Supplementary Notes to "Identifying Implicitly Abusive Remarks about Identity Groups using a Linguistically Informed Approach"

April 10, 2022

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 on our crowdsourcing experiments for building classifiers (§2)
- details on the configuration of the supervised classifiers we used in our experiments (§3)
- details on the linguistically informed classifier as implemented for English data (§4)
- details on running and replicating baselines (§5)
- details on how we replicated the linguistically informed classifier on German data (§6)

2 Crowdsourcing Experiments for Building Classifiers

All our manual annotation for this research to be used for building classifiers was produced with the help of crowdsourcing using *Prolific*.[1] We made use of crowdsourcing rather than expert annotations since we assume that a gold standard produced by crowdsourcing contains a more representative judgment for abusive language detection, which is a highly subjective task.

For the main task, i.e. determining whether a sentence is abusive or not, we also allowed the crowdworkers to flag sentences they thought were not proper English or they did not understand. Such sentences were subsequently removed from the final gold standard.

[1] `https://www.prolific.co/`

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 one of our surveys were the following:

- Annotators had to be native speakers of English.
- Annotators had to have a task approval rate of 95% or higher.
- Annotators were not allowed to have any literacy difficulties.

In the advertisement for each task on Prolific, we explicitly encouraged crowdworkers with a genuine interest in linguistics to participate in this task.

For each annotation task we also produced detailed 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. If an annotator incorrectly labeled a significant proportion of test instances, the annotator had not understood the annotation task. Subsequently, the annotations of this annotator were not included in our final gold standard.

Each annotator from Prolific was only admitted to one single annotation task. This measure was considered particularly important for our oracle ratings of the individual components of our linguistically informed classifier (i.e. aspectual classification, perpetrator detection and detection of non-conformist views). Only if crowdworkers were restricted to one particular task, would they not be influenced by other competing concepts we wanted to measure separately. We considered this the most unbiased annotation procedure possible.

The final label for our resulting gold standard was the majority label of our 5 different crowdworkers. In the case of the lexicons (e.g. perpetrator detection and fine-grained sentiment), we were even more restrictive. Here, we required 4 out of 5 different crowdworkers to agree in their judgment. We decided to be more restrictive for these tasks since the resulting word lists were used as seeds for a bootstrapping algorithm to create two large lexicons for the respective tasks. For such a setting, a high precision of the initial word seeds is imperative. This procedure is in line with related lexicon induction schemes, such as the induction of abusive words from a large list of negative polar expressions as proposed by Wiegand et al. [2018].

3 Supervised Classifiers

For all supervised classifiers we used in this research we refrained from heavy tuning of hyperparameters. This is due to the fact that most of our experiments were evaluated in a cross-dataset setting, i.e. the training and test data originated from different datasets. As a consequence, tuning hyperparameters would only be possible by using some development data from the source domain. This, however, would mean that the resulting models would be tuned

for the wrong domain. By running the classifiers with frequently used settings of hyperparameters, we hope to produce models that are overall more robust across different domains (i.e. different datasets) than models finetuned on the wrong domain. Thus, we follow the strategy that was proposed for the large scale cross-dataset evaluation reported in Wiegand and Ruppenhofer [2021].

3.1 Computing Infrastructure and Running Time

Our experiments were carried out on a server (Lenovo ThinkSystem SR665; 1TB RAM; 2x32 Core AMD CPU) that is also equipped with a GPU (NVIDIA RTX A40, 48GB RAM). Particularly, the classification experiments on the many transformer-based baselines were very time-consuming. We estimate a total computational budget of 300 GPU hours.

3.2 Monolingual Transformer for English: RoBERTa

We used RoBERTa [Liu et al., 2019] as a representative learning method for state-of-the-art (generic) supervised classification. We made exploratory experiments with both RoBERTa-large and and RoBERTa-base. In general, for each experiment involving a transformer we carried out 5 different runs and considered the average performance of these 5 runs as the overall performance. For most datasets, RoBERTa-large was much more unstable than RoBERTa-base, possessing a high fluctuation in classification performance between the 5 different runs. We also got a considerable amount of runs that just produced a majority-classifier. Our observation was that the more different training and test data were (and since we exclusively focus on cross-dataset classification in this paper, this accounts for many experiments), the more majority classifiers we obtained. We initially assumed the underlying class distribution to be the reason for this behaviour. However, on datasets that have almost a balanced class distribution similar to our test set, such as the dataset from Vidgen et al. [2021], we also observed this issue.

