Euphemistic Abuse – A New Dataset and Classification Experiments for Implicitly Abusive Language

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

Jana Kampfmeier
Digital Age Research Center (D!ARC)
Alpen-Adria-Universität Klagenfurt
AT-9020 Klagenfurt, Austria
janaka@edu.aau.at

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

Josef Ruppenhofer
Center of Advanced Technology for
Assisted Learning and Predictive Analytics
FernUniversität in Hagen
D-58097 Hagen, Germany
josef.ruppenhofer@fernuni-hagen.de

Abstract

We address the task of identifying euphemistic abuse (e.g. *You inspire me to fall asleep*) paraphrasing simple explicitly abusive utterances (e.g. *You are boring*). For this task, we introduce a novel dataset that has been created via crowdsourcing. Special attention has been paid to the generation of appropriate negative (non-abusive) data. We report on classification experiments showing that classifiers trained on previous datasets are less capable of detecting such abuse. Best automatic results are obtained by a classifier that augments training data from our new dataset with automatically-generated GPT-3 *completions*. We also present a classifier that combines a few manually extracted features that exemplify the major linguistic phenomena constituting euphemistic abuse.

1 Introduction

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

(1) Stop editing this, you dumbass.
(2) you stupid fucking idiot, fucking kill yourself

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.

Due to the increasing amount of abusive language on the Web, NLP methods are required that are able to detect such content automatically.

Currently, only explicit abuse can be reliably detected (van Aken et al., 2018; Wiegand et al., 2019, 2021b). By **explicit abuse** we understand **abusive language that is conveyed by unambiguously abusive words** (e.g. *bimbo*, *scum*, *tosser*). The automatic detection of more implicit forms of abusive language (3)-(5) remains challenging (van Aken et al., 2018; Wiegand et al., 2019, 2021b).

(3) Did Stevie Wonder choose these models?
(4) She still thinks she matters.
(5) You look like the back end of a bus.

Despite some recent research efforts introducing new datasets for implicitly abusive language (Sap et al., 2020; ElSherief et al., 2021; Vidgen et al., 2021b; Hartvigsen et al., 2022), there are still some subtypes that have not been explored. This particularly concerns implicit abuse that does not target identity groups (e.g. *Jews*, *immigrants*, *women*, *black people* etc.). In this paper, we address such **a form of implicit abuse**. We address implicitly abusive sentences such as (6) that paraphrase some common explicitly abusive utterances (7). Such paraphrases are a form of euphemism (Felt and Riloff, 2020) since they are perceived as less harsh than their explicit counterparts. Therefore, we call them **euphemistic abuse**.

(6) euphemistic abuse:
 (a) I see that good hygiene does not sit well with you.
 (b) You make watching paint dry an activity to look forward to.
 (c) Your love of yourself is astounding.
 (d) No one would describe you as a hero.
 (e) I bet you slow down for car crashes.
(7) explicit abuse:
 (a) You stink.

[1] http://thelawdictionary.org

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 16280–16297
December 6-10, 2023 ©2023 Association for Computational Linguistics

(b) You are boring.
(c) You are arrogant.
(d) You are a coward.
(e) You are sadistic.

In their roadmap on implicit abuse, Wiegand et al. (2021b) identify euphemistic abuse as one subtype next to abuse towards identity groups, dehumanization, call for action, multimodal abuse and comparisons. The authors establish euphemistic abuse as a rare phenomenon, with a frequency similar to dehumanization or comparisons, which have been examined in previous work (Mendelsohn et al., 2021; Wiegand et al., 2021a).

We focus on euphemistic sentences that are **perceived as abusive without any additional context** when being uttered in an everyday situation. The target of our sentences is a generic individual addressed by the second person pronoun. The task is a **binary (sentence-level) classification problem** in which abusive euphemisms are to be distinguished from similar but non-abusive sentences.

We introduce a novel dataset for this task that has been created via crowdsourcing. A special property of that dataset is that the negative instances, i.e. the non-abusive sentences, are *contrast sets* (Gardner et al., 2020; Sen et al., 2022). They represent sentences that are syntactically and semantically similar to the instances of euphemistic abuse. This makes them very difficult to distinguish from euphemistic abuse for automatic classifiers.

Apart from detailing the creation of our dataset, we report the performances of various automated classifiers. The most effective is an approach that augments the data by automatically generated completions by GPT-3. We also introduce a feature-based approach that exemplifies the major linguistic phenomena that constitute euphemistic abuse.

Our **contributions** are the following:

- We introduce a novel dataset for the task of detecting euphemistic abuse and describe the complex creation procedure.
- We demonstrate that euphemistic abuse cannot be detected effectively by classifiers trained on previous datasets.
- We report on the performance of various classification approaches including a classifier based on augmented text instances and a classifier combining high-level linguistic features.

All data and annotation guidelines created as part of this research are **publicly available**.[2]

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

2 Related Work

Much of the previous work on 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).

Quite recently, several new datasets for supervised classification have been introduced that focus on implicitly abusive language (Sap et al., 2020; ElSherief et al., 2021; Vidgen et al., 2021a,b; Hartvigsen et al., 2022; Wiegand et al., 2022). Wiegand et al. (2022) report that most of these datasets suffer from heavy biases. Moreover, all of these datasets focus on identity groups. To the best of our knowledge, the only exception is the dataset by Wiegand et al. (2021a). Like our dataset, it has been constructed via crowdsourcing. Yet it includes only abusive sentences in the form of *like*-comparisons (e.g. *Your hair looks like you have been electrocuted*). Our dataset of euphemistic abuse is much more heterogeneous in terms of sentence structure.

Another work that is also closely related to ours is the collection of *microaggressions* introduced in Breitfeller et al. (2019) which are *subtle, often veiled, manifestations* of abusive language. However, that dataset exclusively comprises abusive instances. That is, there are no negative (non-abusive) data, which is a prerequisite for supervised classification. Moreover, in that dataset, each example typically represents a larger discourse in which a situation involving abusive language is described. On average, instances are 7 times longer than the instances in our dataset (Table 1). Unlike our dataset of euphemistic abuse, the abusive utterances are also very context-dependent.

Our instances of euphemistic abuse can be considered as paraphrases of explicit abuse. Paraphrases in abusive language detection have been studied for the task of translating abusive sentences to non-abusive (*civil*) paraphrases (Laugier et al., 2021; Logacheva et al., 2022). In this paper, we address paraphrases that remain abusive.

3 Data

Our dataset was produced via crowdsourcing. As a platform we used Prolific academic.[3] Using existing datasets for our task was not an option as instances of euphemistic abuse are too rare in them and have also a very limited lexical variability. Fol-

[3]www.prolific.co

lowing previous work on creating a dataset for a rare subtype of implicit abuse (Wiegand et al., 2021a), we therefore asked crowdworkers to **invent** instances of euphemistic abuse.

