode polar intensity information or that allow polar intensity information to be derived using the methods proposed by Wiegand et al. [2018a].

8.2.4 Wiktionary

While for German, we could use JWKTL [Zesch et al., 2008] as an API to access information from German Wiktionary (similar to how Wiegand et al. [2018a] used that API to access English data), this API currently does not support Portuguese. Although there is an alternative API that also allows access to the Portuguese Wiktionary called WiktionaryParser[11], its functionality is considerably more restricted. A fundamental difference is that while JWKTL operates on a local Wiktionary dump allowing efficient traversal of all lexical entries of a particular language, WiktionaryParser retrieves lexical entries from the online version. This makes processing considerably slower.

8.2.5 FrameNet

Already the German version of FrameNet was too small for being incorporated in the induction approach. Since the Portuguese version[12] is even much smaller than the German version (just 32 frames and 38 lexical units are available), we could not use this resource for Portuguese either.

9 Motivation for the Non-English Datasets Chosen for our Crosslingual Experiments

We chose German and Portuguese as languages for our crosslingual experiments primarily because there are datasets for abusive language detection in these languages available. While there are also datasets for this task in other languages, most prominently Spanish [Álvarez-Carmona et al., 2018, Fersini et al., 2018b] and Italian [Bosco et al., 2018, Fersini et al., 2018a], the raw data of these alternative datasets, which in all cases are tweets from Twitter, have been extracted by querying the Twitter-streaming API with a set of abusive words. While this results undoubtedly in a dataset with a significant proportion of abusive language, the lexical diversity is rather limited. Typically, on such datasets the abusive words are those of the initial queries. Since our approach aims for capturing the lexical diversity of abusive language, those datasets seem inappropriate for our evaluation.

The Portuguese [Fortuna et al., 2019] and German [Wiegand et al., 2018b, Struß et al., 2019] datasets, on the other hand, which also represent tweets from

[11] `https://github.com/Suyash458/WiktionaryParser`
[12] `https://framenet.icsi.berkeley.edu/fndrupal/framenets_in_other_languages`

Twitter, have been constructed by looking for user profiles that tend to contain abusive language. On such datasets, there is a much smaller lexical bias than on those datasets created via abusive words. The construction process of the German datasets also includes many measures to avoid any other kind of data bias which has been reported on several datasets for abusive language detection [Wiegand et al., 2019a].

Another reason for choosing these particular datasets is that they comprise different subtypes of abusive language, while many of the other datasets just focus on one particular subtype, e.g. misogyny [Fersini et al., 2018a,b]. Specific subtypes are not the focus of our study.

10 Creating Crosslingual classifier with multilingual BERT

We use the multilingual version of BERT in the same fashion we also used the monolingual version of BERT (§6.2). The only difference is that our training and test data differ in the language (i.e. training: English; test: Portuguese or German) and the underlying BERT model. We thus follow the methodology proposed by Pires et al. [2019]. As a multilingual model, we use the model recommended by Google research: BERT-Base, Multilingual Cased for 104 languages using 12-layer, 768-hidden, 12-heads, 110M parameters.[13] (Notice that there is no multilingual version of BERT-Large available yet.)

All classifiers other than BERT used in our crosslingual experiments are lexicon-based experiments. BERT is the only supervised classifier. In order to produce a very strong baseline classifier using BERT, we adjusted for this supervised classifier the class distribution of the training data to the class distribution of the test data.

In our exploratory experiments, we also tested a classifier based on crosslingual embeddings as a more lightweight alternative to multilingual BERT. These experiments were carried out on English and German data. For inducing crosslingual embeddings, we used MUSE [Conneau et al., 2017] from Facebook-research. As a supervised classifier that uses those embeddings, we used FastText [Joulin et al., 2017]. Both MUSE and FastText are tools that work well in their default configuration. The monolingual embeddings were induced on the English and German version of Web as Corpus [Baroni et al., 2009] using word2vec [Mikolov et al., 2013] in its default configuration.[14] Despite using the most advanced alignment method with MUSE, namely a method which starts with a bilingual dictionary, we obtained lower scores with the classifier based on multilingual embeddings than with multilingual BERT. Therefore, we only used the stronger multilingual classifier, i.e. BERT, in the paper (due to lack of space).