The runs of RoBERTa-large that did not result in a majority classifier were in a similar range as the results from RoBERTa-base. Therefore, we decided to carry out all experiments using RoBERTa-base since this was the most stable classifier that also produced the overall best performance.

For classification, we fine-tuned RoBERTa using the implementation for text classification within the FLAIR framework (version 0.8) [Akbik et al., 2019]. In order **not to overfit the model**, it was trained with **standard hyperparameter settings**:

- learning rate=3e-5, mini batch size=16, mini batch chunk size=4, maximal epochs=5

We maintained the original class distribution of the datasets (both of training and test data) since this is the most realistic setting. Moreover, this is also the way in which recent cross-domain evaluations were conducted [Wiegand et al., 2018, Wiegand and Ruppenhofer, 2021].

For the sake of comparability, for each existing dataset we always trained on the official training set. If no such partition had been defined, we trained on the entire dataset.

3.3 Multilingual Transformer: XLM-RoBERTa

We use the multilingual version of RoBERTa, i.e. XLM-RoBERTa [Conneau et al., 2020] in the same fashion in which we also used the monolingual version of RoBERTa, including the choice of hyperparameter settings (§3.2). The only difference is that our training and test data differ in the language (i.e. training: English; test: German). We thus follow the methodology proposed in previous work [Pires et al., 2019, Zampieri et al., 2020]. As a specific model, we used XLM-RoBERTa-base rather than XLM-RoBERTa-large for reasons of consistency and also since XLM-RoBERTa-large behaved in a similarly unstable manner as RoBERTa-large (§3.2).

3.4 Monolingual Transformer for German

As a monolingual transformer for German, we took the best performing transformer according to Chan et al. [2020]. Again, we fine-tuned the transformer using the implementation for text classification within the FLAIR framework. We used the same hyperparameter settings as in our experiments on English data (§3.2).

3.5 Logistic Regression

For our feature-based classifiers we decided to use logistic regression. We used the implementation within **LIBLINEAR** [Fan et al., 2008] **with L1 regularization**. The advantage of logistic regression is that it is a robust classifier which does **not require any hyperparameter tuning**. This property is particularly suited to our experiments since logistic regression is also used for one of our aspectual classifiers whose training data is created via *distant supervision* [Mintz et al., 2009]. In distant supervision, one typically has training data that is noisy and that may also differ from the actual test data. If we relied on a supervised classifier that required heavy tuning of hyperparameters here, we would likely overfit to the distantly-supervised training set.

4 Linguistically Informed Classifier

Our linguistically-informed classifier consists of three components:

- a classifier which distinguishes between episodic and non-episodic sentences (§4.1)
- a classifier which is able to detect whether an identity group is depicted as a perpetrator (§4.2)
- a classifier based on fine-grained sentiment analysis which is able to detect non-conformist views (§4.3)

With regard to tools and resources, we had to improvise since standard tools and resources mostly produced poor results on the text source we were processing, i.e. tweets. This issue concerns, for example, syntactic parsing, semantic role labeling or further semantic categorization. Our solution to this problem was to approximate the output of more complex (deeper) NLP processing with

simpler (more shallow) NLP tools that exist for Twitter, e.g. a part-of-speech tagger, word lists and regular expressions.

Such procedure is not unheard of. For instance, Riloff et al. [2013] heavily rely on such heuristics in their linguistic approach to identify sarcasm on tweets. Our set of rules was deliberately kept simple and was in no way tuned for our dataset. Thus we wanted to avoid any overfitting. Moreover, since even such a lower bound of our linguistically-informed approach outperformed all classifiers trained on all sorts of previous datasets (as shown in the main paper), this should further highlight the effectiveness of our proposed approach.

4.1 Aspectual Classification

4.1.1 How Training Data were Extracted

We decided to create a new training dataset for aspectual classification via distant supervision [Mintz et al., 2009] since we thought it to be the cheapest solution for our task.

The data for the different classes originate from different text sources. As training data for episodic sentences, we extracted tweets from **news feeds**. Tweets from news feeds typically discuss specific events. They are very similar to news headlines. Therefore, most of these tweets should be episodic.

As news feeds we chose organizations that also report on topics related to our targets (e.g. gazaapost, IsraelNewsNow, LGBT_news).