All crowdworkers admitted to any of our tasks had to be native speakers of English with some basic academic education (at least undergraduate level) without dyslexia and an overall approval rate of 95% or higher. We had to enforce these criteria since, without them, we had far too many crowdworkers that failed to follow our annotation instructions (e.g. they did not understand the concept of implicitly abusive language; they provided many ungrammatical sentences, etc.). In general, our aim was to recruit a sample of crowdworkers that represents a wide spectrum of the English-speaking society. This was achieved by dividing the creation of the dataset into many smaller tasks, thus allowing many different crowdworkers to participate in the creation process. We did not specifically sample from a narrow range of age or a specific subset of countries (other than those in which English is spoken as a first language).

Figure 1 illustrates the order of the individual tasks that we set up. Each of these tasks is described in the following. We repeatedly ran through this pipeline until no more significantly new sentences were obtained.

Cue Phrases. The cue phrases represent the input given to the crowdworkers. They were selected by the authors and consist of explicitly abusive sentences (e.g. *You are ugly*) that the crowdworkers are to paraphrase. Many common abusive words, such as slurs (e.g. *cunt*, *tosser*), do not have a sufficiently specific semantics that can be paraphrased. Therefore, they are unsuitable for being used in cue phrases. We consulted the lexicon of abusive words from Wiegand et al. (2018) that contains a large number of common nouns and adjectives with a specific semantics (e.g. *wannabe*, *boring*, *stupid*). We aimed for a list with a wide semantic spectrum of abuse. Therefore, we avoided including multiple similar words. We identified 97 abusive words that complied with the above criteria and used them for our cue phrases. The cue phrases were selected manually since we did not see any possibility to identify abusive words with a sufficiently specific semantics automatically (i.e. one of the criteria these words have to fulfil).

Some of our cue phrases may only be mildly abusive. However, each paraphrase would be validated at a later stage as being abusive.

❶ **Generating Euphemistic Abuse.** Given an abusive cue phrase, crowdworkers were asked to invent some paraphrase that was not to contain any abusive word. It should comprise a single sentence that is unambiguously abusive. No further context or world knowledge should be required to understand its meaning or perceive its abusive intent.

❷ **Filtering of Euphemistic Abuse.** Despite our above criteria for recruiting crowdworkers the sentences produced by them still required some post-processing. Apart from removing sentences that contained abusive words or that required special world knowledge (e.g. sentences referring to a movie that is not generally known), we checked whether the paraphrase matched the given cue phrase. We also removed sentences that are *like*-comparisons since there is already a dataset specializing in these sentences (Wiegand et al., 2021a). The filtering was done semi-automatically. The presence of abusive words could be checked automatically. However, checking whether a response paraphrases a given cue phrase accurately and checking whether no further context or world knowledge is required to understand the paraphrase had to be accomplished manually.

In order to avoid duplicates and near-duplicates being repeatedly added to our dataset, we semi-automatically removed those new paraphrases with a high similarity to any of the sentences already included in our sentence pool with the help of Sentence-BERT (Reimers and Gurevych, 2019).

❸ **Validation of Euphemistic Abuse.** Each sentence that successfully passed the previous steps was validated by another 5 crowdworkers. To ensure an unbiased validation, these crowdworkers were not involved in the previous generation task. Neither were they given the original cue phrase for each sentence. The crowdworkers could label each sentence as either being an *abusive sentence*, some form of *criticism*, some *other sentence* or *not being proper English*. *Criticism* was included in order to prevent crowdworkers from labeling instances of heavy criticism as abuse. This was only considered as an auxiliary category which we did not consider for automatic classification.

Only those sentences were retained that were not flagged as improper English by any crowdworker. A sentence was considered abusive if this was the judgement of the majority of the 5 crowdworkers. All remaining sentences were kept in the dataset

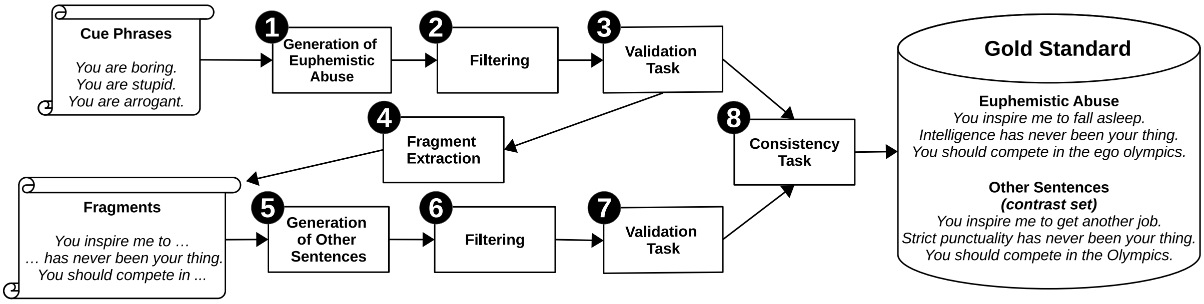

Figure 1: Illustration of how the novel dataset is created.

as non-abusive sentences. In order to have an even more varied and larger set of non-abusive instances, however, we applied the following steps:

❹ **Fragment Extraction.** The non-abusive sentences we created should be as close to the language of euphemistic abuse as possible. Essentially, we were aiming to produce *contrast sets* (Gardner et al., 2020) since they are known to result in training data for effectively learning class distinctions.

The first step of producing such data was to isolate **fragments** from the existing euphemistic abusive sentences. These fragments (9), which were manually derived, should preserve as much of the syntax and semantics of the original sentence as possible (8). However, they should also allow crowdworkers to use them to produce a non-abusive sentence that does not sound convoluted.

(8) *euphemistic abuse:* You'll let anyone between your legs.
(9) *fragment:* You'll let anyone ...

The fragments were created by starting from the complete sentence and removing constituents until the remaining fragment was no longer considered abusive. The constituents that were removed first were those that hardly had an impact on the syntactic and semantic structure of the sentence. Thus, the resulting fragments only differ by degrees from the original sentence in terms of syntax and semantics.

❺ **Other Sentences.** We asked an additional set of crowdworkers to form complete sentences using the fragments we had derived from the euphemistic abusive sentences. The resulting sentences should be negative in polarity but not abusive. For instance, given the fragment in (9), crowdworkers should come up with a sentence such as (10). We focused on negative polarity as the resulting sentences would be semantically similar to the euphemistic abuse.

(10) You'll let anyone get away with their actions.

For some fragments, inventing negative but non-abusive sentences was impossible. This concerns sentences, such as (11), where the fragment itself (12) is biased towards a completion having a positive sentiment (13). For these cases, we also added another task in which we asked crowdworkers to produce a positive sentence.[4]

(11) I am glad we can only meet as often as we do.
(12) I am glad we can ...
(13) I am glad we can spend time together.

❻ **Filtering of Other Sentences.** Similar to step ❷, we applied filtering steps to remove unwanted sentences (e.g. sentences requiring world knowledge) from our pool of *other* sentences.

