Modeling Profanity and Hate Speech in Social Media with Semantic Subspaces

Vanessa Hahn, Dana Ruiter, Thomas Kleinbauer, Dietrich Klakow
Spoken Language Systems Group
Saarland University
Saarbrücken, Germany
{vhahn|druiter|kleiba|dklakow}@lsv.uni-saarland.de

Abstract

Hate speech and profanity detection suffer from data sparsity, especially for languages other than English, due to the subjective nature of the tasks and the resulting annotation incompatibility of existing corpora. In this study, we identify profane subspaces in word and sentence representations and explore their generalization capability on a variety of similar and distant target tasks in a zero-shot setting. This is done monolingually (German) and cross-lingually to closely-related (English), distantly-related (French) and non-related (Arabic) tasks. We observe that, on both similar and distant target tasks and across all languages, the subspace-based representations transfer more effectively than standard BERT representations in the zero-shot setting, with improvements between F1 $+10.9$ and F1 $+42.9$ over the baselines across all tested monolingual and cross-lingual scenarios.

1 Introduction

Profanity and online hate speech have been recognized as crucial problems on social media platforms as they bear the potential to offend readers and disturb communities. The large volume of user-generated content makes manual moderation very difficult and has motivated a wide range of natural language processing (NLP) research in recent years. However, the issues are far from solved, and the automatic detection of profane and hateful contents in particular faces a number of severe challenges.

Pre-trained transformer-based (Vaswani et al., 2017) language models, e.g. BERT (Devlin et al., 2019), play a dominant role today in many NLP tasks. However, they work best when large amounts of training data are available. This is typically not the case for profanity and hate speech detection where few datasets are currently available (Waseem and Hovy, 2016; Basile et al., 2019; Struß et al., 2019) with moderate sizes at most. In addition, these tasks are known to be highly subjective (Waseem, 2016). Annotation protocols for hate speech and profanity often rely on different assumptions that make it non-trivial to combine multiple datasets. In addition, such datasets only exist for few languages besides English (Ousidhoum et al., 2019; Abu Farha and Magdy, 2020; Zampieri et al., 2020).

For such low-resource scenarios, few- and zero-shot transfer learning has seen an increased interest in the research community. One particular approach, using semantic subspaces to model specific linguistic aspects of interest (Rothe et al., 2016), has proven to be effective for representing contrasting semantic aspects of language such as e.g. positive and negative sentiment.

In this paper, we propose to learn **semantic subspaces to model profane language** on both the word and the sentence level. This approach is especially promising because of its ability to cope with sparse profanity-related datasets confined to very few languages. Profanity and hate speech often co-occur but are not equivalent, since not all hate speech is profane (e.g. *implicit* hate speech) and not all profanity is hateful (e.g. *colloquialisms*). Despite being *distantly related* tasks, we posit that modeling profane language via semantic subspaces may have a positive impact on downstream hate speech tasks.

We analyze the efficacy of the subspaces to encode the profanity (*neutral* vs. *profane* language) aspect and apply the resulting subspace-based representations to a **zero-shot transfer classification** scenario with both similar (*neutral/profane*) and distant (*neutral/hate*) target classification tasks. To study their ability to generalize across languages we evaluate the zero-shot transfer in both a **monolingual** (German) and a **cross-lingual** setting

with closely-related[1] (English), distantly-related (French) and non-related (Arabic) languages.

We find that subspace-based representations outperform popular alternatives, such as BERT or word embeddings, by a large margin across all tested transfer tasks, indicating their strong generalization capabilities not only monolingually but also cross-lingually. We further show that semantic subspaces can be used for **word-substitution** tasks with the goal of generating automatic suggestions of neutral counterparts for the civil rephrasing of profane contents.

2 Related Work

Semantic subspaces have been used to identify gender (Bolukbasi et al., 2016) or multiclass ethnic and religious (Manzini et al., 2019) bias in word representations. Liang et al. (2020) identify multiclass (gender, religious) bias in sentence representations. Similarly, Niu and Carpuat (2017) identify a stylistic subspace that captures the degree of formality in a word representation. This is done using a list of minimal-pairs, i.e. pairs of words or sentences that only differ in the semantic feature of interest over which they perform principal component analysis (PCA). We take the same general approach in this paper (see Section 3).

Conversely, Gonen and Goldberg (2019) show that the methods in Bolukbasi et al. (2016) are not able to identify and remove the gender bias entirely. Following this, Ravfogel et al. (2020) argue that semantic features such as gender are encoded non-linearly, and suggest an iterative approach to identifying and removing gender features from semantic representations entirely.

Addressing the issue of data sparseness, Rothe et al. (2016) use ultradense subspaces to generate task-specific representations that capture semantic features such as abstractness and sentiment and show that these are especially useful for low-resourced downstream tasks. While they focus on using small amounts of labeled data of a specific target task to learn the subspaces, we focus our study on learning a generic profane subspace and test its generalization capacity on similar and distant target tasks in a zero-shot setting.