For non-episodic sentences, we sampled **implied statements** from the Social Bias Frames Corpus (SBIC) [Sap et al., 2020]. In that dataset, the annotators added for each abusive instance (1) the stereotype that the remark alludes to (2).

(1) What do you call a movie with an all-Muslim cast? A box office bomb.
(2) Muslims are all terrorists

Implied statements from SBIC are generalizations. They describe habituality or states. Therefore, they can be considered non-episodic.

Notice that for the tweets to represent our episodic training data, we only considered tweets with a negative sentiment, e.g. by checking whether the tweet contains a negative polar verb. If we had not filtered the tweets, we would have run the risk of classifiers learning a sentiment bias. (That is, the non-episodic training data is exclusively negative while the episodic training data would also contain a significant amount of positive and neutral tweets.)

The use of SBIC as training data for aspectual classification may raise the question why this data was not used for learning the type of implicit abuse we focus on in this paper. First, many of the implied statements contain explicitly abusive words (e.g. *muslims are idiots*). Secondly, most of those examples are not verb-centered (2). Thirdly, Wiegand et al. [2021] observed several significant biases on this dataset which makes targeted learning of implicitly abusive language difficult.

For our final distantly-supervised training set, we randomly sampled 1000 instances from both classes. This should be a fairly unbiased class distribution and size of training data. Larger amounts of training data would have been difficult to acquire since filtering the episodic training data for negative sentiment reduced the amount of available tweets significantly.

4.2 Perpetrator Classifier

The classifier to detect identity groups depicted as perpetrators largely relies on a list of 495 perpetrator-evoking verbs, e.g. *assassinate*, *murder*, *sabotage*, *terrorize*. We identified an identity group as a perpetrator, if the sentence contained a verb according to this verb list and, additionally, the identity group was the agent (i.e. logical subject) of that verb.

As existing tools for semantic role labeling perform too poorly on tweets for detecting agents in our dataset, we simply identified the agent as the noun (phrase) preceding the perpetrator-evoking verb. Instances of this pattern can be easily detected with the help of a part-of-speech tagger that is available for this text domain [Owoputi et al., 2011].

4.3 Fine-Grained Sentiment Analysis for Detecting Non-Conformist Views

In order to detect non-conformist views attributed to identity groups in our dataset, we need to detect two different things:

- Determine the sentiment of the patient of the verb in the sentence.
- Determine the sentiment of the agent toward the patient of the verb in the sentence.

Once these two forms of sentiment have been computed, we simply follow the rules as proposed in our main paper (see Table 5 in the main paper).

4.3.1 Determine the a Priori Sentiment of the Patient

Similar to the detection of agents in §4.2, patients were identified using a set of simple rules operating on the part-of-speech tagged sentence. The patient was identified as a noun phrase immediately following the verb of the sentence.

In order to classify the sentiment of the patient of a verb, we ran a sentiment text classifier on the extracted phrase. As a sentiment text classifier, we used *TweetEval* [Barbieri et al., 2020]. We could slightly increase the detection accuracy of sentiment by further adding the prediction of a transformer-based sentiment classification model provided within *FLAIR* [Akbik et al., 2019] and the sentiment lexicons from Wilson et al. [2005] (*Subjectivity Lexicon*) and Mohammad and Turney [2013] (*NRC lexicon*). Though TweetEval is generally robust, we still found that it is biased towards predicting neutral and positive sentiment. So, we overwrote these sentiment labels if one of the alternative resources suggested a negative sentiment.

Moreover, an additional rule for *protected groups* (e.g. *disabled people*, *poor people*) had to be added. While for several of these groups, traditional sentiment analysis would predict a negative sentiment (e.g. since the negative polar expression *poor* is contained in *poor people*) in our context they are considered as positive. This is so since protected groups are groups our society feels to have an obligation to protect from any harm. In other words, in the context in which we study the mention of protected groups, i.e. (3) and (4), their sentiment is similar to positive polar expressions, like *peace* in (5). Disliking peace is a

classifier	Acc	Prec	Rec	F1
PerspectiveAPI: identity attack	62.0	65.7	57.7	61.4
PerspectiveAPI: toxic	62.6	62.1	62.2	62.2

Table 1: Evaluation of classification performance on our English dataset using the prediction of PerspectiveAPI; two categories are considered *identity attack.* and *toxicity.*

non-conformist view as is disliking poor or disabled people. All these sentences are typically perceived abusive.[2]

(3) Jews dislike [poor people]$^+$.

(4) Jews dislike [disabled people]$^+$.

