Originally published in: Toutanova, Kristina/Rumshisky, Anna/Zettlemoyer, Luke/Hakkani-Tur, Dilek/Beltagy, Iz/Bethard, Steven/Cotterell, Ryan/Chakraborty, Tanmoy/Zhou, Yichao (Eds.): Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – Stroudsburg: Association for Computational Linguistics, 2021. Pp. 576-587.

Implicitly Abusive Language – What does it actually look like and why are we not getting there?

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

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

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

Abstract

Abusive language detection is an emerging field in natural language processing which has received a large amount of attention recently. Still the success of automatic detection is limited. Particularly, the detection of implicitly abusive language, i.e. abusive language that is not conveyed by abusive words (e.g. *dumbass* or *scum*), is not working well. In this position paper, we explain why existing datasets make learning implicit abuse difficult and what needs to be changed in the design of such datasets. Arguing for a divide-and-conquer strategy, we present a list of subtypes of implicitly abusive language and formulate research tasks and questions for future research.

1 Introduction

Abusive or offensive language is commonly defined as hurtful, derogatory or obscene utterances made by one person to another person or group of persons.[1] Examples are (1)-(3). In the literature, closely related terms include *hate speech* (Waseem and Hovy, 2016) or *cyberbullying* (Zhong et al., 2016). While there may be nuanced differences in meaning, they are all compatible with the general definition above.[2]

(1) stop editing this, you dumbass.
(2) Just want to slap the stupid out of these bimbos!!!
(3) 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.

[1] http://thelawdictionary.org/abusive-language

[2] The examples in this work are included to illustrate the severity of abusive language. They are taken from actual web data and in no way reflect the opinion of the authors.

Though there has been much work on abusive language detection in general, comparatively little work has been focusing on **implicit** forms of abusive language (4)-(5) (Waseem et al., 2017). By *implicit* abuse we understand abusive language that is not conveyed by (unambiguously) abusive words (e.g. *dumbass*, *bimbo*, *scum*).

(4) I haven't had an intelligent conversation with a woman in my whole life.
(5) Why aren't there any Mexicans on Star Trek? Because they don't work in the future either.

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).

In this position paper, we want to shed more light on the nature of implicitly abusive language. We identify subtypes of implicit abuse that can be found in existing datasets and the literature. We also outline shortcomings that prevent implicitly abusive language from really being learned on its own terms. With this study, we hope to guide future research on implicitly abusive language.

Our **contributions** in this paper are:

- We present a list of subtypes of implicit abuse. This is accompanied by quantitative information from publicly available datasets.
- We derive research tasks and questions regarding those subtypes for future research.
- We detail properties of existing datasets that make them less suitable for training classifiers to detect implicit abuse.
- We propose key issues that need to be considered when building new datasets for implicitly abusive language.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 576–587
June 6–11, 2021.
Publikationsserver des Leibniz-Instituts für Deutsche Sprache
URN: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-104498

2 The Story So Far

By far the most prominent classification approaches applied to abusive language detection are supervised learning methods. Whereas initially, traditional learning algorithms, such as SVMs or logistic regression, were among the most popular methods for this task (Warner and Hirschberg, 2012; Burnap et al., 2015; Nobata et al., 2016), at present, best results are obtained by deep-learning methods, particularly transformers (Struß et al., 2019; Kumar et al., 2020; Zampieri et al., 2020). A more detailed summary of the methods explored can be found in Schmidt and Wiegand (2017) and Fortuna and Nunes (2018).

Unfortunately, so far there has been little error analysis of system output for abusive language detection. As a consequence, the community is fairly unaware of what types of errors are made and why.

The most notable exception is van Aken et al. (2018) who carry out experiments on the dataset of Google's Toxic Comment Classification Challenge[3] and the dataset by Davidson et al. (2017).