Zero-shot transfer, where a model trained on a set of tasks is evaluated on a previously unseen task, has recently gained a lot of traction in NLP. Nowadays, this is done using large-scale transformer-based language models such as BERT, that share parameters between tasks. Multilingual varieties such as XLM-R (Conneau et al., 2020) enable the zero-shot cross-lingual transfer of a task. One example is sentence classification trained on a (high-resource) language being transferred into another (low-resource) language (Hu et al., 2020).

[1]Both English and German belong to the West-Germanic language branch, and are thus closely-related. French, on the other hand, is only distantly related to German via the Indo-European language family, while Arabic (Semitic language family) and German are not related.

w (profane)	**ŵ (neutral)**
Arschloch [asshole]	Mann [man]
Fotze [cunt]	Frau [woman]
Hackfresse [shitface]	Mensch [human]

Table 1: Examples of word-level minimal pairs.

3 Method: Semantic Subspaces

A common way to represent word-level semantic subspaces is based on a set P of so-called *minimal pairs*, i.e. N pairs of words $(w, \hat{w})$ that differ only in the semantic dimension of interest (Bolukbasi et al., 2016; Niu and Carpuat, 2017). Table 1 displays some examples of such word pairs for the profanity domain. Each word w is encoded as a word embedding $e(w)$:

$$P = \{(e(w_1), e(\hat{w}_1)), \ldots, (e(w_N), e(\hat{w}_N))\}$$

Then, each pair is normalized by a mean-shift:

$$\bar{P} = \{(e(w_i) - \mu_i, e(\hat{w}_i) - \mu_i) | 1 \leq i \leq N\}$$

where each $\mu_i = \frac{1}{2}(e(w_i) + e(\hat{w}_i))$.

Finally, PCA is performed on the set $\bar{P}$ and the most significant principal component (PC) is used as a representation of the semantic subspace.

We diverge from this approach in four ways:

Normalization We note that there is no convincing justification for the normalization step. As our experiments in the following sections show, we find that the profanity subspace is better represented by P than by $\bar{P}$. For our experiments, we thus distinguish three different types of representations:

- **BASE**: The raw featurized representation r.
- **PCA-RAW**: Featurized representation r projected onto the non-normalized subspace $S(P)$.

- **PCA-NORM**: Featurized representation r projected onto the normalized subspace $S(\bar{P})$.

Here, projecting a vector representation r onto a subspace is defined as the dot product $r \cdot S(P)$.

Number of Principal Components c The use of just a single PC as the best representation of the semantic subspace is not well motivated. This is recognized by Niu and Carpuat (2017) who experiment on the first $c = 1, 2, 4, \ldots, 512$ PC and report results on their downstream-task directly. However, a downside of their method for determining a good value for c is the requirement of a task-specific validation set which runs orthogonal to the assumption that a good semantic subspace should generalize well to many related tasks.

Instead, we propose the use of an *intrinsic evaluation* that requires no additional data to estimate a good value for c. Rothe et al. (2016) have shown that semantic subspaces are especially useful for classification tasks related to the semantic feature encoded in the subspace. Here, we argue the inverse: if a semantic subspace with c components yields the best performance on a related classification task, c should be an appropriate number of components to encode the semantic feature.

More specifically, we apply a classifier function $f(x) = y$, which learns to map a subspace-based representation $x = e \cdot S(P)$ to a label $y \in \{\text{profane}, \text{neutral}\}$. We learn $f(x)$ on the same set P used to learn the subspace. In order to evaluate on previously unseen entities, we employ 5-fold cross validation over the available list of minimal pairs P and evaluate Macro F1 on the held-out fold. Due to the simplicity of this intrinsic evaluation, the experiment can be performed for all values of c and the c yielding the highest average Macro F1 is selected as the final value. The above holds for P and $\bar{P}$ equally.

Sentence-Level Minimal Pairs We move the word-level approach to the sentence level. In this case, minimal pairs are made up of vector representations of sentences $(e(s), e(\hat{s}))$.

In order to standardize the approach and to focus the variation in the sentence representations on the profanity feature, sentence-level minimal pairs are constructed by keeping all words contained equivalent except for *significant words* that in themselves are minimal pairs for the semantic feature of interest. For instance, a sentence-level minimal pair for the *profanity* feature with significant words:

The food here is shitty.
The food here is disgusting.