(5) Jews dislike [peace]$^+$.

4.3.2 Determining the Sentiment of the Agent toward the Patient

This information is encoded in the semantics of the main verb (as detailed in the main paper). We therefore require a resource that reliably encodes this information for verbs. As detailed in the main paper, we produced such a resource by bootstrapping a lexicon of about 1,700 negative verbs from a list of 500 manually annotated verbs.

5 Baselines

5.1 PerspectiveAPI

In our evaluation of the task of abusive language detection, we used *PerspectiveAPI*[3] as one baseline. This tool runs on unrestricted text and is currently considered the state of the art [Röttger et al., 2021]. The tool predicts several subtypes of abusive language. We examined how well the two categories *toxicity* and *identity attack* correlate with abuse on our dataset. That is, we compared the predictions of the tool for both categories against the actual abusive instances in our gold standard. *Toxicity* is the most general type of abusive language that this tool recognizes, while *identity attack* focuses on abusive language towards identity groups. The results are displayed in Table 1. In our main paper, we could only list the performance of one of these classifiers due to space limitations. Since this classifier serves as a baseline and we wanted to include the strongest baseline possible, we chose the slightly more predictive category, i.e. *toxicity.*

[2]If we marked protected groups as negative polar expressions, we could not identify that (3) and (4) convey non-conformist views since disliking something negative is considered acceptable by our society and not considered a non-conformist view.

[3]`www.perspectiveapi.com`

5.2 Aspectual Classification

5.2.1 Why the Tool *sitent* could not be Used for Our Experiments

The only publicly available tool for aspectual classification for English, i.e. *sitent* [Friedrich et al., 2016], has a heavy tense bias. That is, on the data on which that classifier was trained, episodic sentences usually co-occur in the past tense. However, in our dataset virtually all sentences are in present tense, since abusive (stereotypical) remarks are typically written in present tense. As a consequence, *sitent* fails to detect many episodic sentences in our data. Moreover, the feature extraction requires the input text to be syntactically parsed. Unfortunately, the parsing quality on tweets is fairly poor. This further explains the poor performance of this tool.

Feature-based Baseline Classifier The distantly-supervised training data we acquired for the two different classes originates from two different text sources (i.e. news feeds and implied statements). Unlike conventional labeled training data, our resulting dataset is more heterogeneous. As a consequence, we cannot rule out (topic) biases in our training data. For example, there may be some words/topics just occurring in the training data for one of the two classes and this co-occurrence may just be a result of the sampling process. Even state-of-the-art supervised text classifiers, such as transformers, are known to learn these spurious correlations which deteriorate classification performance on unseen test data.

A classifier relying on high-level predictive features, on the other hand, is known to be a more effective solution for noisy and out-of-domain training data [Dias et al., 2009, Mohammad, 2012, Wiegand et al., 2018]. The feature space is considerably lower and by using task-specific features, there is only a low likelihood that spurious correlations are learned by a classifier. Although in the end, we used RoBERTa as a supervised classifier for aspectual classification, we had to evaluate the feature-based classifier in order to compare performance between the two approaches. If the feature-based classifier had outperformed RoBERTa, this would have been an indication of spurious correlations in our training data to which RoBERTa would have overfitted. Luckily, this was not the case. We assume that our measures to produce appropriate (i.e. fairly unbiased) training data (as described in §4.1.1), such as focusing on sentences with a negative sentiment in order to have exclusively negative sentiment in both classes, were successful.

As a reference, Table 2 lists the features we used in our work.

The simplest features were the detection of progressive tense and quantification. These features could easily be detected with the help of part-of-speech tagging and some very few regular expressions.

In order to **avoid any overfitting**, we only wrote word lists ourselves in case they represented a small set of functional words. For instance, for the detection of generalizing adverbial phrases, we devised a small set of such expressions (e.g. *always*, *again*, *all*, *every*). However, for all other features that required some word lists, we employed publicly available lists from the web. For instance, we used such publicly available list for our set of concrete nouns.[4]

For the detection of person names and locations, we did not use a named-entity recognizer due to the lacking performance of standard recognizers on