❼ **Validation of Other Sentences.** We had the *other* sentences validated using the same task description as for validating euphemistic abuse.

❽ **Consistency Task.** There were sentences in our dataset with a fairly similar meaning that had been assigned opposing class labels (possibly due to the fact that different crowdworkers had annotated them). For example, (14) was labeled as not abusive but (15) and (16) were.

(14) You have testosterone coming out your ears.
(15) Your testosterone is showing.
(16) You have all the social skills of a neanderthal.

The consistency task consists of two steps:

We first manually identified potentially inconsistent sets of sentences in our dataset, such as (14)-(16). These sets were identified in a semi-automatic fashion. We first built clusters of similar sentences automatically. Such clusters were generated with the help of Sentence-BERT. Then, we manually compiled sets of instances from those clusters that were indeed semantically similar and also had contradicting class labels. We could not

[4] Since they only serve as a back-off, their proportion is low (11% compared to the number of negative abusive sentences).

conduct this identification fully automatically as the sentences in the automatically created clusters often were not similar enough.

In the second step, we had these inconsistent sets validated by the crowdworkers. More specifically, crowdworkers were to score the entire group (either as *abusive sentence* or *other sentence*) and indicate when they considered some member of the group to deviate from that group label. The label of that particular sentence would then be updated in case it did not correspond with the original label. Thus, about 7% of the labels of the final dataset were changed.

The Final Dataset. In our final dataset, **each sentence is either labeled as *abusive* or *other* (i.e. non-abusive)**. Table 1 provides some statistics on our crowdsourcing experiments and the final dataset we used in our forthcoming experiments. More than 600 crowdworkers were recruited in order to ensure a sufficiently high lexical variability of the sentences generated. About 80% of the invented 10K sentences were removed as a result of the above filtering measures in order to produce a dataset with sufficient annotation quality. This particularly meant removing a large number of near-duplicates. Retaining them in our dataset would have meant that classifiers memorizing a few frequently occurring instances of euphemistic abuse would already yield a high classification performance. However, our aim is to measure the performance on a broad range of euphemistic abuse.

A random sample of 200 sentences from our final dataset was also annotated by one co-author. We compared these labels with the majority vote of the crowdworkers resulting in a substantial agreement (Landis and Koch, 1977) of Cohen's $\kappa = 0.72$.

For more details regarding the set-up of our crowdsourcing tasks, we refer you to Appendix §B. *The instructions for each task given to the crowdworkers is part of the github repository.*

4 Classifiers

In the following, **we always used RoBERTa** (Liu et al., 2019) **for supervised classification unless stated otherwise**. We considered this model since it is regarded as a robust transformer for text classification. We fine-tuned the pretrained model on the given training data using the FLAIR-framework (Akbik et al., 2019) with default settings. ***The appendix (§A) contains more details regarding the settings of all classifiers.***

crowdsourcing		
tasks (overall)	114	100.0%
generation tasks	59	51.8%
validation tasks	45	39.5%
consistency tasks	10	8.8%
crowdworkers	631	
invented sentences (overall)	10538	100.0%
invented euphemistic abusive sentences	3950	37.5%
invented negative (non-abusive) sentences	5918	56.2%
invented positive (non-abusive) sentences	670	6.4%
final dataset		
cue phrases	97	
sentences	1797	100.0%
euphemistic abusive sentences	640	35.6%
other (non-abusive) sentences	1157	64.4%
avg. sent. length of euph. abus. sentences	9.8 tokens	
avg. sent. length of other sentences	10.0 tokens	

Table 1: Statistics on crowdsourcing and final dataset.

For some approaches we had to **identify explicit abuse**. In order to detect this kind of abuse accurately, we fine-tuned RoBERTa on the dataset by Founta et al. (2018). This dataset was chosen since it comprises short text instances like our dataset and has been reported to contain a high degree of explicit abuse (Wiegand et al., 2019).

4.1 Cross-Dataset Classifiers

We now list some baselines that address tasks related to ours and that were trained on other datasets.

Definition-based Classifier. Dictionary definitions of explicitly abusive words (17)-(18) bear some resemblance to euphemistic abuse in that they also paraphrase abusive words. However, euphemistic abuse is much more varied in style. We want to show that a definition-based classifier does not suffice to detect euphemistic abuse.

As training data for euphemistic abuse, our definition-based classifier uses word definitions from online lexicons for the abusive words in the cue phrases. We obtain our definitions from Wiktionary[5] (17) and WordNet (18) (Miller et al., 1990). In order to collect the largest training set possible, we use the union of the entries of both resources.

(17) *coward*: a person who lacks courage (*Wiktionary*)
(18) *ugly*: displeasing to the senses (*WordNet*)

As negative training data, we take word definitions (from the above lexicons) of negative polar expressions that are not abusive. These expressions are obtained from the Subjectivity Lexicon (Wilson et al., 2005). Since we do not want to overfit the class distribution of the training data to our

[5] www.wiktionary.org

test set, we assume an equal distribution between definitions of abusive and non-abusive expressions.

Euphemism Classifier. All our abusive sentences are instances of euphemism (Felt and Riloff, 2020). However, not all euphemisms paraphrase abusive words (e.g. *make someone redundant* is a non-abusive euphemism of *lay someone off*). We train a classifier on the dataset from the Shared Task on Euphemism Detection (Lee et al., 2022) and predict an abusive sentence if the classifier detects a euphemism.

Previous Classifiers for Abusive Language Detection. We consider a classifier that is trained on the union of all available datasets for implicitly abusive language detection (**implicitUnion**), i.e. all those mentioned in §2, a classifier trained on a typical dataset for explicit abuse (we use Founta et al. (2018)), and the publicly available classifier **PerspectiveAPI**[6]. We also consider the most recent and most robust publicly available transformer for implicitly abusive language detection from Hartvigsen et al. (2022), i.e. HateBERT finetuned on **ToxiGen**.

4.2 Within-Dataset Classifiers

In the following, we describe classifiers that are all trained and tested on our new dataset.

4.2.1 Feature-based Classifier

Our feature-based approach is a logistic regression trained on our new dataset that combines a small set of high-level (not necessarily mutually exclusive) binary features. We want to examine in how far the concepts represented by these features predict euphemistic abuse. Since the features are difficult to produce automatically, we extract them manually.

The annotation was done by a co-author. For each feature, we measured a substantial agreement to another co-author on a sample of 200 sentences.[7] *The github repository contains all annotation guidelines.*

Negated Antonyms of Abusive Words. Our first feature extracts negated antonyms (e.g. *not beautiful*) of abusive words (e.g. *ugly*). Examples are (19)-(21). Negated antonyms are perceived almost as equivalent to the respective abusive word. In our manual annotation, we use a broad notion of negation: We do not only consider negation words (e.g. *not* (19) or *nothing* (20)) but also *shifters* (e.g. *lack* (21)), i.e. content words that, similar to negation words, can affect the polarity of a phrase (Wilson et al., 2005; Polanyi and Zaenen, 2006; Schulder et al., 2017).