Zero-Shot Transfer In order to evaluate how well profanity is encoded in the resulting word- and sentence-level subspaces, we test their generalization capabilities in a zero-shot classification setup. Given a subspace $S(P)$ (or $S(\bar{P})$), we train a classifier $f(x) = y$ to classify subspace-based representations $x = e \cdot S(P)$ as belonging to class $y \in \{\text{profane}|\text{neutral}\}$. The x used to train the classifier are the same entities in the minimal pairs used to learn $S(P)$. This classification task is the *source task* $\mathcal{T} = \{x, y\}$. As the classifier is learned on subspace-based representations, it should be able to generalize significantly better to previously unseen profanity-related tasks than a classifier learned on generic representations $x = e$ (Rothe et al., 2016). Given a previously unseen task $\bar{\mathcal{T}} = \{\bar{x}, \bar{y}\}$, we follow a **zero-shot transfer** approach and let classifier f, learned on source task $\mathcal{T}$ only, predict the new labels $\bar{y}$ given instances $\bar{x}$ without training it on data from $\bar{\mathcal{T}}$. The zero-shot generalization can be quantified by calculating the accuracy of the predicted labels $\hat{\bar{y}}$ given the gold labels $\bar{y}$. The extend of this zero-shot generalization capability can be tested by performing zero-shot classification on a variety of unseen tasks $\bar{\mathcal{T}}$ with variable task distances $\bar{\mathcal{T}} \Leftrightarrow \mathcal{T}$.

4 Experimental Setup

4.1 Data

Word Lists The minimal-pairs used in our experiments are derived from a German slur collection[2].

Fine-Tuning We use the German, English, French and Arabic portions of a large collection of tweets[3] collected between 2013–2018 to fine-tune BERT. For the German BERT model, all available German tweets are used, while the multilingual BERT is fine-tuned on a balanced corpus of 5M tweets per language. For validation during fine-tuning, we set aside $1k$ tweets per language.

Target Tasks We test our sentence-level representations, which are used to train a *neutral/profane* classifier on a subset of minimal pairs, on several hate speech benchmarks. For all four languages, we focus on a distant task DT (*neutral/hate*). For

[2] www.hyperhero.com/de/insults.htm
[3] www.archive.org/details/twitterstream

Corpus	# Sentences	# Tokens
Fine-Tuning		
Twitter-DE	5(9)M	45(85)M
Twitter-EN	5M	44M
Twitter-FR	5M	58M
Twitter-AR	5M	75M
Target Tasks		
DE-ST	111/111	1509/1404
DE-DT	2061/970	14187/9333
EN-ST	93/93	1409/1313
EN-DT	288/865	8032/3647
AR-ST	12/12	164/84
AR-DT	46/54	592/506
FR-DT	5822/302	49654/2660

Table 2: Number of sentences and tokens of the data used for fine-tuning BERT for the sentence-level experiments. Target task test sets are reported with their respective *neutral*/*hate* (DT) and *neutral*/*profane* (ST) distributions.

German, English and Arabic we additionally evaluate on a similar task ST (*neutral*/*profane*), for which we removed additional classes (*insult*, *abuse* etc.) from the original finer-grained data labels and downsampled to the minority class (*profane*).

For German (DE), we use the test sets of GermEval-2019 (Struß et al., 2019) Subtask 1 (*Other*/*Offense*) and Subtask 2 (*Other*/*Profanity*) for DT and ST respectively. For English (EN), we use the HASOC (Mandl et al., 2019) Subtask A (*NOT*/*HOF*) and Subtask B (*NOT*/*PRFN*) for DT and ST respectively. French (FR) is tested on the hate speech portion (*None*/*Hate*) of the corpus created by Charitidis et al. (2020) for DT only, while Arabic (AR) is tested on Mubarak et al. (2017) for DT (*Clean*/*Obscene+Offense*) and ST (*Clean*/*Obscene*). As AR has no official train/test splits, we use the last 100 samples for testing. The training data of these corpora is not used.

Table 2 summarizes the data used for fine-tuning as well as testing.

Pre-processing The Twitter corpora for fine-tuning were pre-processed by filtering out incompletely loaded tweets and duplicates. We also applied language detection using `spacy` to further remove tweets that consisted of mainly emojis or tweets that were written in other languages.

4.2 Model Specifications

To achieve good coverage of profane language, we use 300-dimensional German FastText embeddings (Deriu et al., 2017) trained on 50M German tweets for the word-level experiments in Section 5.

The BERT models (Devlin et al., 2019) used in Section 6 are `Bert-Base-German-Cased`[4] and `Bert-Base-Multilingual-Cased` for the monolingual and multilingual experiments respectively, since they pose strong baselines. We fine-tune on the Twitter data (Section 4.1) using the masked language modeling objective and early stopping over the evaluation loss ($\delta = 0$, $\text{patience} = 3$). All classification experiments use Linear Discriminant Analysis (LDA) as the classifier.

5 Word-Level Subspaces

Before moving to the lesser explored sentence-level subspaces, we first verify whether word-level semantic subspaces can also capture complex semantic features such as profanity.

5.1 Minimal Pairs

Staying within the general low-resource setting prevalent in hate speech and profanity domains, and to keep manual annotation effort low, we randomly sample a small amount of words from the German slur lists, namely 100, and manually map these to their neutral counterparts (Table 1). We focus this list on nouns describing humans.

