Identifying Implicitly Abusive Remarks about Identity Groups using a Linguistically Informed Approach

Michael Wiegand
Digital Age Research Center (D!ARC)
Alpen-Adria-Universität Klagenfurt
AT-9020 Klagenfurt, Austria
michael.wiegand@aau.at

Elisabeth Eder
Institut für Germanistik
Alpen-Adria-Universität Klagenfurt
AT-9020 Klagenfurt, Austria
elisabeth.eder@aau.at

Josef Ruppenhofer
Leibniz Institute for the German Language
D-68161 Mannheim, Germany
ruppenhofer@ids-mannheim.de

Abstract

We address the task of distinguishing implicitly abusive sentences on identity groups (*Muslims terrorize the world daily*) from other group-related negative polar sentences (*Muslims despise terrorism*). Implicitly abusive language are utterances not conveyed by abusive words (e.g. *bimbo* or *scum*). So far, the detection of such utterances could not be properly addressed since existing datasets displaying a high degree of implicit abuse are fairly biased. Following the recently proposed strategy to solve implicit abuse by separately addressing its different subtypes, we present a new focused and less biased dataset that consists of the subtype of atomic *negative* sentences about identity groups. For that task, we model components that each address one facet of such implicit abuse, i.e. depiction as perpetrators, aspectual classification and non-conformist views. The approach generalizes across different identity groups and languages.

1 Introduction

Abusive language is commonly defined as hurtful, derogatory or obscene utterances made by one person to another person.[1] Examples are (1)-(2).

In the literature, closely related terms include *hate speech* (Waseem and Hovy, 2016) or *cyber bullying* (Zhong et al., 2016). While there may be nuanced differences in meaning, they are all compatible with the general definition above.

(1) stop editing this, you dumbass.
(2) Go lick a pig you arab muslim piece of scum.

Due to the rise of user-generated web content, the amount of abusive language is growing. NLP methods are required to focus human review efforts towards the most relevant microposts. Though there has been much work on abusive language detection in general, there has been little work focusing on **implicit** forms of abusive language (3)-(4) (Waseem et al., 2017). By *implicit* we understand abusive language that is not conveyed by (unambiguously) abusive words (e.g. *bimbo*, *scum*).

(3) Did Stevie Wonder choose these models?
(4) You inspire my inner serial killer.

Detailed analyses of the output of existing classifiers have also revealed that currently only explicit abuse can be reliably detected (van Aken et al., 2018; Wiegand et al., 2019, 2021b).

In this paper, we define implicit abuse as those abusive utterances that lack any abusive word according to the largest lexicon of abusive words available, i.e. the lexicon by Wiegand et al. (2018).

In particular, datasets focusing on abuse towards **identity groups** (*Jews*, *gay people* etc.) contain a high degree of implicit abuse. For example, according to Wiegand et al. (2021b), on the dataset from Waseem and Hovy (2016), 56% of the abusive instances are implicit, while on the dataset from Sap et al. (2020), as many as 62% are.

So far, existing research on implicitly abusive language detection on identity groups has been limited by various biases on existing datasets (Arango et al., 2019; Wiegand et al., 2019), most prominently the **identity-group bias** (Dixon et al., 2018): mentions of identity groups almost exclusively occur in microposts that are considered abusive. As a consequence, most classifiers erroneously learn identity groups as clues for abusive language.

Given that implicit abuse is a challenging problem, Wiegand et al. (2021b) argue that the only

[1] http://thelawdictionary.org

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5600 - 5612

reasonable approach to solve this problem is to address specific subtypes individually rather than consider all types of implicit abuse at once.

In this paper, we address the **task** of distinguishing implicitly abusive remarks on identity groups (5)-(7) from other negative polar sentences (8)-(10). The task is a **binary classification problem**. Each instance is an individual atomic sentence.

(5) Jews succumb to cultural degeneracy. *(ABUSIVE)*
(6) Gay people are contaminating our planet. *(ABUSIVE)*
(7) Women fabricate menopausal symptoms. *(ABUSIVE)*
(8) Jews grieve for Orlando. *(OTHER)*
(9) Gay people are defying stereotypes! *(OTHER)*
(10) Women dread return of Taliban overseas. *(OTHER)*

We create a novel less biased dataset for this task. In NLP, there is an increasing awareness of the importance of producing such data (Gardner et al., 2020). Moreover, Zhou et al. (2021) find that ensuring the quality of datasets during their creation is considerably more effective than even the most sophisticated statistical debiasing techniques.

Unlike previous work, we focus on a **linguistically informed classification approach** and show that this approach is equally effective for different identity groups and can be used to outperform supervised classifiers trained on existing datasets.

We consider only negative polar utterances, since implicitly abusive microposts have a predominantly negative sentiment. For instance, on a random sample of 200 implicitly abusive instances from the dataset by Sap et al. (2020), we could not find a single remark with a positive or neutral sentiment.

Our **contributions** are the following:

- We present the first extensive study on how to detect implicitly abusive remarks among negative atomic remarks on identity groups.
- We establish the predictiveness of 3 linguistic features, namely, aspectual classification, the detection of perpetrators and non-conformist views. The latter two features are addressed for the first time, in general.
- We present a new dataset for this task.
- We introduce new lexical resources for detecting perpetrators and non-conformist views.

This paper only addresses one subset of implicit abuse. However, we consider this focus appropriate, since it is not trivial to detect these instances. As a comprehensive classifier that can detect all these types, we envisage a meta-classifier that collects predictions of individual classifiers designed for different subtypes of abusive language.