(19) You are <u>not beautiful</u>. (→ *ugly*)
(20) There is <u>nothing of interest</u> in your life. (→ *boring*)
(21) You <u>lack humility</u>. (→ *pompous*)

Opposing Sentiment. We annotate sentences in which there is an obvious pairing of opposing sentiments (22)-(24). Such a pairing is a stylistic device meant to trigger a particular reaction on the part of the reader. Such utterances are typically contradictions and thus, though they appear to be positive in sentiment (since the positive polar expression is usually the more salient expression), they are often meant in a derogatory way. This phenomenon, which has received some considerable attention in sarcasm detection (Riloff et al., 2013; Hee et al., 2018), is also often perceived as abusive.

(22) You are excellent$^+$ at breaking$^-$ things.
(23) You must love$^+$ having people hate$^-$ you.
(24) You are unique$^+$ in your ability to disappoint$^-$.

Taboo Topics. Abusive language often employs words associated with taboo topics, such as specific bodily organs, physical and mental abnormity, to express offensiveness. Allen and Burridge (2006) define taboo as a proscription of behavior that affects everyday life. They also provide a list of semantic fields, such as death (25), sex (26) or bodily functions (27) which form the basis of the manual extraction of our taboo feature.

(25) I'd prefer you were in a <u>grave</u>.
(26) You would fit well in a <u>brothel</u>.
(27) Your <u>smell</u> greeted me five minutes before you arrived.

Extremes. We annotate sentences that can be considered instances of extreme or absolute language. Such language bears some similarity to abusive language. Often, these two linguistic phenomena even coincide. There can be many linguistic realizations of extreme or absolute language. E.g. it may be conveyed by the usage of superlatives (28), generalizations (29) or hyperbole (30).

(28) You are truly the <u>best</u>$_{superlative}$ at doing nothing.
(29) You are not very good <u>at anything</u>$_{generalization}$.
(30) If you get any thinner, <u>you'll be transparent</u>$_{hyperbole}$.

Lexicalization. Some of the euphemistic abusive sentences in our dataset are lexicalizations (31)-(33). That is, they contain derogatory idioms that one could also potentially find in a

[6] https://perspectiveapi.com/

[7] The agreement scores (Cohen's κ) are in Appendix §B.3.

dictionary, such as *thefreedictionary.com*. With this category, we want to measure the degree to which euphemistic abuse is actually lexicalized. We anticipate a low proportion of lexicalization in our dataset since, otherwise, the task of detecting euphemistic abuse could be easily reduced to a dictionary-based approach.

(31) You are not the sharpest tool in the box.
(32) You are a thorn in my side.
(33) You don't have a backbone.

Unusual Properties. This feature is used for any utterance which includes some unusual property, behaviour or situation:

(34) Your main hobby must be letting life pass you by.

By *unusual*, we mean the following:

- The addressee is attributed unusual properties or displays some unusual behaviour. This could be strange hobbies (e.g. *staring at an empty wall for hours*), preferences (e.g. *fancy other people failing*) or beliefs (e.g. *believing in fairy tales*).
- The addressee causes unusual situations or events (e.g. *causing others to scream and run away*) or unusual behaviour on the part of the speaker (35).
- The unusual properties may also be conveyed by the usage of non-standard language, i.e. unusual imagery (36) or some other creative wording (37).

(35) You will make me want to do very nasty things.
(36) Your heart made an iceberg look warm.
(37) You are the leader of Boredville.

Our concept of *unusual properties* should cover all cases in which the abused target is meant to be alienated from the reader. Thus, this feature actually captures diverse instantiations of *othering*, i.e. a means of stigmatizing the target as not fitting in within the norms of a social group, which has been observed to coincide with abusive language very often (Burnap and Williams, 2016).

4.2.2 Classifiers using GPT-3 Completions

We devise classifiers that operate on the dataset being augmented by additional text. We use each sentence of our dataset as a *prompt* for GPT-3 (Brown et al., 2020) and generate a larger number of text *completions* for that prompt. As an individual completion we let GPT-3 generate a sentence of similar size as the input sentence. GPT-3 completions have been shown to be an effective way of augmenting datasets for abusive language detection (Wullach et al., 2021; Hartvigsen et al., 2022). We expect that among the generated completions of euphemistic abusive sentences, e.g. (38), there will be instances of abuse bearing some semantic relation to the original sentence (they may even be paraphrases of that sentence) that are easier to detect automatically, e.g. since they are instances of explicit abuse (39).

(38) euphem. abuse: You are a master of doom and gloom.
(39) completion: You are a self-absorbed $\text{whiner}_{explicit}$.

(Appendix §A.4 describes in detail how prompts in GPT-3 are constructed along illustrations.) GPT-3 completions are used in **2 different variations**:

inclusive. This classifier is trained on the original sentences where a fixed number of GPT-3 completions are concatenated to each original sentence. (The test data are equally augmented.) In our initial experiments we observed that larger numbers of completions have a more beneficial impact on classification than smaller numbers. In our evaluation, we used 100 completions for each sentence since this number would exhaust the maximum text length of standard transformers (i.e. 512 tokens).

exclusive. This classifier does not use our novel dataset as training data.[8] Instead, we run a classifier trained for explicit abuse on the set of 100 completions for each original sentence. If the classifier predicts abuse within the set of completions[9], the original sentence is considered abusive.

4.2.3 Seq2Seq using T5

Instead of framing our task as a binary *classification* problem, we can also consider it a **sequence-to-sequence** (seq2seq) task in which new sentences are *generated* from the sentences of our dataset. We want to learn directly how implicitly abusive remarks [input] can be paraphrased by their explicit counterparts [output]. This process reflects how humans often realize the abusive nature of some implicitly abusive remark. We examine whether there is some advantage to this modelling approach.

[8]This variant is actually a *cross-dataset classifier*. We introduced it in this subsection due to its similarity to the *inclusive*-variant, which is a *within-dataset classifier*.

[9]We actually set the threshold to 3 since it might happen that a single completion out of the 100 completions is misclassified as abusive by chance. The threshold of 3 was motivated by Manning and Schütze. (1999) who proposed it for removing noise in frequency-based metrics. It may well be that a different threshold produces even better results. However, we refrained from trying this since we did not want to overfit this classifier.

During training, for the euphemistic abusive input sentences (6), the output to be learned are their explicit counterparts (7), which are represented by the original cue phrases. For the non-abusive input sentences of our dataset, we consider as output the input sentence as there is no hidden explicit abuse to be uncovered. Since the output of the resulting seq2seq-classifier can be any arbitrary sentence, its evaluation is less straightforward. In order to compare this output with those of the previous approaches (§4.1-§4.2.2) we consider a generated output sentence of an abusive input sentence as correct if it contains some explicit abuse. A generated output sentence of a non-abusive input sentence is correct if no explicit abuse is contained. As a seq2seq-model, we use T5 (Raffel et al., 2020).