Each word in our minimal pairs is featurized using its word embedding, this is our BASE representation. We learn PCA-RAW and PCA-NORM representations on the embedded minimal pairs.

5.2 Classification

We evaluate how well the resulting representations BASE, PCA-RAW and PCA-NORM encode information about the profanity of a word by focusing on a related word classification task where unseen words are classified as *neutral* or *profane*. To evaluate how efficient the subspaces can be learned in a low-resource setting, we downsample the list of minimal pairs to learn the subspace-based representations and the classification task to 10–100 word pairs. After the preliminary exploration of the number of principal components (PC) required to represent profanity, the number of PC for the final representations lie within a range of 15–111. Each experiment is run over 5 seeded runs and we report the average F1 Macro with standard error. As each seeded run resamples the training and test data, the standard error is also a good indicator

[4] www.deepset.ai/german-bert

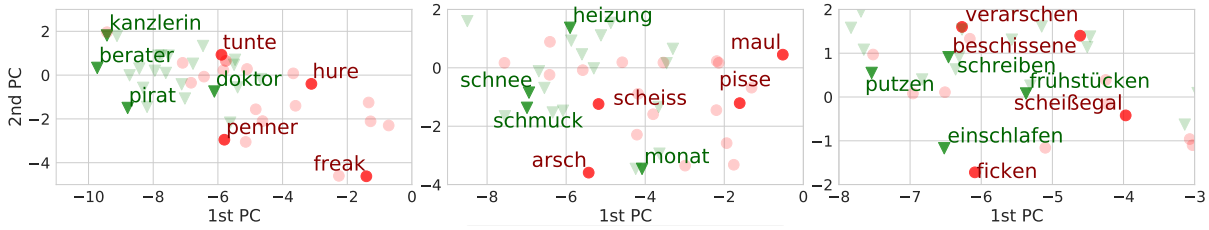

Figure 1: Projections of profane and neutral words from TL-1 (left), TL-2 (middle) and TL-3 (right) onto a word-level profane subspace learned by PCA-NORM on 10 minimal pairs (● Profane, ▼ Neutral).

of the variability of the method when trained on different subsets of minimal pairs.

Test Lists For this evaluation, we create three test lists (TL-{1,2,3}) of profane and neutral words. The contents of the three TLs are defined by their decreasing relatedness to the list of minimal pairs used for learning the subspace, which are nouns describing humans. TL-1 is thus also a list of nouns describing humans, TL-2 contains random nouns not describing humans, and TL-3 contains verbs and adjectives. The three TLs are created by randomly sampling from the word embeddings that underlie the subspace representations and adding matching words to TL-{1,2,3} until they each contain 25 profane and 25 neutral words, i.e. 150 in total.

Projecting the TLs onto the first and second PC of the PCA-NORM subspace learned on 10 minimal pairs suggests that a separation of profane and neutral words can be achieved for nouns describing humans (TL-1), while it is more difficult for less related words (TL-{2,3}) (Figure 1).

Results Across all TLs, the subspace-based representations outperform the generalist BASE representations (Figure 2), with PCA-NORM reaching F1-Macro scores of up to 96.0 (TL-1), 89.9 (TL-2) and 100 (TL-3) when trained on 90 word pairs. This suggests that they generalize well to unseen nouns describing humans as well as verbs and adjectives, while generalizing less to nouns not describing humans (TL-2). This may be due to TL-2 consisting of some less frequent compounds (e.g. Großmaul [big mouth]). PCA-NORM and PCA-RAW perform equally on TL-1 and TL-3, while PCA-NORM is slightly stronger on the mid-resource (50-90 pairs) range on TL-2. This suggests that the normalization step when constructing the profane subspace is only marginally beneficial. Even when the training data is very limited (10–40 pairs), the standard errors are decently small (F1 ±1–5), indicating that the choice of minimal pairs has only a small impact on the downstream model performance. When more training data is available (80–100 pairs), the influence of a single minimal pair becomes less pronounced and thus the standard error decreases significantly.

5.3 Substitution

We use the profane subspace S_{prf} to substitute a profane word w with a neutral counterpart $\hat{w}$. We do this by removing S_{prf} from w,

$$\hat{w} = \frac{w - S_{\text{prf}}}{||w - S_{\text{prf}}||} \quad (1)$$

and replacing it by its new nearest neighbor $\text{NN}(\hat{w})$ in the word embeddings. Here, we focus on the PCA-NORM subspace learned on 10 minimal pairs only. We use this subspace to substitute all profane words in TL-{1,2,3}.

Human Evaluation To analyze the similarity and profanity of the substitutions, we perform a small human evaluation. Four annotators were asked to rate the similarity of profane words and their substitutions, and also to give a profanity score between 1 (not similar/profane) and 10 (very similar/profane) to words from a mixed list of slurs and substitutions.