All resources created as part of this research are made **publicly available**. They are contained in the supplementary material[2] to this paper, which also includes implementation details.

2 Related Work

Much of the previous work in abusive language detection follows a one-size-fits-all approach (Fortuna and Nunes, 2018). Surveys on existing datasets do not address implicit abuse (Vidgen and Derczynski, 2020; Poletto et al., 2021).

Wiegand et al. (2021b) present a roadmap on implicit abuse arguing that this type of abusive language has not adequately been addressed in previous work. No classification experiments are presented. Next to implicit abuse towards identity groups, they identify as subtypes dehumanization, euphemisms, call for action, multimodal abuse and comparisons. Comparisons are also addressed by Wiegand et al. (2021a) who present the first dataset for this subtype along with classification experiments. The comparisons do not target identity groups. Therefore, our novel dataset and the comparison dataset comprise different sentence types.

Breitfeller et al. (2019) present a study on *microaggressions* which are comments or actions expressing a prejudiced attitude towards marginalized groups unconsciously. Such instances are cases of implicit abuse. Since this is a descriptive study no data for classification are introduced.

Han and Tsvetkov (2020) propose a classification approach for what they call *veiled toxicity*, an umbrella term for many different subtypes of implicit abuse. The approach is evaluated on the dataset by Sap et al. (2020) which Wiegand et al. (2021b) report to have considerable biases.

ElSherief et al. (2021) introduce a general dataset for implicit abuse which is sampled from tweets by hate groups. The authors report biases in the dataset, such as the identity-group bias.

3 Data

As a source for our data, we chose Twitter since it is a platform that contains a high degree of abusive language. We focused on 4 identity groups that cover a range of different characteristics (religion, sexual orientation and gender) and that can also be frequently found in existing datasets. Moreover, they need to occur with sufficient frequency in both

[2] https://github.com/miwieg/naacl2022_identity_groups

languages we are going to examine. The groups are *gay people*[3], *Jews*, *Muslims* and *women*.[4]

The abusive utterances we are looking for are essentially stereotypical sentences on identity groups. Such remarks typically realize the abused target, i.e. the identity group, as the agent (i.e. logical subject) of the verb (5)-(7). Our new dataset focuses only on this argument position since stereotypical remarks usually depict identity groups as the entities performing some action (agent) rather than being affected by it (patient, i.e. logical object). We obtain such utterances by extracting tweets containing mentions of our identity groups followed by a negative polar verb. (This strategy has been proposed by Wiegand et al. (2021b) in order to ensure lexical variability.) The focus on verbs rather than on nouns and adjectives was motivated by the fact that the latter two are more likely to be explicitly abusive words. For example, these parts of speech compose 91% of the lexicon by Wiegand et al. (2018). In this work, we are interested in implicit abuse, however. To test the recall of our sampling approach, we inspect two random samples of 200 abusive (atomic) instances from two popular datasets that focus on identity groups (Sap et al., 2020; Waseem and Hovy, 2016). We find that 80/84% of the instances realize the identity group as an agent. 70/70% of the predicates are verbs, the remainder being adjectives and nouns. Of the verbal predicates, 79/92% were negative polar verbs.

Vidgen et al. (2021b) recently introduced a dataset similar to ours: It focuses on identity groups and also aims at having annotators create suitable non-abusive data. Their goal is to reduce the identity-group bias on their data by a large degree. We refer to this dataset as ***DynaB***. We examined the non-abusive instances in *DynaB* for our 4 identity groups (Table 1) and found that more than 80% of the instances are cases of *reported* abuse (Chiril et al., 2020), as in (11), negations (12), or simply positive or neutral utterances (13).

(11) **It's rude to keep saying** Jews own the media.
(12) Jews do **not** drive climate change.
(13) Jews are **industrious**.

Our dataset, however, consists of **atomic** sentences, i.e. there is no negation or reported abuse (5)-(10). Further, all sentences convey a **negative** sentiment. We believe this to be more challenging since a classifier needs a proper understanding of the atomic utterances themselves rather than looking for positive/neutral sentiment (13) or context clues indicating a non-abusive nesting, such as negation words (e.g. *not* (11)) or reporting verbs (e.g. *say* (12)).

We implemented the following measures proposed by Wiegand et al. (2021b) for producing less biased data for the detection of implicit abuse.

- Our data is sampled from one textual source, i.e. Twitter. Both abusive and non-abusive sentences are sampled by the same pattern (i.e. mention of identity group preceding a negative verb). Thus no biases are caused by merging instances from different text sources.
- In order to avoid any user biases, tweets were sampled from a wide set of different users. The average number of tweets per user is 1.1.
- In order to avoid a focus on frequently occurring verbs, we sampled our dataset from a wide set of negative polar verbs.[5] On average, each verb occurs twice in the final dataset. Unlike previous datasets, this sampling strategy thus puts due emphasis on the "long tail" of the verb distribution.
- We only included sentences that do **not** contain explicitly abusive words. Otherwise, classifiers could easily detect the respective abusive utterances since they would just have to focus on these explicit clues.
- We remove any text co-occurring with our sentences that might give rise to spurious correlations, e.g. hashtags or user names. We observed that particularly hashtags, such as *#banIslam* or *#feminismIsCancer*, often strongly correlate with abusive tweets. Such hashtags display a behaviour similar to explicitly abusive words.

We created a gold standard for English and another, less-resourced language, German. Exactly the same sampling procedure was applied to both datasets. However, due to the sparsity of German language content on Twitter (Hong et al., 2011), the German dataset is smaller.