5 Evaluation

As evaluation measures, we use macro-average precision, recall and F1-score. For **within-dataset** classification, we carry out a 5-fold cross-validation. Unless stated otherwise, we arrange the folds in such a way that the euphemistic abusive sentences for the same cue phrase are restricted to one fold. Thus, the test data will always contain euphemistic abuse for **unseen cue phrases**. The alternative in which sentences are assigned to the folds at **random** is less strict since euphemistic abuse may originate from cue phrases observed in training.

For all classifiers that we trained on a non-deterministic model, e.g. RoBERTa, we report the average over 5 training runs (+ standard deviation).

As an upper bound, we also tested a **human classifier**. We randomly sampled the judgment of one individual annotator from the crowdsourced gold-standard annotation. This individual judgement may notably differ from the gold standard label which is the majority label of 5 annotators.

5.1 Comparison of Different Classifiers

Table 2 shows the performance of the different classifiers on our new dataset. All classifiers that have been trained on other existing datasets perform very poorly. Surprisingly, the performance of the union of all known datasets for implicitly abusive language is no exception. We ascribe this to biases reported on these datasets (Wiegand et al., 2022) and the fact that they mostly focus on identity groups whereas abuse on our dataset is unrelated to membership in identity groups. By far the best cross-dataset classifier is the one that classifies the GPT-3 completions of the test instances according to explicit abuse. Evidently, GPT-3 is a good way to uncover the abusive nature in euphemisms.

classifier	**Prec**	**Rec**	**F1** (*std*)
majority-class classifier	32.2	50.0	39.2
cross-dataset evaluation			
definition-based classifier*	52.8	51.7	52.2 (±1.6)
(general) euphemism detection*	52.9	52.8	52.6 (±0.7)
implicitUnion*	61.7	53.4	56.6 (±1.7)
explicit abuse (Founta et al., 2018)*	67.4	50.8	57.9 (±1.7)
ToxiGen (Hartvigsen et al., 2022)	61.5	62.5	62.0
PerspectiveAPI	78.2	52.2	62.6
GPT-3::exclusive*	68.5	68.3	68.4 (±0.3)
within-dataset evaluation (5-fold cross-validation)			
standard text classification*	68.1	61.3	64.5 (±2.4)
seq2seq using T5	70.9	59.2	64.5 (±1.2)
standard text classific. (*random*)*	68.4	64.5	66.2 (±1.9)
GPT-3::inclusive*	**73.7**	**69.1**	**71.3** (±0.6)
feature-based classifier	74.6	76.6	75.6
human classifier	79.3	77.3	78.3

Table 2: Evaluation of different classifiers on our novel dataset (*: RoBERTa has been used as classifier).

	anton.	**lex.**	**taboo**	**senti.**	**extreme**	**unusual**	**all**
F1	60.3	61.0	63.0	65.4	68.5	72.3	**75.6**

Table 3: Performance of the different features (§4.2.1).

As far as the within-dataset classifiers are concerned, a **standard text classification** approach, i.e. training on the sentences as they are, is only 2% points above the generic cross-dataset classifier of PerspectiveAPI. Even if we relax our strict setting and evaluate on random folds so that the test data also comprise euphemistic abuse of cue phrases observed in the training data, the increase in performance is still rather limited. Seq2seq is on a par with the standard classification approach. Potentially, it is more sophisticated since it generates new sentences rather than simply producing a binary class label. However, our manual inspection of those sentence outputs revealed that only about 28% of the generated sentences preserved the meaning of their input sentence. (*Some illustrations can be found in Table 6 of the appendix.*)

The only notable increase in performance for within-dataset evaluation is obtained by training on sentences augmented by completions (*GPT-3::inclusive*). This approach also outperforms the variant in which we classify the completions but do not train on our dataset (*GPT-3::exclusive*). The feature-based classifier notably outperforms all within-dataset classifiers. It is not much below the human baseline. Table 3 shows the performance

	manual	automatic	(Founta et al., 2018)
F1	64.5 (±2.4)	82.2 (±1.1)	95.7 (±0.2)

Table 4: Evaluation of standard text classification (RoBERTa) on different negative data.

	lex.	taboo	anton.	senti.	extreme	unusual
coverage	70.7	59.2	51.0	45.1	55.6	52.2
prevalence	4.4	11.3	20.2	22.8	59.2	72.7

Table 5: Coverage of *GPT-3::inclusive* on each linguistic feature (*prevalence* is the percentage of that feature on the set of all *actual* euphemistic abusive sentences).

of the individual features of that feature set. The strongest feature is the one that detects *unusual properties* but it is still less effective than the combination of features.

5.2 Impact of Negative (Non-Abusive) Data

We measure the impact of our expensive method to produce negative instances **manually** (§3) by comparing it against 2 alternative methods.

The first, like our proposed method, employs newly generated sentences. However, unlike our proposed method, we do not produce them manually but have them generated **automatically** by having the fragments derived from euphemistic abuse (step ❹ in §3) completed to full sentences by GPT-3. The fragments are used as prompts for GPT-3. They had been created in such a way that they do not contain any abusive content from their source sentences (cf. (8) & (9)). Since GPT-3 very rarely generates an abusive completion given a non-abusive prompt, i.e. in our case a fragment, we are likely to obtain non-abusive sentences.

Our second alternative simply uses non-abusive instances from an existing dataset for abusive language detection. We used the dataset from **Founta et al. (2018)** since it is known to be fairly unbiased (Wiegand et al., 2019). As a dataset comprising tweets it also consists of instances that are of similar length as our manually produced instances. We removed Twitter-specific tokens (e.g. hashtags) as they might create spurious correlations in the resulting dataset (Ramponi and Tonelli, 2022).

For both alternatives the same amount of negative data was sampled as in our original dataset. We replaced the negative data both in the training and test set. Table 4 shows the F-scores of the standard text classifiers trained on the 3 different datasets. There is a huge increase in performance, the highest for *Founta*. The lower the score, the more negative instances that are difficult to distinguish from the abusive instances the dataset contains. Such instances are obviously included in our manually compiled set of negative instances that comprise our proposed dataset. On the other hand, the performance of *Founta* suggests that its negative instances are (almost) trivial to distinguish from euphemistic abuse. This is the result of using inappropriately sampled data, a phenomenon that has often been reported in abusive language detection (Arango et al., 2019). **This experiment suggests that training a classifier on our dataset represents the hardest setting possible**. A classifier performing well on such data should also perform well on other data.

6 Discussion

The performance of our best automated classifier (i.e. *GPT-3::inclusive*) is still notably below that of the manual but very predictive feature-based classifier. Since the latter classifier is an explainable classifier we measured the overlap of our best automated classifier with the feature-based classifier in order to find out which linguistic aspects that classifier manages to detect to what specific degree. Table 5 presents the results. It shows the coverage of *GPT-3::inclusive* on the subset of euphemistic abuse for which a specific linguistic feature was extracted. In order to get an idea of the significance of each individual feature we also list its percentage among the set of euphemistic abuse. The table shows that *GPT-3::inclusive* manages to cover euphemistic abuse from all feature categories to some extent. The coverage on *lexicalization* is notably higher than on the remaining features. However, this is also the least frequent feature category.