[13] `https://github.com/google-research/bert/blob/master/multilingual.md`

[14] Although our test data are tweets from Twitter, we deliberately decided against taking embeddings from Twitter corpora. The reason is that we are not aware of existing word embeddings for both English and German data that have been induced from similar types of Twitter corpora. Neither are we aware of any publicly available Twitter corpora that would have allowed us to induce the word embeddings ourselves.

11 Gold Standard for Disambiguating Potentially Abusive Words

All our manual annotation for this research was produced with the help of crowdsourcing using *Prolific Academic*.[15] In *all* experiments, we had each instance rated by 5 different annotators. The most important requirements we specified for annotators in order to be eligible to participate in our survey were the following:

- Annotators had to be a native speaker of English.
- Annotators had to have a task approval rate of 90% or higher. (At the time we ran our experiments, this was the highest threshold possible on Prolific Academic.)
- Annotators were not allowed to have any literacy difficulties.

For each annotation task we also produced some annotation guidelines (*they are also included in the supplementary data*). Annotators had to read those guidelines before the actual annotation. The guidelines were also made available as a separate pdf-file, so that the annotators could always refer to them during their annotation.

All guidelines were illustrated with examples. We also used these examples as test instances. They were randomly interspersed with the regular instances to be annotated. However, they were not marked as such so that the annotators could not recognize them. We only informed them in the guidelines that we had included test instances. The order of the instances presented to the annotators was randomized. If a significant proportion of test instances was incorrectly labeled by some annotator, then this meant that the annotator either had not read the annotation guidelines or had not understood the annotation task. Subsequently, the annotation of this annotator was rejected and not included in our final gold standard.

In all annotation tasks, the set of instances was organized in bins of about 100 to 200 instances. Each bin was considered a separate task posted on *Prolific Academic*. We felt that a larger amount of instances would result in a loss of concentration on the part of the annotators. The tweets to be annotated via crowdsourcing were pre-filtered by a Native speaker of English who excluded tweets that were spam or totally incomprehensible.

In order to have the most unbiased annotation possible, each annotator from *Prolific Academic* was only admitted to one single annotation task of our work (i.e. one bin).

12 Classifiers for Contextual Disambiguation of Abusive Words

In our experiments for contextual disambiguation, we employ two different types of supervised classifiers: text classifiers and word-specific classifiers. The first type of classifier is a generic classifier in which complete microposts are processed with no specific focus on any words occurring in them. The second type

[15] `www.prolific.ac`

of classifier, on the other hand, comprises task-specific features that aim to disambiguate a particular target word (i.e. *fuck* or *bitch*). We describe the two classifiers in the two following subsections.

For all supervised classifiers on the task of contextual disambiguation of abusive words, we adjusted the class distribution of the training data to the class distribution of the test set. This adjustment of the training data results in much stronger performance on *all* chosen supervised classifiers including the baseline classifiers. It is common practice in evaluating cross-domain classifiers, in general [Daxenberger et al., 2017].

12.1 Text Classification

For the text classifiers, we employ the most advanced form of supervised classifier, namely BERT. The configuration of that classifier mostly follows the configuration of our cross-domain classifier (§6.2). That is, we use the same model, i.e. BERT-Large, Uncased: 24-layer, 1024-hidden, 16-heads, 340M parameters. Again, we also just fine-tune that model by adding a further layer on top of the existing model. The fine-tuning is done on the respective training data, i.e. either the dataset from Davidson et al. [2017] or the dataset from the Kaggle-challenge.[16]

12.2 Word-Specific Classification

For our word-specific classifier, we employ a feature-based approach. Thus, we follow Holgate et al. [2018] who report worse performance with BiLSTMs than with a feature-based approach on their dataset, which (like ours) is a dataset for word-sense disambiguation in abusive language detection. The feature set we use for our present task is depicted in Table 2. This table also provides a motivation for each of the chosen features. We initially also considered the feature set by Holgate et al. [2018]. However, in our exploratory experiments we found that a lightweight feature set as we proposed is sufficient for this task.