Both datasets were **annotated via the crowdsourcing platform Prolific**.[6] The label of each

[3]For this group, we used the terms *gay people* and *lesbians*. Other expressions, such as *gays* or *queer*, were too infrequent.

[4]Ideally, we would also have included *black people* as an additional identity group. However, it was not possible to obtain a sufficient amount of implicitly abusive data for this identity group in both languages that we consider in this paper.

[5]We used the list of negative polar verbs contained in the resources by Wiegand et al. (2018).

[6]https://www.prolific.co/

property	English	German
sentences	2221	970
abusive sentences	56.24%	52.16%
non-abusive sentences	43.76%	47.84%
sentences on gay people	403	154
sentences on Jews	545	184
sentences on Muslims	782	367
sentences on women	491	265
no. of unique verbs	965	534
avg. frequency of verbs	2.30	1.82
avg. sentence length (in tokens)	7.75	7.00
avg. no. of sentences per user	1.05	1.10

Table 1: Statistics of the datasets

instance represents the majority vote of 5 different crowdworkers, who were native speakers. We opted for a very high approval rate (i.e. 95% or higher) in order to guarantee a sufficiently high annotation quality. (*The supplementary material contains annotation guidelines.*) Table 1 offers some descriptive statistics.

On a random sample of 200 sentences, we computed the agreement between the majority vote of our crowdsourced judgments and one co-author of this paper. We measured substantial agreement of $\kappa = 0.87$ on the English and $\kappa = 0.82$ on the German dataset (Landis and Koch, 1977).

4 Supervised Classifiers and Evaluation

We consider RoBERTa (Liu et al., 2019) as a baseline for generic supervised classification for English data. For our German data, we use the best transformer according to Chan et al. (2020). We fine-tune the pretrained models on the given task using the FLAIR framework (Akbik et al., 2019). (*The supplementary notes contain more details on all classifiers including hyperparameter settings.*)

As evaluation measures, we use macro-average precision, recall, F1-score. For all classifiers built with transformers, we report the average over 5 training runs (including standard deviation). All other classifiers produce deterministic output.

5 Linguistically Informed Classifier

We propose a linguistically informed classifier which models 3 component tasks. We describe how this classifier is built for English. The component tasks represent concepts which have been suggested to be predictive for this task (Wiegand et al., 2021b) but, so far, could not be tested due to the lack of data. In order to avoid overfitting, each component comes with a separate classifier being built on training data different to the test data of our main task. Since **we manually labeled our dataset also for each of the component tasks[7] we can conduct an intrinsic evaluation of each component, too**. In order to have an unbiased annotation, each crowdworker was only allowed to participate in exactly one of our annotation tasks.

5.1 Component 1: Aspectual Classifier

The Task. In our first task we address aspectual classification. Abusive utterances regarding identity groups are usually stereotypes (Sap et al., 2020). Per definition, stereotypes coincide with habitual (or non-episodic) aspect (14)-(15). On the other hand, episodic aspect (16)-(17), i.e. utterances that express information about a single event (Friedrich and Pinkal, 2015), despite the fact that they may be tendentious (Mendelsohn et al., 2021) or even be cases of fake news (Zhou and Zafarani, 2020), is more likely to be non-abusive. We **distinguish between episodic and non-episodic sentences**.

(14) Muslims are vandalising Hindu temples every day. (*non-episodic*)
(15) The Jews damage our souls. (*non-episodic*)
(16) Muslims vandalise newspaper offices in Odisha over publication of Mohammed's images. (*episodic*)
(17) Jews damage olive trees in West Bank. (*episodic*)

The Method. Aspectual classification was investigated by Friedrich and Pinkal (2015) and an implementation of their classifier is available as part of ***sitent*** (Friedrich et al., 2016). However, we observed substantial issues with *sitent* when applied to our data. The tool was trained on Wikipedia and MASC (Ide et al., 2008). On these datasets, episodic aspect is biased towards past tense. However, our data originates from Twitter and both episodic and non-episodic sentences co-occur in present tense.

As a consequence, we decided to build a classifier from scratch. As no suitable labeled training data for our domain (i.e. social media) is available, we decided to apply a form of *distant supervision* (Mintz et al., 2009). As a proxy for episodic sentences, we sampled tweets from news feeds (e.g. *LGBT_news* or *GazaTVNews*) from Twitter. Such tweets typically report on specific events (18)-(19).

(18) Israel strikes Iranian targets inside of Syria.
(19) North Texas Student Expelled for Being Gay

For the non-episodic sentences, we considered the *implied statements* (21) from the *social bias frames*

[7] For all component tasks, we obtained a substantial agreement with the lowest being at $\kappa = 0.65$ (detection of perpetrators) using the same random sample as for the main task.

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 of the feature-based aspectual (baseline) classifier (*more details in the supplementary notes*).

	majority-class	sitent	feature-based	RoBERTa
F1	38.4	53.0	76.6	**76.9** (±1.0)

Table 3: Evaluation of aspectual classification.

corpus (Sap et al., 2020). In that dataset, the annotators added for each abusive instance (20) the stereotype that the remark alludes to (21).

(20) What do you call a movie with an all-Muslim cast? A box office bomb.