7 Conclusion

We presented the task of identifying euphemistic abuse. We introduced a novel dataset that has been created via crowdsourcing and also includes suitable negative data. Our classification experiments revealed that classifiers trained on previous datasets are insufficient. Best results are obtained by a classifier that augments training data from our new dataset with automatically-generated completions. As an explainable upper bound, we also presented a classifier that combines a set of manually extracted features that helps us better understand what linguistic concepts are involved in that task.

Limitations

Our dataset comprises single sentences that are **not context-dependent**. Superficially, this seems to be a simple task setting. However, our experiments revealed that already euphemistic abuse that is not context-dependent represents a notable challenge for state-of-the-art classification approaches. Besides, existing datasets are not a good source from which to train a classifier for this task.

The dataset we introduce in this paper is created via crowdsourcing rather than by extracting text from existing datasets or the Web. Therefore, it may be criticized for lacking **authenticity**. However, as with many other subtypes of implicit abuse, existing datasets do not provide a sufficiently representative set of instances for such long-tail phenomena. For other subtypes of implicitly abusive language, the same strategy was pursued (Vidgen et al., 2021b; Wiegand et al., 2021a). It was also applied to plagiarism detection (Potthast et al., 2010) and deception detection (Ott et al., 2011). As a matter of fact, the whole idea of contrast sets (Gardner et al., 2020) is also based on the principle of having annotators create sentences with specific linguistic content. Contrast sets are increasingly becoming a standard procedure for producing hard datasets for specific NLP tasks since they allow us to determine whether classification models sufficiently learned the actual task rather than memorizing some spurious correlations.

Our dataset **only addresses one subtype of abusive language**. Therefore, classifiers trained on our new data will only be capable to detect this type of abusive language rather than abusive language, in general. Thus, we do not follow a *one-size-fits-all approach* but a *divide-and-conquer approach*, instead. This is in line with Wiegand et al. (2021b) who claim that this is the only reasonable approach to complex phenomena such as the detection of implicitly abusive language. The fact that our dataset is a focused dataset that specializes in only one subtype of implicitly abusive language may also explain that its size with less than 2K instances is comparatively small. Datasets specializing in other subtypes of implicitly abusive language are of similar size if not smaller. As a matter of fact, the dataset for implicitly abusive comparisons (Wiegand et al., 2021a) comprises only 1K instances.

Quite recently, Röttger et al. (2021) introduced HATECHECK, a test suite that should be used for identifying weaknesses in classification models for abusive language detection. Unfortunately, we cannot apply this test suite on classifiers trained on our new dataset. The functionalities of HATECHECK, i.e. the groups of instances that are designed to check the capability of a classification model with regard to a specific linguistic phenomenon, address abusive language targeting identity groups or mentions of explicitly abusive words. The subset of implicitly abusive language that we address in our new dataset does not target identity groups. Neither does our dataset contain explicitly abusive words.

Our feature-based approach exclusively comprises **manually extracted features**. Therefore, the performance of the resulting classifier trained on these features does not reflect the state of the art of the automatic feature extraction. We also tried to produce an automatic version of that classifier. However, either there were no plausible training data for these features or the available training data only resulted in poor classification performance. For instance, for the feature of detecting opposing sentiment, we fine-tuned RoBERTa on the dataset from SemEval 2018 Task 3: Irony detection in English tweets (Hee et al., 2018).[10] The resulting classifier overfit heavily to Twitter-specific data artifacts of the SemEval dataset and therefore generalized poorly to our dataset. Our feature-based approach should be regarded as an *upper-bound baseline*. We included it to show what major linguistic phenomena are involved in the formulation of euphemistic abuse and what performance can be reached by a supervised classifier that perfectly identifies these phenomena.

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 at the time of annotation (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 whereas the number of sentences to be invented was restricted to 30 sentences. Furthermore, we allowed each crowdworker to take part in one single

[10]The category *Verbal irony by means of a polarity contrast* corresponds exactly to our concept of opposing sentiment.

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.

Having crowdworkers invent abusive language may look ethically debatable. However, we think that this is justifiable in this particular context since we do not think that there is an alternative method that would yield a dataset with a comparable quality. Moreover, the type of abusive language that our work focuses on does not target any specific individual or any specific identity group. Therefore, the creation of our dataset does not immediately harm anyone. By producing an additional dataset for abusive language detection, our intention is not to encourage people to use this form of language. On the contrary, our new dataset should result in an improved basis for building effective classifiers to detect abusive language thus enabling better control of such unwanted language on the Web. In principle, creating morally disputable content as part of research is not unheard of. 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.

8 Acknowledgements

The authors were partially supported by the Austrian Science Fund (FWF): P 35467-G. The authors would like to thank Sybille Sornig for contributing to the manual annotation of this research.

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.

Keith Allen and Kate Burridge. 2006. *Forbidden Words: Taboo and the Censoring of Language*. Cambridge University Press.

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.

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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.

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.

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.

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

Christian Felt and Ellen Riloff. 2020. Recognizing Euphemisms and Dysphemisms Using Sentiment Analysis. In *Proceedings of the Workshop on Figurative Language Processing (Fig-Lang)*, pages 136–145, Online.

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 Behaviour. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, Stanford, CA, USA.

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.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3309–3326, Dublin, Ireland.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 Task 3: Irony Detection in English Tweets. In *Proceedings of The International Workshop on Semantic Evaluation (SemEval)*, pages 39–50, New Orleans, LA, USA.

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

Léo Laugier, John Pavlopoulos, Jeffrey Sorensen, and Lucas Dixon. 2021. Civil Rephrases Of Toxic Texts With Self-Supervised Transformers. In *Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL),*, pages 1442–1461, Online.

Patrick Lee, Anna Feldman, and Jing Peng. 2022. A Report on the Euphemisms Detection Shared Task. In *Proceedings of the Workshop on Figurative Language Processing (FLP)*, pages 184–190, Abu Dhabi, United Arab Emirates (Hybrid).

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*.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. ParaDetox: Detoxification with Parallel Data. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 6804–6818, Dublin, Ireland.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press.

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.

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

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.

Livia Polanyi and Annie Zaenen. 2006. Contextual Valence Shifters. In *Computing Attitude and Affect in Text: Theory and Applications*, pages 1–10. Springer-Verlag.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Alan Ramponi and Sara Tonelli. 2022. Features or Spurious Artifacts? Data-centric Baselines for Fair and Robust Hate Speech Detection. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 3027–3040, Seattle, WA, USA.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. 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.

Paul Röttger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet B. Pierrehumbert. 2021. 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.

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.