[4] `https://7esl.com/concrete-nouns/`

feature	example	episodic?
is the sentence in progressive tense?	*Women are unbalancing the world.*	no
is there a mention denoting a specific point in time?	*Lesbians are wrestling right now on Jerry springer.*	yes
is there a generalizing adverbial phrase?	*Muslims slander Christians all the time.*	no
is there some quantification?	*Muslims assassinate 2 Christian aid workers.*	yes
does the verb describe a state?	*Women hate short men.*	no
is there a concrete noun?	*Muslims Steal Ambulance.*	yes
is there a mention of a person name?	*Jews Censor David Duke's Youtube Channel.*	yes
is there a mention of a (specific) location?	*Muslims Brawl At NY Amusement Park.*	yes

Table 2: Feature set for the feature-based aspectual baseline classifier.

tweets. Instead, we used a part-of-speech tagger developed for Twitter [Owoputi et al., 2011] in order to detect mentions of proper nouns (that tagger has a special part-of-speech tag for this category). In addition, in order to check the specific type of proper noun (i.e. person or location), we used supersenses[5] from WordNet [Miller et al., 1990]. For both types of named entities, there is a corresponding supersense, i.e. `noun.person` and `noun.location`. For our feature to detect mentions of a specific point in time, we also relied on the respective supersense in WordNet (i.e. `noun.time`) that contains nouns such as *monday*, *christmas*, *winter*. In order to avoid temporal specifications referring to generalizing events (e.g. *every monday* or *each christmas*), a temporal specification must occur in the absence of a generalizing adverbial.

The detection of *state verbs* was difficult. Unfortunately, we could not find a comprehensive resource that accurately lists all English state verbs.[6] However, we thought it necessary to include a dedicated feature since the fact whether a verb denotes a state or not seems to matter very much for the detection of episodic aspect. State verbs (e.g. *believe*, *love*, *hate*), per definition, typically denote non-episodic aspect. The best approximation we could make out are the emotion verbs as included in the supersense `verb.emotion` of WordNet (e.g. *despise, fear, trouble, worry, upset, yearn*). All emotion words denote states. Still, there may also be states that do not relate to emotions. So our lexicon will not cover all possible types of states. However, given that we are only dealing with sentences that convey negative sentiment, the proportion missed by our approximation should be fairly small.

Ideally, we would also have included a list of *action verbs* (e.g. *fight*, *shout*, *laugh*) since such verbs typically denote episodic aspect. However, finding an appropriate word list of such aspect was even more difficult than finding state verbs.

[5] `https://wordnet.princeton.edu/documentation/lexnames5wn`

[6] For example, the related supersense of WordNet `verb.stative` also includes many verbs that often denote actions, e.g. *crash*, *kill*, *strike*.

5.3 Fine-Grained Sentiment Analysis for Detecting Non-Conformist Views

5.3.1 Shortcomings of Existing Resources for Determining the Sentiment of the Agent toward the Patient

The two resources ***EffectWordNet*** [Choi and Wiebe, 2014] and the ***connotation-frames lexicon*** [Rashkin et al., 2016] contain verb-information that describe the sentiment of the agent towards the patient in a sentence. While the connotation-frames lexicon explicitly encodes such information as evoked by verbs, EffectWordNet uses a different terminology. Still, we found that those verbs labeled as *+effect* can be considered a proxy of verbs having a positive sentiment towards the patient. In a similar fashion, verbs labeled as *-effect* can be considered a proxy of verbs having a negative sentiment towards the patient.

Unfortunately, as mentioned in the paper, we observed significant issues with these two resources. EffectWordNet is a resource that contains an annotation on the sense level. In order to work with these sense-level entries of EffectWordNet directly, we would need some open-domain fine-grained word-sense disambiguation to process our dataset of negative sentences on identity groups. However, we are not aware of any robust tool of that kind. As a practical alternative, we convert the existing sense-level annotation to the lemma level, and subsequently work on the lemma level. Our conversion only considers those verbs in EffectWordNet whose synsets, i.e. the sense units within EffectWordNet they are a member of, all have the same effect type (i.e. *+effect*, *-effect* or none). With this conversion, there is a fairly low likelihood of introducing incorrect effect types. However, we cannot guarantee it will not happen because EffectWordNet does not contain an exhaustive annotation of all synsets and often not all synsets for a lemma are annotated. There is at least the possibility that for some lemma some senses have not been annotated and these senses would actually have a different effect type than the ones that have been annotated. To make it worse, by our restrictive conversion from the sense level to the lemma level, a significant number of lemmas (i.e. more than 25% of the resource) cannot be assigned a definite effect-label.