(21) *implied statement:* Muslims are all terrorists

For our training set, we randomly sampled 1000 news tweets (=episodic) and 1000 implied statements (=non-episodic). As classifiers, we trained RoBERTa and a feature-based baseline. The latter was included since generic supervised classifiers (such as RoBERTa) are susceptible of learning spurious correlations contained in training data. Such correlations cannot be ruled out as our training data for the two classes was sampled from different sources. Our feature-based baseline, which is a logistic regression trained on high-level features that are fairly domain independent, makes such overfitting less likely. The features for detecting episodic sentences check for mentions of concrete entities or a specific point in time, while features for non-episodic sentences try to detect states and generalizations. Table 2 lists the full feature set.

Table 3 shows the result of the different classifiers on our English dataset (§3). *sitent* performs poorly. We attribute it to the tense bias reported above. The feature-based baseline is strong but it does not outperform RoBERTa. Therefore, RoBERTa does not seem to be seriously affected by spurious correlations. We use the output of RoBERTa in all subsequent experiments. In order to facilitate the combination with other components of our classifier, we use the majority vote of the 5 runs of this classifier.

5.2 Component 2: Perpetrator Classifier

The Task. A common stereotype that can be observed with every identity group is the depiction as perpetrators (22)-(24). By perpetrators, we understand persons who commit an illegal, criminal, or evil act.[8] Although different identity groups are typically depicted as different perpetrators (e.g. Muslims are depicted as terrorists (22), women are considered to be dishonest (23), while gay men are accused of being pedophiles (24)), all these stereotypes describe actions that involve criminal offenses (e.g. raping, stealing) or morally contemptible behaviour (e.g. adultery, lying). We think it is most economical to frame the detection of perpetrators as a single task.

(22) [Muslims]$_{agent}$ terrorize the world daily.

(23) [Women]$_{agent}$ betray their partners.

(24) [Gay people]$_{agent}$ are raping our children.

We consider the task a form of semantic role labeling (Gildea and Jurafsky, 2002), i.e. perpetrators are specific entities evoked by particular verbs. Therefore, we need to **find perpetrator-evoking verbs** (e.g. *terrorize*, *betray*, *rape*) **and the respective argument position of the perpetrator**.

The Method. In order to obtain a labeled dataset of perpetrator-evoking verbs, we randomly sampled 500 negative polar verbs from the *Subjectivity Lexicon* (Wilson et al., 2005) and asked crowdworkers to form simple sentences (only a main clause) in which the given verb evokes an event that includes some perpetrator. The 500 verbs are in no way tuned for our test data (§3).[9] Since we do not want crowdworkers to invent any anti-Semitic, homophobic, Islamophobic or misogynist sentences, we invented a fictitious people whose name has no phonetic resemblance to existing identity groups. The crowdworkers were asked to depict these people as perpetrator, if possible. Obviously, plausible sentences can only be formed with the subset of perpetrator-evoking verbs we are looking for. For other verbs, such as *grieve* or *dread*, forming such

[8] www.dictionary.com/browse/perpetrator

[9] Our English dataset contains 965 verbs of which only 373 can be found among the 500 from the Subjectivity Lexicon.

sentences is not possible. Therefore, crowdworkers were asked not to provide a sentence in case they felt that they were unable to meet the criterion of constructing a context with a perpetrator being a participating entity of the event evoked by the given verb. Only if the majority of 5 crowdworkers managed to produce such sentences for the same verb, did we consider it as a perpetrator-evoking verb. This setting also allowed us to identify the semantic role of the perpetrator. Overall, 165 out of 500 verbs were identified as perpetrator-evoking verbs. In 96% of the respective sentences, the semantic role of the perpetrator was the agent of the verb (as in (22)-(24)).

In a second step, we extended the list of perpetrator-evoking verbs. Our aim is to obtain a (nearly) exhaustive list of perpetrator-evoking verbs. Therefore, we train a classifier on our 500 verbs (each verb labeled as either *perpetrator-evoking* or *other*) and classify each verb from the largest list of publicly available negative polar verbs. We took the verbs from the set of negative polar words from Wiegand et al. (2018) (totaling 1,700 negative verbs). We trained a logistic regression classifier where each verb was represented by its (publicly available) word embedding induced on Common Crawl (Mikolov et al., 2018).[10] We ended up with 491 perpetrator-evoking verbs. Our **lexicon-based classifier** identifies a perpetrator if it is observed as an agent of one of these 491 verbs. This classifier is run on our dataset (§3). The output is evaluated against the gold annotation for this component task.

As a baseline, we run a very fine-grained semantic-role labeling system based on **FrameNet** (Baker et al., 1998) on our data. We chose *open sesame* (Swayamdipta et al., 2017) which is the most recent publicly available tool for semantic-role labeling based on FrameNet. Due to its fine-grained inventory, there are frame elements (this is the term for semantic roles in FrameNet) which semantically correspond to our concept of perpetrators. More precisely, we considered text spans as perpetrators if they are predicted to be one of the following frame elements: *Abuser*, *Assailant*, *Counter_actor*, *Destroyer*, *Invader*, *Killer*, *Manipulator*, *Offender*, *Perpetrator* and *Wrongdoer*.

Table 4 shows the performance of the different classifiers to detect mentions of perpetrators in our English dataset (§3). Our lexicon-based classifier outperforms FrameNet, which is known to have a limited lexical coverage (Das and Smith, 2011).

	majority-class	FrameNet	lexicon-based classifier
F1	33.9	60.1	**70.5**

Table 4: Evaluation of perpetrator classification.