As prominent errors that a supervised classifier makes, van Aken et al. (2018) list *toxicity without swearwords*, *rhetorical questions* and *comparisons/metaphorical language*. All these phenomena can be subsumed by implicit abuse. Unfortunately, the study by van Aken et al. (2018) is only of limited help since one of two datasets considered, namely the dataset from the Toxic Comment Classification Challenge, contains a high degree of explicitly abusive language (Table 1). The other dataset, i.e. the dataset by Davidson et al. (2017), is not considered in our work, since it is not a dataset for the *detection* of abusive language but the *disambiguation* of potentially abusive words.[4]

Wiegand et al. (2019) find that supervised classifiers with a reasonable cross-domain performance are those that are trained on datasets with a high degree of explicit abuse. Classifiers trained on datasets with a high degree of implicit abuse perform poorly on other datasets, no matter whether one deals with implicit or explicit abuse. From that the authors conclude that classifiers are not effectively learning implicit abuse.

[3] www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/overview

[4] In other words, it deals with the question in which contexts a mention of a potentially abusive word (e.g. *fuck*) is really used in an abusive manner and what type of abuse is conveyed, i.e. *hate speech* or mere *profanity*.

dataset	publication	size	%expl.
Kumar	(Kumar et al., 2018)	15,000	32.7
SBFrames	(Sap et al., 2020)	45,318	37.6
Waseem	(Waseem and Hovy, 2016)	16,165	44.4
Warner	(Warner and Hirschberg, 2012)	3,438	51.3
OffensEval	(Zampieri et al., 2019)	13,240	54.0
Razavi	(Razavi et al., 2010)	1,525	64.7
Founta	(Founta et al., 2018)	59,357	75.9
Kaggle	(Wulczyn et al., 2017)	312,737	76.7

Table 1: Statistics of datasets (*size*: number of microposts; *expl.*: explicit abuse).

3 Datasets Considered in this Study

Recent years have seen a notable increase of datasets for abusive language detection. Since a survey would be beyond the scope of this section, we refer the reader to Poletto et al. (2020) and Vidgen and Derczynski (2020). However, implicit abuse is not covered in these publications.

Due to the limited space, we only focus on English datasets in this paper. We also just consider the common binary classification task of whether a micropost is abusive or not. Table 1 shows the proportion of explicit abuse on the different datasets. We compute these scores by checking each abusive micropost from a dataset for the presence of an abusive word according to the lexicon of abusive words from Wiegand et al. (2018). The complementary proportion to each score can be considered a proxy for the degree of implicit abuse (e.g. 67.3% for *Kumar*). However, such scores should just be considered an upper bound for implicit abuse since we will have missed explicitly abusive microposts. (Even the lexicon from Wiegand et al. (2018) is not exhaustive.) From the scores in Table 1, we can conclude that the datasets *Kumar*, *SBFrames*, *Waseem*, *Warner* and *OffensEval* have a fairly high proportion of implicit abuse, which is why we focus on these datasets in the remainder of this paper.

4 Different Subtypes of Implicit Abuse

For each dataset, we manually annotated a random sample of 500 implicitly abusive instances (according to our proxy described in §3) for their subtypes, i.e. 2,500 instances in total. The subtypes we used were either mentioned in previous work (van Aken et al., 2018) or frequently observed in the examined datasets. In the following, we describe these subtypes:

4.1 Stereotypes

By stereotypes, we understand a fixed, overgeneralized belief about a particular group or class of

people (Cardwell, 1999):

(6) Jews have undue influence.

Stereotypes are difficult to detect as there are quite a few stereotypes per identity group. Moreover, stereotypes need not be negative in tone:

(7) Jews are good at making money.

As a consequence, using sentiment analysis as a pre-filtering step by isolating only negative statements may miss a substantial fraction of stereotypical remarks. However, as a research task it may be a reasonable starting point since not every negative sentence focusing on some identity group conveys some (abusive) stereotype (e.g. (8)-(10)). A first research question could be how to detect stereotypical statements among negative statements.

(8) Gay people fight for the right to be accepted.
(9) Muslims groan under the recession.
(10) Jews mourn the loss of a member of their community.