Original profane words were rated with an average of 6.1 on the **profanity** scale, while substitutions were rated significantly lower, with an average rating of 1.9. Minor differences exist across TL splits, with TL-1 dropping from 6.8 to 1.3, TL-2 from 6.1 to 3.1 and TL-3 from 5.4 to 2.1.

The average **similarity** rating between profane words and their substitution differs strongly across different TLs. TL-1 has the lowest average rating of 2.8, while TL-2 has a rating of 3.3 and TL-3 a rating of 5.1. This is surprising, since the subspaces generalized well to TL-1 on the classification task.

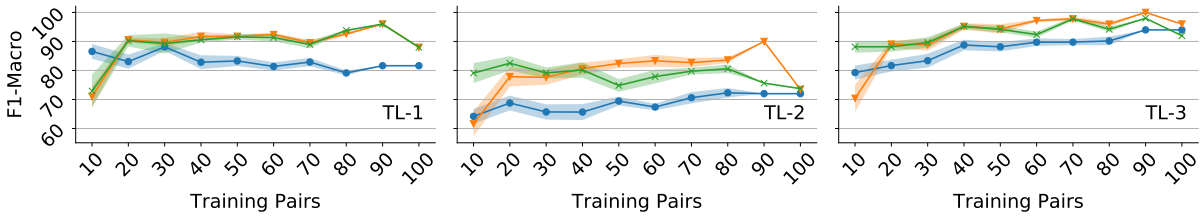

Figure 2: F1-Macro of the LDA models, using BASE or PCA-{RAW,NORM} representations on the word classification task based on 10 to 100 training word pairs (—● BASE, —▼ PCA-NORM, —× PCA-RAW).

Word w	**NN**(w)	**NN**($\hat{w}$)
Scheisse [shit]	Scheiße, Scheissse, Scheissse04, Scheißee	schrecklich, augenscheinlich, schwerlich, schwesterlich [horrible, evidently, hardly, sisterly]
Spast [dumbass]	Kackspsst, Spasti, Vollspast, Dummerspast	Mann, Mensch, Familienmensch, Menschn [man, person, family person, people]
Bitch	x6bitch, bitchs, bitchin, bitchhh	Frau, Afrikanerin, Mann, Amerikanerin [woman, african, man, american]
Arschloch [asshole]	Narschloch, Arschlochs, Arschloc, learschloch	Mann, Frau, Lebenspartnerin, Menschwesen [man, woman, significant other, human creature]
Fresse [cakehole]	Fresser, Schnauze, Kackfufresse, Schnauzefresse	Frau, Mann, Lebensgefährtin, Rentnerin [woman, man, significant other, retiree]

Table 3: Profane words w with top 4 NNs before (NN(w)) and after (NN($\hat{w}$)) removal of the profane subspace.

Qualitative Analysis To understand the quality of the substitutions, especially on TL-1, which has obtained the lowest similarity score in the human evaluation, we perform a small qualitative analysis on 3 words sampled from TL-1 (*Spast, Bitch, Arschloch*) and 1 word sampled from TL-2 (*Fresse*) and TL-3 (*Scheiss*) each. Before removal, the nearest neighbors (NNs, Table 3) of the sampled offensive words were mostly orthographic variations (e.g. *Scheisse [shit]* vs. *Scheiße*) or compounds of the same word (e.g. *Spast [dumbass]* vs. *Vollspast [complete dumbass]*). After removal, the NNs are still negative but not profane (e.g. *Scheisse* →*schrecklich [horrible]*). While the first NNs are decent counterparts, later NNs introduce other (gender, ethnic, etc.) biases, possibly stemming from the word embeddings or from the minimal pairs used to learn the subspace. The counterparts to *Scheisse [shit]* seem to focus around the phonetics of the word (all words contain *sch*), which may also be due to the poor representation of adjectives in embedding spaces. *Fresse [cakehole]* is ambiguous[5], thus the subspace does not entirely capture it and the new NNs are neutral, but unrelated words.

While human similarity ratings on TL-1 were low, qualitative analysis shows that these can still be reasonable. The low rating on TL-1 may be due to annotators' reluctance to equate human-referencing slurs to neutral counterparts.

The ability to automatically find neutral alternatives to slurs may lead to practical applications such as the suggestion of alternative wordings.

6 Sentence-Level Subspaces

In Section 5, we identified profane subspaces on the word-level. However, abuse mostly happens on the sentence and discourse-level and is not limited to the use of isolated profane words. Therefore, we move this method to the sentence-level, exploring the two subspace-based representation types PCA-RAW and PCA-NORM. Concretely, we learn sentence-level profane subspaces that allow a context-sensitive representation and thus go beyond isolated profane words, and verify their efficacy to represent *profanity*. Similarly to the word-level experiments, we focus our analysis on the ability of the subspaces to generalize to similar (*neutral/profane*) and distant (*neutral/hate*) tasks. We compare their performance with a BERT-encoded BASE representation, which does not use a semantic subspace.