For this type of features, we again chose SVM^{Light} [Joachims, 1999] which is widely used for NLP-related tasks and is particularly effective on the detection of abusive language [Schmidt and Wiegand, 2017, Wiegand et al., 2018a,b]. We trained the SVM on the respective training data, i.e. either the dataset of Holgate et al. [2018], the training data produced via our simple heuristic or the training data generated with the help of the middle-finger emoji.

Since in virtually all gold standards for abusive language detection, the class of abusive language is always a minority class, the SVM needs to be adjusted to the given class distribution. Otherwise we are very likely to obtain a majority-class classifier, that is, a classifier that will always predict the non-abusive category. For that purpose, SVM^{Light} offers a j-parameter that represents a cost-factor by which training errors on positive examples outweigh errors on negative examples. Unfortunately, we had no development set for this type of classification and in order to offer a fair comparison against our baselines, we refrained from tuning this parameter on our test set. Instead, we simply took the setting used in cross-domain classification from Wiegand et al. [2018a], i.e. $j = 2$ (*the supplementary notes of Wiegand et al. [2018a] document this parameter value*).

[16] `www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge`

feature	motivation
words immediately preceding and following target word	may be helpful in order to learn phrases such as *fuck off*; larger context is avoided since we are likely to overfit to particular domains
presence of abusive words (according to the lexicon from Wiegand et al. [2018a]) in context?	target word is likely to be abusive if it co-occurs with other (unambiguously) abusive words
presence of positive/negative polar expressions (according to the *Subjectivity Lexicon* [Wilson et al., 2005]) in context?	positive polar expressions rarely co-occur with abusive language, negative polar expressions, however, do
which pronouns are in context?	2nd person pronouns are typical of abusive usage: *you are a bitch*; 1st person pronouns are likely to indicate non-abusive usage: *I am a bitch*
quotation signs in tweet?	quotation signs indicate reported speech; a tweet may report an abusive remark, however, this reported tweet may not be abusive, for example: *just played the album for my sister and the first thing she says is "what have you done **bitch**"*
presence of exclamation sign?	a typical means of expressing high emotional intensity (similar to interjections)

Table 2: Features for disambiguating a potentially abusive word (referred to as *target word*); *context* is defined as a window of 4 words neighbouring the target word.

It may not be optimal but it is fair to other baselines and, at the same time, it prevented the learned models to become majority-class classifiers.

All parameters of the SVM other than the j-parameter were left in their default configuration.

References

Eneko Agirre and Aitor Soroa. 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, 2009.

Miguel A. Álvarez-Carmona, Estefanía Guzmán-Falcón, Manuel Montes y Gómez, Hugo Jair Escalante, Luis Villaseñor-Pineda, Verónica Reyes-Meza, and Antonio Rico-Sulayes. Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets. In *Proceedings of the Evaluation of Human Language Technologies for Iberian Languages Workshop (IberEval)*, Sevilla, Spain, 2018.

Ramón F. Astudillo, Silivio Amir, Wang Lin, Mario Silva, and Isabel Trancoso. Learning Word Representations from Scarce and Noisy Data with Embedding Subspaces. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL/IJCNLP)*, pages 1074–1084, Beijing, China, 2015.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet Project. In *Proceedings of the International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*, pages 86–90, Montréal, Quebec, Canada, 1998.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetti. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226, 2009.

Cristina Bosco, Felice Dell'Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. Overview of the EVALITA 2018 Hate Speech Detection Task. In *Proceedings of the Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA)*, Turin, Italy, 2018.