We believe that specific linguistic properties may be indicative for automatic classification. For example, stereotypes are more likely to co-occur with habitual aspect (11) rather than non-habitual aspect (12) (Friedrich and Pinkal, 2015).

(11) Jews always support terror instability.
(12) Jews currently fear displaying their faith in public.

One should also examine whether generic phrases regarding identity groups (13) correlate with stereotypes (Reiter and Frank, 2010). Previous work already established that the definite article, which represents a subset of such generic phrases, is predictive for abusive language (Burnap and Williams, 2015; Palmer et al., 2017).

(13) The jew does not care about the humankind.

Further, the same stereotype can be expressed in different ways. For example, (14)-(17) convey the sexist stereotype that *women belong in the kitchen*.

(14) Men should drive and women should cook.
(15) This is how America should be. 5 women slaving over a hot stove.
(16) Get back in the kitchen.
(17) Women should all stay at home in an apron, chained to the oven!

We could also envisage the following research task for stereotype classification. Rather than classifying arbitrary statements on identity groups as stereotype or not, one could start with a set of existing stereotypes (e.g. *black people are criminal*) and learn to align all statements on identity groups crawled from a social-media site to these existing stereotypes. Such a setting may be simpler than the plain classification since the stereotypes are known in advance. This setting would even be in line with reality since the set of stereotypes regarding identity groups is not infinite.

To date, the only significant work analyzing the detection of stereotypes is Cryan et al. (2020) who compare a lexicon-based classifier with a supervised classifier for detecting gender stereotypes. The dataset used for supervised classification and the lexicon are created via crowdsourcing. No deeper linguistic analysis of the data is performed.

4.2 Perpetrators

(18)-(21) depict some identity group as perpetrators.

(18) Jews are **raping** kids.
(19) Muslims are **looting** old people's homes.
(20) Black people **steal** everything.
(21) Jews **scheme** on world domination daily.

By perpetrator, we understand a person who commits, an illegal, criminal, or evil act.[5] An individual being depicted as a perpetrator (22) may be just considered an accusation. However, if identity groups (i.e. *Jews*, *Muslims* etc.) are presented in this way, (18)-(21) this is perceived as abusive language, particularly if the victims are protected classes (e.g. *kids*, *old people*) as in (18) and (19).

(22) Our neighbour is **beating** his wife.

This form of implicit abuse is a proper subset of stereotypes (§4.1). However, we think that abuse conveyed by depicting someone as a perpetrator has some notably different properties than the other stereotypes. These properties justify a separate category. The actions that characterize perpetrators are often criminal offenses (e.g. raping, murdering, stealing) or are at least morally contemptible (e.g. adultery, lying, scheming). Thus, we consider them to be universal actions that can apply to different targets (i.e. identity groups). In contrast, the other stereotypes are target-specific and less universal. Switching identity groups does not necessarily preserve the abusiveness as shown in (23) and (24).

(23) Jews belong in the kitchen.
(24) Women are good at making money.

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

We assume that the depiction as a perpetrator is also largely tied to (fairly unambiguous) lexical units, i.e. a subset of action predicates (primarily verbs) being negative polar expressions. From a computational perspective, it should, therefore, be feasible to detect such cases reliably. The depiction of other stereotypes may be less tied to specific lexical items. Therefore, we believe the detection of those stereotypes to be more challenging.

4.3 Comparisons

Abusive comparisons are comparisons in which the vehicle (*you* in (25)) is compared to some offensive entity, action or state (*idiot* in (25)). Abusive comparisons need not be explicitly abusive (25) but can also be implicitly abusive (26)-(27).

(25) You talk like an **idiot**.
(26) You look like someone only a mother could love.
(27) You sing like a dying bird.

A research question that would need to be answered is whether detecting abusive comparisons is not (almost) identical to the detection of comparisons conveying a negative sentiment. Such classification of comparisons into positive (28), neutral (29) and negative comparisons (30) has already been addressed by Qadir et al. (2015).

(28) You look like a princess.
(29) You look like your brother.
(30) You look like a crackhead.