Fortunately, unlike EffectWordNet, the annotation of the connotation-frames lexicon is already on the lemma level. However, this resource comes with its own challenges. Rather than a categorical label for sentiment, each lemma is assigned a score between -1 and 1. While we assumed positive scores for positive sentiment and negative scores for negative sentiment, we still found several entries whose annotation appeared counter-intuitive to us. For instance, the verbs *lower*, *diminish* and *disrupt* were categorized as conveying a positive sentiment of the agent towards the patient. However, examples from our dataset clearly suggest that the opposite sentiment is true:

(6) $[\text{Jews}]_{agent}$ are lowering $[\text{crime}]_{patient}$.

(7) $[\text{Muslims}]_{agent}$ are diminishing $[\text{people's bad experiences within Islam}]_{patient}$.

(8) $[\text{Jews}]_{agent}$ are disrupting $[\text{the world peace}]_{patient}$.

The verbs *interrupt* and *disturb* that are almost synonymous to *disrupt* have the opposite sentiment in the connotation-frames lexicon. This is inconsistent

and symptomatic for the noise contained in the resource.

Both the connotation-frames lexicon and EffectWordNet suffer from sparsity in general. EffectWordNet covers about 42% of the verbs in our dataset for abusive language detection while the connotation-frames lexicon only covers about 22% of the verbs. The former resource also includes an extension where the annotation of the manually annotated senses was automatically projected to the unlabeled senses of WordNet via some label propagation algorithm. Manual inspection, however, also revealed that the extension contains a high degree of noise.

6 Replicating the Linguistically Informed Classifier on German

In the following subsections, we detail how we replicated the individual components of our linguistically-informed classifier on German data.

6.1 Aspectual Classification for German

Ideally, we would just have trained a multilingual transformer (i.e. XLM-RoBERTa) on our English dataset for this task and classified our German dataset with the resulting model. Unfortunately, the performance of that classifier was too low. Obviously, aspectual classification is too difficult for generic cross-lingual classifiers, such as transformers. Therefore, we built a cross-lingual feature-based classifier for this task. This was motivated by the fact that the feature-based classifier operates on very high-level features that, with only one exception[7], are also language universal (Table 2). Consequently, it was possible to produce a German equivalent without additional training data in German. This was very convenient since we had no dedicated training data for German. (Due to the sparsity of German tweets [Hong et al., 2011] it is much more difficult to produce German training data equivalent to the ones we acquired for English.) Another reason for building a feature-based classifier was that in our evaluation on English data, the feature-based classifier was almost as effective as the (English) transformer.

The feature-based classifier uses a model we trained on English data but can subsequently also be used to classify German data. This is possible since English and German data are converted into the same feature space. All we have to do is to replicate a feature extraction for our German (test) data.

Creating the German equivalent to the word lists was a small effort since the novel lists specially created for this work were fairly small and could be manually translated in a very short amount of time. Reproducing word lists based on WordNet supersenses was not difficult either, since there exists a German equivalent to the English WordNet called *GermaNet* [Hamp and Feldweg, 1997] which also includes a similar set of supersenses. As a part-of-speech tagger, we used the German model of the statistical toolkit *TreeTagger* [Schmid, 1994]. Despite the fact that this model is not specifically tailored to the language of Twitter, we found that it produced sufficiently accurate analyses.

[7]The only high-level feature of our feature set that has no German equivalent is the feature that checks whether the verb is in the progressive. (German does not have such a form.)

6.2 Perpetrator Classifier for German

For the German classifier to detect whether identity groups are depicted as perpetrators, we manually translated the set of 500 negative polar seed verbs that were rated as either *perpetrator-evoking* or *other* to German. Similar to the procedure we applied to English, the resulting set of labeled German verbs was used to train a supervised classifier to predict perpetrator-evoking verbs on a large list of negative polar German expressions. As a set of negative polar expressions for German we used the set that was also considered in the experiments for abusive language detection by Wiegand and Ruppenhofer [2021]. For the supervised classifier (i.e. logistic regression) to predict novel perpetrator-evoking verbs, we represented the verbs by their word embeddings induced on the German *Web as Corpus* [Baroni et al., 2009]. This distributional representation was also proposed for German data by Wiegand and Ruppenhofer [2021] as a substitute for pretrained English embeddings.