6.1 Minimal Pairs

Using the German slur collection, we identify tweets in Twitter-DE containing swearwords, from which we then take 100 random samples. We create a neutral counterpart by manually replacing significant words, i.e. swearwords, with a neutral

[5] *Fresse* can mean *shut up*, as well as being a pejorative for *face* and *eating*.

variation while keeping the rest of the tweet as is:

a) *ich darf das nicht verkacken!!!*
[I must not fuck this up!!!]]
b) *ich darf das nicht vermasseln!!!*
[I must not mess this up!!!]

6.2 Monolingual Zero-Shot Transfer

We validate the generalization of the German sentence-level subspaces to a similar (*profane*) and distant (*hate*) domain by zero-shot transferring them to unseen German target tasks and analyzing their performance.

6.2.1 Representation Types

We fine-tune `Bert-Base-German-Cased` on Twitter-DE (9M Tweets). Each sentence in our list of minimal pairs is then encoded using the fine-tuned German BERT and its sentence representation $s = \text{mean}(\{h_1, ..., h_T\})$ is the mean over the T encoder hidden states h. This is our BASE representation. We further train PCA-RAW and PCA-NORM on a subset of our minimal pairs. We chose 14–96 PCs for PCA-RAW and 9–94 PCs for PCA-NORM depending on the size of the subset of minimal pairs used to generate the subspace.

6.2.2 Results

We train the PCA-RAW and PCA-NORM representations on subsets of increasing size $(10, 20, \ldots, 100$ minimal pairs). For each subset and representation type (BASE, PCA-RAW, PCA-NORM), we train an LDA model to identify whether a sentence in the subset of minimal pairs is neutral or profane. These models are zero-shot transferred to the German similar task ST (*neutral/profane*) and distant task DT (*neutral/hate*). We report the average F1-Macro and standard error over 5 seeded runs, where each run resamples its train and test data.

ST: Similar Task Despite the fact that the LDA models were never trained on the target task data, the PCA-RAW and PCA-NORM representations yield high peaks in F1 when trained on 50 (F1 68.9, PCA-RAW) minimal pairs and tested on DE-ST (Figure 3). PCA-RAW outperforms PCA-NORM for almost all data sizes. PCA-RAW outperforms the BERT (BASE) representations especially on the very low-resource setting (10–60 pairs), with an increase of F1 +14.2 at 40 pairs. Once the training size reaches 70 pairs, the differences in F1 become smaller. The subspace-based representations are especially useful for the low-resource scenario.

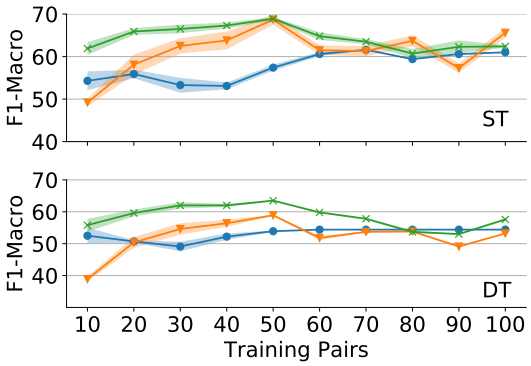

Figure 3: F1-Macro of the LDA models, zero-shot transferred to the similar (top) and distant (bottom) German tasks (BASE, PCA-NORM, PCA-RAW).

DT: Distant Task For the distant task DT, the general F1 scores are lower than for the similar task ST. However, PCA-RAW still reaches a Macro-F1 of 63.5 at 50 pairs for DE-DT. This indicates that the profane subspace found by PCA-RAW partially generalizes to a broader, offensive subspace. Similar to ST, the projected PCA-RAW representations are especially useful in the low-resource case up to 50 sentences. The F1 of the BERT baseline is well below the PCA-RAW representations when data is sparse, with a major gap of F1 +10.9 at 30 pairs for DE-DT. The classifier using BASE representations stays around F1 53.0 (DE-DT) and does not benefit from more data, indicating that these representations do not generalize to the target tasks. However, once normalization (PCA-NORM) is added, the generalization is also lost and we see a drop in performance around or below the baseline. As for ST, all three representation types level out once higher amounts of data (70–80 pairs) are reached.

The standard errors show a similar trend to those in the word-level experiments: we observe a small standard error when training data is sparse (10–40 pairs), indicating that the choice of minimal pairs has a small impact on the subspace quality, which decreases further when more minimal pairs are available for training (50–100 pairs).

6.3 Zero-Shot Cross-Lingual Transfer

To verify whether the subspaces also generalize to other languages, we zero-shot transfer and test the German BASE, PCA-RAW and PCA-NORM representations on the similar and distant tasks of closely-related (English), distantly-related (French) and non-related (Arabic) languages. For French, we only test on DT due to a lack of data for ST.