Marc Schulder, Michael Wiegand, Josef Ruppenhofer, and Benjamin Roth. 2017. Towards Bootstrapping a Polarity Shifter Lexicon using Linguistic Features. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, pages 624–633, Taipei, Taiwan.

Indira Sen, Mattia Samory, Claudia Wagner, and Isabelle Augenstein. 2022. Counterfactually Augmented Data and Unintened Bias: The Case of Sexism and Hate Speech Detection. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 4677–4695, Seattle, WA, USA.

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 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, Elisabeth Eder, and Josef Ruppenhofer. 2022. Identifying Implicitly Abusive Remarks about Identity Groups using a Linguistically Informed Approach. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 5600–5612, Seattle, WA, 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.

Tomer Wullach, Amir Adler, and Einat Minkov. 2021. Fight Fire with Fire: Fine-tuning Hate Detectors using Large Samples of Generated Hate Speech. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4699–4705, Punta Cana, Dominican Republic.

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.

Appendix Overview

This appendix provides more detailed information regarding certain aspects of our research for which there was not sufficient space in the main paper. **This is optional, to a large extent illustrative material which is not necessary to consult in order to understand the main paper.**

We focus on the following aspects:

- details on the configuration of the statistical models we used in our experiments (§A)
- details on the annotation experiments, particularly crowdsourcing and the manual annotation of high-level features (§B)

A Hyperparameters of Statistical Models

For all statistical models we used in this research we **refrained from heavy tuning of hyperparameters**. This is due to the fact that several 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 tools with frequently used (default) settings of hyperparameters, we hope to produce models that are overall more robust across different domains (i.e. different datasets) than models fine-tuned on the wrong domain. Thus, we follow the strategy that was proposed for the large-scale cross-dataset evaluation reported in Wiegand et al. (2022).

A.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). We estimate a total computational budget of 100 GPU hours.

A.2 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 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, displaying a high fluctuation in classification performance between the 5 different runs. We also got a considerable amount of runs that just produced a majority-class classifier. Our observation was that the more different training and test data were (and since we include a considerable number of cross-dataset classifiers in this paper, this accounts for many experiments), the more majority-class classifiers we obtained.

The runs of RoBERTa-large that did not result in a majority-class 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-dataset evaluations were conducted (Wiegand et al., 2018, 2022).

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.

A.3 T5

We used T5 (Raffel et al., 2020) for learning translations from our euphemistic abusive sentences to explicit cue phrases. The code we used was based on the summarization code example provided by huggingface.[11] We used this transformer with its default hyperparameter settings in order to have an unbiased setting:

- learning rate=5e-05
- number of epochs=3
- batch size=4
- number of beams=4
- dropout rate=0.1

The specific model variant we used was T5-large. We also experimented with smaller variants but performance was notably inferior. We could not test even larger models, e.g. T5-3b, as those models had resource requirements that exceeded our computing infrastructure (§A.1).

Table 6 illustrates some sentences generated on our test data. Our manual evaluation revealed that only about 28% of the sentences are semantically equivalent to the input sentence (upper half of Table 6). We observed that very often T5 generates some sentences that bear some semantic relation to the input sentence yet they are not equivalent (lower half of Table 6).

[11] https://github.com/huggingface/transformers/blob/main/examples/pytorch/summarization/run_summarization.py

input sentence	sentences generated by T5
semantically plausible sentences	
You would not be considered pretty.	You are ugly.
I see no value in you.	You are useless.
You start more fights than the ringing of a bell.	You are belligerent.
Your main hobby must be letting life pass you by.	You are lazy.
You are brilliant at putting your own needs first.	You are selfish.
You inspire no one.	You are boring.
Reason is foreign to you.	You are stupid.
Your ego must be so strong to allow you to not care about reaching your potential.	You are arrogant.
implausible sentences	
I'd prefer you were in a grave.	You are a murderer.
I knew a stick that had more fat on than you.	You are fat.
You are old fashioned.	You are senile.
You are skilled at taking what is not yours.	You are a hobo.
You like to sing like no one is listening.	You are a show off.
You make me want to hide behind my sofa.	You are pathetic.
I bet you slow down for car crashes.	You are reckless.
You don't care about appearances at all.	You are ugly.

Table 6: Illustration of sentences generated by a T5-model trained (i.e. fine-tuned) on our novel dataset of euphemistic abuse.

A.4 GPT-3

In this research, we use GPT-3 for two different purposes, (1) for augmenting each individual sentence of our dataset by completions (§4.2.2), and (2) for obtaining automatically generated negative data instances for our fragments we extracted from the (original) euphemistic abusive sentences (§5.2). Although GPT-3 supports outputs for both tasks, the model has to be run in two different modes. For obtaining text completions, we use the **Completion-mode**, while for the second task we run the **Insertion-mode**. This second mode was necessary since our fragments are designed in such a way that the part that is to be completed can be at any position of a sentence, i.e. not just the end of a sentence, as exemplified by (40).

(40) Your ability to ... has always inspired me.

For the Insertion-mode, we specify at what part of a given sentence additional text material is to be inserted by the command *[insert]*:

(41) Your ability to [insert] has always inspired me.

Unlike the Insertion-mode, in which we specify what output we want to obtain explicitly by the command in the brackets (i.e. *insert* in (41)), some discussion must be devoted to the creation of the sentence generation in the Completion-mode. Here, no such explicit command can be used. For this mode, we exactly follow the syntax of the **prompt**[12] as proposed by Hartvigsen et al. (2022), that is, the input sentence is preceded by a hyphen and is followed by a newline and another hyphen. Thus, for generating a text completion of (42), we use (43) as a prompt.

(42) You inspire me to fall asleep.

(43) `-You inspire me to fall asleep.\n-`

This format suggests that we want to generate some enumeration. More specifically, we want to generate a list of similar sentences (Hartvigsen et al., 2022). The sentence to follow the prompt should mainly paraphrase the prompt. Our aim is to generate 100 paraphrases for each original sentence (§4.2.2). Therefore, we generate 100 individual sentence completions for each original sentence.

From GPT-3, we also considered the most recent model that was available during the time span our experiments were carried out (i.e. spring to winter 2022): **text-davinci-002**. In order to avoid overfitting, we mainly used the default settings of the hyperparameters.

- temperature=0.7
- top p=1
- frequency penalty=0.0
- presence penalty=0.0

The only parameter which deviates from the default settings is the maximum number of tokens (*max token*). The default is 256. The sentences we want to paraphrase or complete are fairly short since our dataset consists of sentences with an average size of about 10 tokens per sentence (Table 1). Using the default setting of 256 tokens would have resulted in generating pieces of text that are far too

[12] *Prompt* refers to the input or query provided to GPT-3 to generate a specific output or response.

long and would not resemble our original dataset. (Please notice that the sentences we wanted to generate should be of the same text length as their prompts, i.e. the original sentences of our dataset.)