6.3 Fine-Grained Sentiment Analysis for Detecting Non-Conformist Views for German

For the German components to establish non-conformist views on the basis of fine-grained sentiment analysis, the component to determine the a priori sentiment towards the patient was more difficult than in the English task. While the *detection* of the patient (phrase) could be achieved similarly as in the English language version, i.e. by writing a small set of rules operating on the output of the part-of-speech tagged sentence, we could not find any publicly available sentiment text classifier for German that worked sufficiently accurate. As a solution, we automatically translated these phrases into English with the help of *Google Translate*[8] and ran the original English sentiment classification on this data.

For determining the sentiment of the agent towards the patient, we manually translated the English set of 500 negative polar verbs that were rated via crowdsourcing to German. Similar to the procedure we applied to English, the resulting set of labeled German verbs was used to train a supervised classifier to predict the fine-grained sentiment on a large list of negative polar German expressions. As a set of negative polar expressions, similar to the classifier to detect perpetrators in German text (§6.2), we used the set of negative polar German expressions from Wiegand and Ruppenhofer [2021]. For training the supervised classifier (i.e. logistic regression) to predict the fine-grained sentiment on these unlabeled negative polar German expressions, the verbs were represented by their word embeddings induced on the German *Web as Corpus* [Baroni et al., 2009].

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the Human Language Technology Conference*

[8] `https://translate.google.com/`

of the North American Chapter of the ACL (HLT/NAACL), pages 54–59, Minneapolis, MN, USA, 2019.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Findings of Association for Computational Linguistics: EMNLP 2020*, 1644–1650, Online, 2020.

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. doi: 10.1007/s10579-009-9081-4.

Branden Chan, Stefan Schweter, and Timo Möller. German's Next Language Model. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 6788–6796, Barcelona, Spain (Online), 2020. doi: 10.18653/v1/2020.coling-main.598. URL `https://aclanthology.org/2020.coling-main.598`.

Yoonjung Choi and Janyce Wiebe. +/-EffectWordNet: Sense-level Lexicon Acquisition for Opinion Inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1181–1191, Doha, Qatar, 2014. doi: 10.3115/v1/D14-1125.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451, Online, 2020.

Gaël Dias, Dinko Lambov, and Veska Noncheva. High-level Features for Learning Subjective Language across Domains. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, San Jose, CA, USA, 2009. URL `http://aaai.org/ocs/index.php/ICWSM/09/paper/view/172`.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

Annemarie Friedrich, Alexis Palmer, and Manfred Pinkal. Situation entity types: automatic classification of clause-level aspect. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1757–1768, Berlin, Germany, 2016.

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.

Lichan Hong, Gregorio Convertino, and Ed Chi. Language matters in Twitter: A large scale study. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, Barcelona, Catalonia, Spain, 2011.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019.

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. doi: 10.1093/ijl/3.4.235.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant Supervision for Relation Extraction without Labeled Data. In *Proceedings of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL/IJCNLP)*, pages 1003–1011, Singapore, 2009. doi: 10.5555/1690219.1690287.

Saif Mohammad. Portable Features for Classifying Emotional Text. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 587–591, Montréal, Canada, 2012. doi: 10.5555/2382029.2382123.

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

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, and Nathan Schneider. Part-of-Speech Tagging for Twitter: Word Clusters and Other Advances. Technical Report CMU-ML-12-107, Machine Learning Department, Carnegie Mellon University., 2011.

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. doi: 10.18653/v1/P19-1493.

Hannah Rashkin, Sameer Singh, and Yejin Choi. Connotation Frames: A Data-Driven Investigation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–321, Berlin, Germany, 2016. doi: 10.18653/v1/P16-1030.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. Sarcasm as Contrast between a Positive Sentiment and Negative Situation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 704–714, Seattle, WA, USA, 2013.

Paul Röttger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet B. Pierrehumbert. HateCheck: Functional Tests for Hate Speech Detection Models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 41–58, Online, 2021.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. SOCIAL BIAS FRAMES: Reasoning about Social and Power

Implications of Language. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5477–5490, Online, 2020. doi: 10.18653/v1/2020.acl-main.486.

Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, United Kingdom, 1994.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. Learning from the wortst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1667–1682, Online, 2021.

Michael Wiegand and Josef Ruppenhofer. Exploiting Emojis for Abusive Language Detection. In *Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL)*, pages 369–380, Online, 2021.

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, 2018. doi: 10.18653/v1/n18-1095.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. Implicitly Abusive Language – What does it actually look like and why are we not getting there? In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 576–587, Online, 2021.

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, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of SemEval*, 2020. doi: 10.5281/zenodo.3950379.