5.3 Component 3: Non-Conformist Views

The Task. For our third component task, we consider the sentiment of the agent towards the patient (as conveyed by the main verb in the sentence) in combination with the sentiment expected a priori towards the patient. (The agent is always the mention of the identity group.) This is illustrated in Table 5. We observe a systematic relationship between abusive language and fine-grained sentiment: **If the sentiment of the identity group (i.e. the agent) towards the patient is opposite to the prior sentiment of the patient, then this utterance depicts the identity group as having a non-conformist view.**[11] Such views are perceived as abusive utterances: If someone attributes non-conformist views to some identity group, then, one often intends to stigmatize this group as not belonging to their own community. This phenomenon is referred to as *othering* (Burnap and Williams, 2016).

The Method. In order to detect the above pattern indicating non-conformist views, we need the output of two modules: the first determines the prior sentiment of the patient (i.e. the phrase representing the logical object); the second determines the sentiment of the agent towards the patient. The prior sentiment of the patient can be easily detected by running a sentiment text classifier on that phrase. For this, we use *TweetEval* (Barbieri et al., 2020).

The difficult part is to detect the sentiment of the agent towards the patient. Sentiment text classifiers are unable to determine such fine-grained sentiment information. They capture the general sentiment of a given text which may be different. For instance, (25) conveys a positive sentiment of *Muslims* towards *violence*, while the sentiment of the sentence is generally considered negative.

(25) $[\text{Muslims}]_{agent}$ **glorify** $[\text{violence}]_{patient}$.

[10] https://dl.fbaipublicfiles.com/fasttext/vectors-english/crawl-300d-2M.vec.zip

[11] In Table 5, we only distinguish between positive and negative sentiment. There is no neutral sentiment. In the context of these sentiment patterns, we found that neutral sentiment follows the same pattern as positive sentiment. Conflating positive and neutral sentiment facilitated automatic processing.

example sentences *(2 sentences for each type; all sentences are non-episodic)*		fine-grained sentiment: agent to patient	fine-grained sentiment: patient	abuse
[Jews]$_{agent}$ long for [a safe Israel]$_{patient}$. [Women]$_{agent}$ abhor [violence]$_{patient}$.	[Muslims]$_{agent}$ grieve for [their brothers]$_{patient}$. [Jews]$_{agent}$ suffer from [ethnic cleansing]$_{patient}$.	positive negative	positive negative	
[Lesbians]$_{agent}$ pray to [Satan]$_{patient}$. [Muslims]$_{agent}$ dislike [peace]$_{patient}$.	[Muslims]$_{agent}$ revert to [stoning victims]$_{patient}$. [Lesbians]$_{agent}$ disrespect [God's plan]$_{patient}$.	positive negative	negative positive	✓ ✓

Table 5: Implicitly abusive language and fine-grained sentiment; **non-conformist views** are sentences in which *sentiment of agent to patient* and *sentiment of patient* **disagree**; non-conformist views coincide with abuse.

	majority-class	frames	effectWN	novel lexicon
F1	41.9	64.1	64.8	**71.6**

Table 6: Evaluation of fine-grained sentiment analysis.

Instead of a text classifier, we seek a lexicon that specifies for any negative polar verb (out of context) whether it conveys a positive sentiment of the agent towards the patient (e.g. *glorify*, *long (for)*, *pray (to)*) or a negative sentiment towards it (e.g. *abhor*, *dislike*, *suffer*). The only lexicons with such information are *EffectWordNet* (***effectWN***) (Choi and Wiebe, 2014) and the *connotation-frames lexicon* (***frames***) (Rashkin et al., 2016). Unfortunately, both resources only cover about 40% of the verbs in our dataset. We also determined a significant level of noise in these resources (*as detailed in the supplementary notes*). Therefore, we decided to create a novel lexicon with that information. It should cover all possible negative verbs. We first had crowdworkers annotate for some (seed) negative verbs the sentiment of the agent towards the patient. We chose the 500 verbs we already used in §5.2. The majority of the crowdworkers' judgements represent our gold standard annotation. On these annotated 500 verbs we trained a logistic regression classifier. As features, we represented each verb by its word embedding from Common Crawl. (Using such a representation is common practice for this task (Rashkin et al., 2016).) The resulting classifier was run on the same large set of 1,700 negative verbs we used in §5.2. For each verb the classifier predicts the sentiment of the agent towards the patient. The result is our novel lexicon. Since we have also manually annotated the sentiment of the agent towards the patient for each verb in the sentences of our labeled dataset (§3), we can evaluate this lexicon against our labeled dataset.

Table 6 evaluates the different lexicons to determine the sentiment of the agent towards the patient on our novel dataset. The table shows that our novel bootstrapped lexicon produces a notable improvement over the existing resources.

procedure isImplicitlyAbusive(*sentence*)
  abusive ← **FALSE**
  if not (**getAspect**(*sentence*) == EPISODIC) **then**
    if hasPerpetrator(*sentence*) **then**
      abusive ← **TRUE**
    else if hasNonConformistView(*sentence*) **then**
      abusive ← **TRUE**
  return *abusive*

Figure 1: Linguistically informed classifier.

5.4 How the Final Classifier is Built

Figure 1 shows how the component tasks introduced in §5.1-5.3 are combined to produce our linguistically informed classifier: We consider those sentences as abusive that are non-episodic and which either depict the identity group as perpetrator or attribute non-conformist views to it. We use the best-performing component classifiers as determined by our previous evaluation (§5.1-§5.3).

We also experimented with a supervised classifier that uses the predictions from our component classifiers as features. However, since the classification performance was on a par with our proposed (rule-based) classifier (Figure 1), we decided in favor of the latter classifier. It has a clear advantage over supervised classification in that it does not require any labeled training data to combine the predictions of the component classifiers.