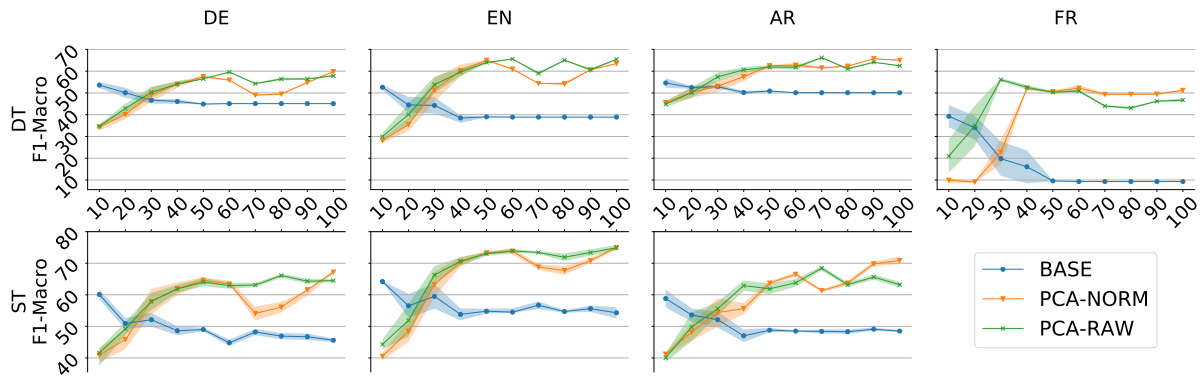

Figure 4: F1-Macro of the LDA model, using BASE or PCA-{RAW,NORM} representations, zero-shot transferred to the similar (bottom) and distant (top) German, English, Arabic and French tasks.

6.3.1 Representation Types

The setup is the same as in Section 6.2.1, except for using `Bert-Base-Multilingual-Cased` and fine-tuning it on a corpus consisting of the 5M {DE,EN,FR,AR} tweets. The resulting model is used to generate the hidden-representations needed to construct the BASE, PCA-RAW and PCA-NORM representations. After performing 5-fold cross validation, the optimal number of PC is determined. Depending on the number of minimal pairs, the resulting subspace sizes lie between 8–67 (PCA-RAW) and 10–44 (PCA-NORM).

6.3.2 Results

As in Section 6.2.2, we train on increasingly large subsets of the German minimal pairs.

ST: Similar Task We test the generalization of the German representations on the similar (*neutral/profane*) task on EN-ST and AR-ST as well as DE-ST for reference. Note that the LDA classifiers were trained on the German minimal pairs only, without access to target task data.

The trends on the three test sets are very similar to each other (Figure 4, bottom), indicating that the German profane subspaces transfer not only to the closely-related English, but also to the unrelated Arabic data. For all three languages, the PCA-{RAW,NORM} methods tend to grow in performance with increasing data until around 40 sentence pairs when the method seems to converge. This yields a performance of F1 66.1 on DE-ST at 80 pairs, F1 74.9 on EN-ST at 100 pairs and F1 68.4 on AR-ST at 70 pairs for PCA-RAW.

Overall, larger amounts of pairs are needed to reach top-performance in comparison to the monolingual case. This trend is also present when testing on DE-ST, leading us to posit that it is caused not by the cross-lingual transfer itself, but by the different underlying BERT models used to generate the initial representations. The differences in F1 between PCA-RAW and PCA-NORM are mere fluctuations between the two methods. The BASE representations are favorable only at 10 training pairs, with more data they overfit on the source task and are outperformed by the subspace representations, with differences of F1 +20.6 at 100 sentence pairs (PCA-RAW) on EN-ST, and F1 +22.4 at 100 sentence pairs (PCA-NORM) on AR-ST.

DT: Distant Task Similar trends to ST are observed on the distant (*neutral/hate*) tasks (Figure 4, top). While the BASE representations are strongest at 10 sentence pairs, they are outperformed by the subspace-based representations at around 30 pairs. PCA-RAW outperforms PCA-NORM and peaks at F1 59.6 (60 pairs), F1 65.6 (60 pairs), F1 66.2 (70 pairs) and F1 56.1 (30 pairs) for the German, English, Arabic and French test sets respectively.

We conclude that the German profane subspaces are transferable not only monolingually or to closely-related languages (English) but also to distantly-related (French) and non-related languages (Arabic), making a zero-shot transfer possible on both similar (neutral/profane) and distant tasks (neutral/hate). The BERT embeddings, on the other hand, were not able to perform the initial transfer, i.e. from minimal-pair training to similar and distant target tasks, thus making the transfer to other languages futile. Subspace-based representations are a powerful tool to fill this gap, especially for classifiers trained on small amounts of source target data and zero-shot transfer to related tasks.