It may come as a surprise that we did not set the maximum number of tokens to 10 (which would exactly correspond to the average sentence length in our dataset). This was motivated by the fact that our dataset consists of individual sentences. However, in GPT-3, one cannot specify the sentence length. Instead, with *max token* one specifies the overall length of the text to be generated. The resulting text may comprise an arbitrary number of sentences. Moreover, in the Completion-mode, the last sentence is also often incomplete (as the completion of the last sentence would exceed the given maximum number of tokens). By setting *max token* to 10, we ended up with too many instances being incomplete sentences. Our final setting was *max token=20* for the Completion-mode and *max token=15* for the Insertion-mode. The token size for the Insertion-mode was lower than for the Completion-mode as the size reflects the length of the additional text material to be generated. In the Insertion-mode, parts of the resulting sentence are already given (in the form of the fragment represented as the prompt). Therefore, to obtain a sentence length similar to that of the sentences to be generated by the Completion-mode, fewer tokens need to be generated in the Insertion-mode.

Table 7 illustrates for some euphemistic abusive sentences of our dataset a set of text completions that GPT-3 generates as they were used for augmenting our training data (§4.2.2). (*The github repository contains the entire set of those text completions we used in our experiments.*) Though GPT-3 is a generic language model, that is, it has not been trained for generating explicit abuse from implicitly abusive remarks, it actually often produces some suitable explicitly abusive sentence completions for the given input sentence. From that one can conclude that GPT-3 often associates an implicitly abusive sentence to some general toxic context.

A.5 Logistic Regression

For our feature-based classifier (§4.2.1) 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**.

A.6 PerspectiveAPI

In our evaluation, we also included *PerspectiveAPI*[13] as one baseline. This tool runs on unrestricted text and, from the publicly available classifiers, it is currently considered the state of the art for the general detection of abusive language (Röttger et al., 2021). The tool predicts several subtypes of abusive language. We examined how well the two categories *toxicity* and *insults* correlate with euphemistic abuse on our dataset. That is, we compared the predictions of the tool for both categories against the euphemistic abuse in our gold standard. We considered the above two categories as they bear the greatest similarity concept-wise to euphemistic abuse. *Toxicity* performed slightly better than *insults* which meant that we used the former in the evaluation in our main paper (i.e. Table 2).

B Details on Annotation

B.1 General Remarks

All guidelines for the different annotation and sentence generation tasks are contained in the github repository to this work.

An issue that may raise some concern is that we decided in favor of a low threshold in terms of what counts as abusive language. More specifically, we asked the crowdworkers who eventually validated the sentences that were generated to think of how sensitive people may react to each given sentence. We think that such a low threshold is appropriate for this particular task since most euphemistic abuse in our dataset looks superficially harmless. There is not a single explicitly abusive sentence, i.e. a sentence in which the abusive nature is evoked by an unambiguously abusive word. Therefore, we had to heighten the crowdworkers' awareness of this subtle form of abusive language. Our final dataset may contain the odd sentence being labeled as abusive which may be considered acceptable in specific contexts, such as legal proceedings, e.g. *You have a sexual interest in children*. However, we consider these situations very special. With our dataset we have everyday situations in mind, such as online communication in social media. In this context, such utterances are not acceptable. They are only meant to offend or discredit the person addressed.

[13] www.perspectiveapi.com

input sentence (i.e. prompt)	completions generated by GPT-3
You are a master of doom and gloom.	You are a very dark person. You are a <u>self-absorbed whiner</u>. You are not a people person. You are a bit of a <u>wimp</u>. You never miss an opportunity to be pessimistic.
You'll let anyone between your legs.	You're nothing but a <u>lowlife whore</u>. You're just a big <u>whorish slut</u>. You have no friends, only men you screw. You've been with so many guys, it's hard to keep track. You'll do anything for money.
Your main hobby must be watching television.	You're a huge fan of the show Teen Mom. You must be a man of few interests. You are a <u>couch potato</u>. Television is a great source of entertainment. You must have a television in your bedroom.
Your testosterone is showing.	You're thinking with your little head. You might want to check your zipper. You're a <u>dick</u>. That's pretty <u>macho</u>. Don't show your testosterone.
I can't imagine anyone coming to you when they need a shoulder to cry on.	You were just never there for anyone. You're so <u>selfish</u> and only think about yourself. I know that you don't even know the meaning of love. You're such a <u>bitch</u>. You're really insensitive.

Table 7: Illustration of completions generated by the (generic) GPT-3 model for some euphemistic abusive sentences of the novel dataset. (*In the completions, explicitly abusive expressions are <u>underlined</u>.*)

B.2 Measures to Avoid Biases

In the following, we describe the measures we took in our crowdsourcing experiments in order to avoid any obvious bias.

Employing Many Crowdworkers. We do not want our dataset to be dominated by the responses by some few specific crowdworkers. Therefore, we split all annotation tasks into subtasks of bins of 100-200 instances. Tasks in which crowdworkers were to invent sentences were split into subtasks of about 30 sentences. Each crowdworker was only admitted to one subtask. As a result, many different crowdworkers (more than 600) contributed to our dataset but each individual crowdworker only contributed to a small share of the overall dataset.

Open Task Description. Our task description for creating euphemistic abuse was kept very brief. We did not suggest any strategy how to invent such sentences or any typical constructions to use. Neither did we recommend any auxiliary tool to use. Thus, crowdworkers were fairly free in devising their sentences and we did not impose any restriction towards a particular subtype of euphemistic abuse (apart from avoiding *like*-comparisons since they are already sufficiently covered in the dataset by Wiegand et al. (2021a)). We hope that, as a result, the resulting set of instances of euphemistic abuse is fairly representative of this subtype of abusive language.

Randomization and Balancing. In each subtask, we tried to present individual examples in a randomized order. This was particularly important for the *consistency task* in which groups of similar sentences were presented and the crowdworkers had to identify the sentence being the odd one out. The sentences within this group were always randomized. However, next to randomization we also had to balance the groups in terms of classes. Most inconsistent groups we found were groups that were generally considered abusive but one single instance had previously been labeled as non-abusive. Thus, we felt that we had to add distractors so that the crowdworkers would not recognize that underlying pattern of the groups to be annotated. More specifically, we added further groups with sentences being generally considered non-abusive, and occasionally one sentence was an obvious abusive sentence (which should function as the odd one out). Of course, these *distracting* groups were also *randomly* interspersed in the list of groups to be annotated by the crowdworkers.

feature	Cohen's κ
opposing sentiment	0.89
negated antonyms of abusive words	0.74
unusual properties	0.67
extreme	0.66
lexicalization	0.65
taboo topics	0.61

Table 8: Inter-annotator agreement for the manual annotation of high-level features.

B.3 Details on the Manual Annotation of High-Level Features

For each of the high-level features, we measured the inter-annotator agreement on a random sample of 200 sentences between two co-authors. One of these co-authors annotated the features for the remaining sentences. Table 8 lists the agreement between these two annotators for each feature. It is highest for *opposing sentiment* and lowest for *taboo topics*, though that agreement can still be considered substantial (Landis and Koch, 1977).