6 Evaluation on English Dataset

We evaluate the linguistically informed classifier (Figure 1) on our new English dataset (for implicit abuse) against other classifiers trained on existing datasets. We carry out a **cross-dataset evaluation**: None of the classifiers, including our linguistically informed classifier, has been trained on our English dataset. Given the recent criticism against within-dataset evaluation (Arango et al., 2019; Wiegand et al., 2019) in which high performance is often the result of overfitting, this is a fairly unbiased set up.

As datasets for training supervised baselines, we chose those that focus on implicit abuse (ElSherief et al., 2021) or abuse towards identity groups

	ABUSIVE			OTHER			AVERAGE		
training data	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1** (*std*)
majority-class classifier	56.2	100.0	72.0	0.0†	0.0	0.0†	28.1	50.0	36.0
(Vidgen et al., 2021a)*	50.0	58.1	53.7	53.3	54.4	53.8	51.6	56.2	52.8 (±1.1)
(Waseem and Hovy, 2016)*	63.0	22.1	32.7	45.4	**82.9**	58.7	54.2	52.5	53.3 (±1.3)
(Founta et al., 2018)*	65.5	61.4	63.4	54.1	56.5	55.3	59.8	59.0	59.4 (±2.1)
(Sap et al., 2020)*	61.5	90.4	73.2	71.4	26.1	38.2	66.4	58.3	62.0 (±4.3)
PerspectiveAPI	67.2	65.3	66.2	57.0	59.1	58.0	62.1	62.2	62.2
(ElSherief et al., 2021)*	70.5	57.8	63.5	55.9	67.3	61.1	63.2	62.6	62.9 (±3.4)
DynaB (Vidgen et al., 2021b)*	61.1	**98.0**	75.3	**88.4**	19.6	32.1	**74.8**	58.8	65.8 (±2.2)
linguistically informed classifier	75.2	76.0	75.6	68.7	67.8	68.2	72.0	71.9	71.9
linguistically informed classifier + DynaB*	**78.1**	74.9	**76.5**	69.3	73.0	**71.1**	73.7	**73.9**	**73.8** (±0.5)
linguistically informed classifier (oracle)	81.3	79.1	80.2	74.0	76.5	75.2	77.6	77.8	77.7
human classifier (upper bound)	81.7	85.4	83.5	82.1	77.8	79.9	81.6	81.9	81.8

Table 7: Cross-dataset evaluation on English dataset (†: *strictly speaking the value for this score is not defined, however, following common practice we considered it* 0 *which enables the computation of the average score*; *: *RoBERTa has been used as classifier*).

(Waseem and Hovy, 2016; Sap et al., 2020; Vidgen et al., 2021a,b). We also included Founta et al. (2018) as a more general dataset sampled from Twitter. For each dataset, we fine-tune the pretrained RoBERTa model (§4) on the training partition of the respective dataset. As a further baseline, we run the state-of-the-art classifier for abusive language detection *PerspectiveAPI*[12] on our dataset.

We also include an **oracle** version of our linguistically informed classifier, that combines the gold standard annotation for the component tasks (§5.1-§5.3) rather than the outputs of the respective classifiers. This can be considered the upper bound for the linguistically informed classifier.

Finally, we also consider a **human classifier** as a general upper bound. We randomly sampled the judgment of one individual annotator from the crowdsourced gold-standard annotation for the detection of abusive language. This individual judgement may notably differ from the gold standard label which is the majority label of 5 annotators.

Table 7 displays the results. The classifiers trained on existing datasets do not perform well on our new dataset. The best classifier among them is the one trained on the *DynaB*-dataset. For *DynaB* (unlike the other datasets), special attention was paid to the inclusion of non-abusive instances (§3). Still, our linguistically informed classifier is more effective. *DynaB* suffers from the *identity-group bias* (§1): its recall for non-abusive instances is only at 20%. As detailed in §3, *DynaB* focuses on non-abusive *nesting* of abusive statements (such as (11) or (12)). However, it only contains very few non-abusive *atomic* utterances (8)-(10). With about 68%, our linguistically informed classifier has no perfect recall for non-abusive instances either. Since both this classifier and *DynaB* have in general a high precision (with *DynaB* having the highest of all classifiers), it makes sense to combine them in order to raise the overall recall. We combine the two classifiers by predicting a non-abusive sentence if one of the two classifiers predicts one. This combination further increased performance. Thus we could outperform *DynaB* by 8%-points in macro-average F1.

[12] www.perspectiveapi.com

targets	**perpetr.**	**non-conf. views**	**aspect**	**combined**
gay people	67.6	62.0	68.4	**70.6**
Jews	61.2	62.4	67.0	**71.8**
Muslims	58.5	62.6	66.9	**72.5**
women	63.0	61.2	70.0	**71.4**
all	62.0	62.5	67.7	**71.9**

Table 8: Evaluation of different linguistic components on the different targets (*evaluation measure: F1-score*).

The strong performance of the oracle version of our linguistically informed classifier (77.7% F1) is proof that our 3 linguistic concepts are predictive of abuse on identity groups. The fact that it outperforms our best automatic solution (73.8% F1) suggests that there is still room for improvement.

Table 8 examines the performance of the individual components of our linguistically informed classifier. Since the combined classifier outperforms every individual classifier, we can conclude that the information contained in the components is complementary to a certain degree.

Table 8 also shows that the individual components are effective across the 4 targets which suggests that they are target independent.