Aljoscha Burchardt, Kathrin Erk Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. The SALSA Corpus: a German Corpus Resource for Lexical Semantics. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, pages 969–974, Genoa, Italy, 2006.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word Translation Without Parallel Data. *arXiv preprint arXiv:1710.04087*, 2017.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, Montréal, Canada, 2017.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. What is the Essence of a Claim? Cross-Domain Claim Identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1055–1066, Copenhagen, Denmark, 2017.

Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. OpenWordNet-PT: An Open Brazilian Wordnet for Reasoning. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 353–360, Mumbai, India, 2012.

Jacob Devlin, Ming-Wei Chang, KentonLee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 4171—4186, Minneapolis, MN, USA, 2019.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). In *Proceedings of the Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA)*, Turin, Italy, 2018a.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *Proceedings of the Evaluation of Human Language Technologies for Iberian Languages Workshop (IberEval)*, Sevilla, Spain, 2018b.

Jorge A. Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. The brWaC corpus: A new open resource for Brazilian Portuguese. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, pages 4339–4344, Miyazaki, Japan, 2018.

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. A Hierarchically-Labeled Portuguese Hate Speech Dataset. In *Proceedings of the Workshop on Abusive Language Online (ALW)*, pages 94–104, Florence, Italy, 2019.

Birgit Hamp and Helmut Feldweg. GermaNet - a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15, Madrid, Spain, 1997.

Eric Holgate, Isabel Cachola, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. Why Swear? Analyzing and Inferring the Intentions of Vulgar Expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4405–4414, Brussels, Belgium, 2018.

Thorsten Joachims. Making Large-Scale SVM Learning Practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 169–184. MIT Press, 1999.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of Tricks for Efficient Text Classification. In *Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL)*, pages 427–431, Valencia, Spain, 2017.

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

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. Benchmarking Aggression Identification in Social Media. In *Proceedings of the Workshop on Trolling, Aggression and Cyberbullying (TRAC)*, pages 1–11, Santa Fe, NM, USA, 2018.

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

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

George Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3:235–244, 1990.

Saif Mohammad and Peter Turney. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 39(3):555–590, 2013.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive Language Detection in Online User Content. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 145–153, Republic and Canton of Geneva, Switzerland, 2016.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Dohar, Qatar, 2014.

Telmo Pires, Eva Schlinger, and Dan Garrette. How Multilingual is Multilingual BERT? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4996—-5001, Florence, Italy, 2019.

Peter Prettenhofer and Benno Stein. Cross-Language Text Classification using Structural Correspondence Learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1118–1127, Uppsala, Sweden, 2010.

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

Anna Schmidt and Michael Wiegand. 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, 2017.

Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. Overview of GermEval Task 2, 2019 Shared Task on the Identification of Offensive Language. In *Proceedings of the GermEval Workshop*, pages 352–363, Erlangen, Germany, 2019.

Partha Pratim Talukdar, Joseph Reisinger, Marius Pasca, Deepak Ravichandran, Rahul Bhagat, and Fernando Pereira. 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, 2008.

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

Michael Wiegand and Josef Ruppenhofer. Opinion Holder and Target Extraction based on the Induction of Verbal Categories. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, pages 215–225, Beijing, China, 2015.

Michael Wiegand, Christine Bocionek, and Josef Ruppenhofer. 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)*, pages 800–810, San Diego, CA, USA, 2016.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 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, 2018a.

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

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 602–608, Minneapolis, MN, USA, 2019a.

Michael Wiegand, Maximilian Wolf, and Josef Ruppenhofer. Detecting Derogatory Compounds–An Unsupervised Approach. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 2076–2081, Minneapolis, MN, USA, 2019b.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 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)*, pages 347–354, Vancouver, BC, Canada, 2005.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Koumar. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pages 75–86, Minneapolis, MN, USA, 2019.

Torsten Zesch, Christof Müller, and Iryna Gurevych. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, pages 1646–1652, Marrakech, Morocco, 2008.