External Comparison The transfer capabilities of our subspace-based models can be set into perspective by comparing them to state-of-the-art classification models that were trained directly on our target tasks. For **DT**, the top scoring team on EN-DT reaches higher levels of F1 (75.6) (Mandl et al., 2019) than our best PCA-RAW representations (F1 65.6). Similarly, the top scoring model on CHS-FR (Charitidis et al., 2020) lies at F1 82.0 and thus F1 +25.9 over PCA-RAW. However, PCA-RAW outperforms the best-performing model reported in Mubarak et al. (2017) (F1 60.0) by F1 +6.2. Note, however, that this comparison is vague, as there is no standard train-test split for AR. For ST, no direct comparison to SOTA models can be made, since the profane-neutral classification task is usually part of a larger multi-class classification task. Nevertheless, the success of simple subspace-based LDA models, trained on very small amounts of task-distant German data, at cross-lingually zero-shot transferring to various tasks underlines the generalization capability of our approach.

6.4 Qualitative Analysis

A qualitative per-task analysis of the errors of the best performing models (PCA-RAW) reveals that some of the gold labels are debatable. The subjectivity of hate speech is a well-known issue for automatic detection tasks. Here, it is especially observable for EN, AR and FR, where arguably offensive comments were annotated as neutral but classified as offensive by our model:

> *C'est toi la pute. Va voir ta mère*
> *[You are the whore. Go see your mom]*

We find that the models tend to over-blacklist tweets across languages as most errors stem from classifying neutrally-labeled tweets as offensive. This is triggered by negative words, e.g. *crime*, as well as words related to religion, race and politics, e.g.:

> *No Good Friday agreement, no deals with Trump.*

7 Conclusion and Future Work

In this work, we have shown that a complex feature such as *profanity* can be encoded using semantic subspaces on the word and sentence-level.

On the **word-level**, we found that the subspace-based representations are able to generalize to previously unseen words. Using the profane subspace, we were able to substitute previously unseen profane words with neutral counterparts.

On the **sentence-level**, we have tested the generalization of our subspace-based representations (PCA-RAW, PCA-NORM) against raw BERT representations (BASE) in a zero-shot transfer setting on both similar (*neutral/profane*) and distant (*neutral/hate*) tasks. While the BASE representations failed to zero-shot transfer to the target tasks, the subspace-based representations were able to perform the transfer to both similar and distant tasks, not only monolingually, but also to the closely-related (English), distantly-related (French) and non-related (Arabic) language tasks. We observe major improvements between F1 +10.9 (PCA-RAW on DE-DT) and F1 +42.9 (PCA-NORM on FR-DT) over the BASE representations in all scenarios. As our experiments have shown that the commonly used mean-shift normalization is not required, we plan to conduct further experiments using unaligned significant words/sentences.

The code, the fine-tuned models, and the list of minimal-pairs are made publicly available[6].

Acknowledgements

We want to thank the annotators Susann Boy, Dominik Godt and Fabian Gössl. We also thank Badr Abdullah, Michael Hedderich, Jyotsna Singh and the anonymous reviewers for their valuable feedback. The project on which this paper is based was funded by the DFG under the funding code WI 4204/3-1. Responsibility for the content of this publication is with the authors.

References

Ibrahim Abu Farha and Walid Magdy. 2020. Multi-task learning for Arabic offensive language and hate-speech detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 86–90, Marseille, France. European Language Resource Association.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

[6] www.github.com/uds-lsv/profane_subspaces

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Polychronis Charitidis, Stavros Doropoulos, Stavros Vologiannidis, Ioannis Papastergiou, and Sophia Karakeva. 2020. Towards countering hate speech and personal attack in social media. *Online Social Networks and Media*, 17:100071.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jan Deriu, Aurelien Lucchi, Valeria De Luca, Aliaksei Severyn, Simon Müller, Mark Cieliebak, Thomas Hofmann, and Martin Jaggi. 2017. Leveraging Large Amounts of Weakly Supervised Data for Multi-Language Sentiment Classification. In *WWW 2017 - International World Wide Web Conference*, page 1045–1052, Perth, Australia.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Paul Pu Liang, Irene Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 5502–5515, Online.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA. Association for Computing Machinery.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on Arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56, Vancouver, BC, Canada. Association for Computational Linguistics.

Xing Niu and Marine Carpuat. 2017. Discovering stylistic variations in distributional vector space models via lexical paraphrases. In *Proceedings of the Workshop on Stylistic Variation*, pages 20–27, Copenhagen, Denmark. Association for Computational Linguistics.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777, San Diego, California. Association for Computational Linguistics.

Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of germeval task 2, 2019 shared task on the identification of offensive language.

In *Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), October 9 – 11, 2019 at Friedrich-Alexander-Universität Erlangen-Nürnberg*, pages 352 – 363, München [u.a.]. German Society for Computational Linguistics & Language Technology und Friedrich-Alexander-Universität Erlangen-Nürnberg.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

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

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.