7 Evaluation on German Dataset

Our final experiments focus on our German dataset. As baselines, we consider supervised classifiers (§4) trained on the German datasets for abusive lan-

training data & classifier	Prec	Rec	F1 (std)
majority-class classifier	26.1	50.0	34.3
GermEval-2021* [Facebook]	65.8	55.7	60.2 (±3.7)
GermEval-2019* [Twitter]	69.5	59.3	63.9 (±4.9)
linguistically informed classifier	70.7	70.6	70.6
ling. inf. class.+GermEval-2019*	73.4	72.6	73.0 (±1.6)
English-dataset (XLM-RoBERTa)	**81.1**	**80.7**	**80.9** (±0.8)
ling. informed classifier (oracle)	82.9	83.0	82.9
human classifier (upper bound)	87.9	87.8	87.8

Table 9: Evaluation on German dataset
(*: using best transformer from Chan et al. (2020)).

guage detection, i.e. *GermEval 2019* (Struß et al., 2019) and *GermEval 2021* (Risch et al., 2021). Next to a classifier that replicates our linguistically informed classifier on German data, we also test a cross-lingual classifier. Following previous work (Zampieri et al., 2020), we fine-tune the multilingual transformer XLM-RoBERTa (Conneau et al., 2020) on our English dataset. Since this language model also covers German, the resulting classifier can also be applied on our German dataset.

Table 9 shows the results. The human baseline is notably higher than on the English dataset. German tweets are predominantly posted by native speakers resulting in more fluent language. This makes the manual annotation for the human baseline easier.

With the linguistically informed classifier we outperform both *GermEval* classifiers. The oracle version is even notably better. These results suggest that the linguistic properties of our 3 components are language independent. The fact that the multilingual transformer performs best indicates that, in general, the type of implicit abuse we address in this work, is valid across different languages.

8 Discussion

As the performance of our oracle classifier shows, even a perfect linguistically informed classifier is still below human performance. We could identify two types of ambiguous utterances in our misclassifications that may be responsible: A few sentences are underspecified as to whether they report facts or reflect the author's opinion being biased by their stereotypical views (26)-(27). Only the interpretation as an opinion is perceived abusive.

(26) Women overuse makeup.
(27) Muslims suppress Christian life in Iraq.

Moreover, the prior sentiment of the patient may occasionally depend on the ideology of the reader. For instance, atheists may consider (28) abusive while religious persons would not. Similarly, feminists and non-feminists may have a different perception of (29). It may be debatable that unique class labels as we have assigned to (26)-(29) are adequate. One may argue that without further context these ambiguities cannot be properly resolved.

(28) Muslims surrender to God's will.
(29) Women unmake patriarchy.

A general limitation of our approach is that our data exclusively originate from Twitter. Therefore, we cannot rule out that certain results reported in this paper only hold for data from this platform. Given, however, that we made sure that the data from that platform that we use are not affected by any obvious user or topic biases (§3) and given that our proposed method works across 4 different identity groups and 2 different languages, we estimate the likelihood that this limitation has significantly affected our results to be very low.

Another limitation of our work is the focus on atomic sentences in which the identity group is the agent of some negative verb. As we have motivated in §3, our exploratory data analysis suggests that this is the most frequent surface realization of such abusive remarks. However, implicitly abusive remarks targeting identity groups may also be expressed in other ways, such as (30) where the identity group is not an agent of some negative polar verb.

(30) Once again, we find Jews and money money money.

While constructions such as (30) are possible, we are unaware of any sampling method that would enable us to capture such constructions. We expect these constructions also to be more infrequent than the more prototypical atomic sentences. Therefore, we leave it to future work to address them.

9 Conclusion

We presented a new focused dataset for implicitly abusive remarks among negative polar utterances on identity groups. We identified 3 linguistic properties which allow us to effectively detect such abusive remarks across different identity groups and across different languages. The utterances have to be non-episodic and the identity group is either depicted as a perpetrator or attributed to a non-conformist view. We are also able to notably outperform classifiers trained on previous datasets.

10 Ethical Considerations

Most of our new gold standard data were created with the help of crowdsourcing. All crowdworkers were compensated following the wage recommended by the crowdsourcing platform Prolific (i.e. $9.60 per hour). Since we were aware of the offensive nature of the data that the crowdworkers had to annotate, we inserted a respective warning in the task advertisement. In order to keep the psychological strain of the crowdworkers at an acceptable level, the data to be annotated was split into bins of 100-200 instances. Furthermore, we allowed each crowdworker to take part in one single task only. We also made it very clear in the task description that we follow a linguistic purpose with our crowdsourcing tasks and the opinion expressed in the sentences to be annotated in no way reflects the opinion of (us) researchers designing the tasks.

One of our crowdsourcing tasks included inventing sentences in which a group of people is framed as a perpetrator (§5.2). Since we did not want crowdworkers to invent any anti-Semitic, homophobic, Islamophobic or misogynist content, we introduced the name of a fictitious people which the crowdworkers were to use in their sentences. We also made sure that the particular name did not have any obvious phonetic resemblance to existing identity groups. Although the resulting sentences being invented are not directed against any existing identity groups they may still be considered abusive. However, we think that this is justifiable in this particular context since we are not aware of any existing dataset that contains a similar content (i.e. a focused dataset for learning perpetrator-evoking verbs) that we could have used for our experiments. In principle, creating morally disputable content as part of research is not unusual. Both in plagiarism detection (Potthast et al., 2010), deception detection (Ott et al., 2011) and, quite recently, abusive language detection itself (Vidgen et al., 2021b; Wiegand et al., 2021a) a procedure similar to ours was pursued.

One substantial part of the data we are going to make publicly available as part of this research will include sentences extracted from Twitter. In order to protect the privacy rights of the authors of the tweets and individuals mentioned in them, we anonymized our data by discarding mentions of usernames. The public release of a limited number of tweets as in the range of our dataset is also in accordance with the regulations of Twitter.

A datasheet describing our novel dataset of labeled sentences for the task of detecting implicitly abusive remarks about identity groups (both English and German version) following the specification of Gebru et al. (2018) was added to the supplementary material.

Our current data focuses on the four identity groups *Jews*, *Muslims* and *gay people* and *women*. This choice was mainly motivated by the fact that these groups are among the most abused identity groups on social media. As a consequence, it was also possible to obtain a reasonable amount of data (even with our restrictive measures to ensure less biased datasets). Moreover, these identity groups are well represented in existing datasets. This allows us to compare our proposed classifier against baseline classifiers trained on these existing datasets. We acknowledge that abusive language on the web is also directed against other identity groups. We leave their automatic detection to future work. However, our study suggests that abusive language that targets these other identity groups will follow the same language patterns as the instances of abusive language examined in this paper.

11 Acknowlegdgements

The authors would like to thank Sybille Sornig for manually annotating parts of the data on which our descriptive statistics in Section 3 are based. We are also grateful to Ines Rehbein for feedback on earlier drafts of this paper.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 54–59, Minneapolis, MN, USA.

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate Speech Detection is Not as Easy as You May Think: A Closer Look at Model Validation. In *Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR)*, pages 45–53, Paris, France.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. 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.

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

Luke M. Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1664–1674, Hong Kong, China.

Pete Burnap and Matthew L. Williams. 2016. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, 5(1):11.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's Next Language Model. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 6788–6796, Barcelona, Spain (Online).

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020. He said "who's gonna take care of your children when you are at ACL?": Reported sexist acts are not sexist. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4055–4066, Online.

Yoonjung Choi and Janyce Wiebe. 2014. +/-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.

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 Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451, Online.

Dipanjan Das and Noah A. Smith. 2011. Semi-Supervised Frame-Semantic Parsing for Unknown Predicates. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1435–1444, Portland, OR, USA.

Lucas Dixon, John Li, Jeffrey S. Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 67–73, New Orleans, LA, USA.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 345–363, Online and Punta Cana, Dominican Republic.

Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, 51(4):85:1–85:30.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, Stanford, CA, USA.

Annemarie Friedrich, Alexis Palmer, and Manfred Pinkal. 2016. 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.

Annemarie Friedrich and Manfred Pinkal. 2015. Automatic recognition of habituals: a three-way classification of clausal aspect. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2471–2481, Lisbon, Portugal.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating Models' Local Decision Boundaries via Contrast Sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1320, Online.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for Datasets. In *Proceedings of the Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT ML)*, Stockholm, Sweden.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3):245–288.

Xiachuang Han and Yulia Tsvetkov. 2020. Fortifying Toxic Speech Detection Against Veiled Toxity. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7732–7739, Online.

Lichan Hong, Gregorio Convertino, and Ed Chi. 2011. Language matters in Twitter: A large scale study. In *Proceedings of the International AAAI Conference*

on Weblogs and Social Media (ICWSM), Barcelona, Catalonia, Spain.

Nancy Ide, Collin Baker, Christiane Fellbaum, Charles Fillmore, and Rebecca Passonneau. 2008. MASC: the Manually Annotated Sub-Corpus of American English. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, pages 2455–2461, Marrakech, Morocco.

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

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

Julia Mendelsohn, David Jurgens, and Ceren Budak. 2021. Modeling Framing in Immigration Discourse on Social Media. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, 2219–2263, Online.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advanced in Pre-Training Distributed Word Representations. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, pages 52–55, Miyazaki, Japan.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. 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.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 309–319, Portland, OR, USA.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An Evaluation Framework for Plagiarism Detection. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 997–1005, Beijing, China.

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

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12, Online.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. 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.

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 *Proceedings of the GermEval Workshop*, pages 352–363, Erlangen, Germany.

Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A. Smith. 2017. Frame-Semantic Parsing with Softmax-Margin Segmental RNNs and a Syntactic Scaffold. *arXiv preprint arXiv:1706.09528*.

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for Toxic Comment Classification: An In-Depth Error Analysis. In *Proceedings of the Workshop on Abusive Language Online (ALW)*, pages 33–42, Brussels, Belgium.

Bertie Vidgen and Leon Derczynski. 2020. Directions in Abusive Language Training Data. *PLoS One*, 15(12).

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021a. Introducing CAD: the Contextual Abuse Dataset. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 2289–2303, Online.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021b. Learning from the Worst: 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.

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

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

Michael Wiegand, Maja Geulig, and Josef Ruppenhofer. 2021a. Implicitly Abusive Comparisons – A New Dataset and Linguistic Analysis. In *Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL)*, pages 358–368, Online.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021b. 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.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. 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.

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

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 347–354, Vancouver, BC, Canada.

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 SemEval*, page 1425–1447, Online.

Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J. Miller, and Cornelia Caragea. 2016. Content-Driven Detection of Cyberbullying on the Instagram Social Network. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3952–3958, New York City, NY, USA.

Xinyi Zhou and Reza Zafarani. 2020. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Surveys*, 53(5):109:1–109:40.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2021. Challenges in Automated Debiasing for Toxic Language Detection. In *Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL)*, pages 3143–3155, Online.