Another research question would be to examine whether abusive comparisons are not identical to (negative) comparisons using figurative language (i.e. similes as (31)). Intuitively, comparisons employing literal language should be less abusive (32).

(31) You look like the back end of a bus.
(32) You look like you have slept badly.

4.4 Dehumanization

By dehumanization, we commonly understand the act of perceiving or treating people as less than human (Haslam and Loughnan, 2016). While Haslam and Loughnan (2016) propose a fairly comprehensive set of different properties that characterize dehumanization, we focus on the most commonly accepted property of likening members of the target group to non-human entities (Haslam, 2006), such as machines, animals or diseases.

We observed two different realizations of dehumanization. On the one hand, the target is explicitly equated with non-human entities (33).

(33) Black people are monkeys.

On the other hand, a more difficult form of dehumanization involves metaphorical language in which the target is not explicitly equated to a non-human entity but their actions or properties are reminiscent of such entities as in (34)-(37).

(34) A **wild flock** of Jews is **grazing** outside a bagel store.
(35) Headscarfed muslims **waddle** around our streets all over.
(36) I **own** my wife and her money.
(37) How come bunches of gay people **mushroom out of the ground** these years?

Different classification approaches may be suitable for the detection of this second type of dehumanization. One may compile a corpus with mentions of animals, diseases etc. and learn the language (i.e. how non-human entities are depicted) by supervised learning. Alternatively, one might compile a lexicon that captures predicates describing actions of animals (e.g. *waddle*) or properties of objects/diseases (e.g. *mushroom out*) and then use this resource as a look-up.

Dehumanization in natural language processing has not yet been properly addressed. The only exception is the in-depth descriptive study by Mendelsohn et al. (2020) examining the dehumanizing connotation of the two words *homosexual* and *gay* in different temporally-indexed corpora.

4.5 Euphemistic Constructions

We observed several abusive remarks that were disguised as an euphemistic construction (38)-(40), typically some form of negation (39) & (40).

(38) You inspire my inner serial killer.
(39) Liberals are not very smart.
(40) I'm not excited about your existence.

If we translate these euphemisms into their unequivocal counterparts (41)-(43), the abusive nature of these statements becomes more obvious.

(41) I want to kill you.
(42) Liberals are retarded.
(43) I hate you.

With the exception of Felt and Riloff (2020), euphemisms have not been addressed in natural language processing so far.

As a research question, one would need to answer how abusive euphemisms can be detected and translated to their unequivocal counterpart.

4.6 Call for Action

Calls for action represent another type of implicitly abusive language. By that we understand that the author of a micropost asks that something, typically some form of punishment, needs to be done to the abused target (44)-(46). In particular violent actions may be shrouded in allusion. For example, (46) is an obscure way to demand that someone should be killed by electrocution.

(44) Thank you for your fortitude and perseverance. Please give McConnell a kick in the butt from some of us.
(45) @USER Liberals are so easy to figure out! Make America great again. Get rid of all liberal women.
(46) He should be given 5000 volts!

Given an appropriate dataset with sufficient occurrences, automatic methods should be able to detect this type of abuse, even in microposts, such as (46), although it is not an explicit call for killing someone. The presence of the modal verb *should* and the exclamation mark indicate the presence of an obligation or even command. In addition, the keyword *volt* in combination with a command may be a clear indicator that the author wants some violent action to take place. State-of-the-art classifiers should be able to learn such correlations.

The problem for studying this type of abusive language lies in its sparsity in the publicly available data. In many countries calling for violent actions is considered a crime. This deters many users from expressing such content on the web.

4.7 Multimodal Abuse

Most social-media platforms allow users to embed images or videos in their posts. In many cases, the abusive content of a micropost is hidden in the non-textual components or results as an interplay of text and image/video. One could also regard many of these abusive posts as instances of implicit abuse since many of them do not contain mentions of abusive words. Therefore, a comprehensive classifier to detect implicitly abusive microposts should also consist of a multimodal component that analyses image or video content and fuses this information with text analysis.

Indeed the community is aware of this form of abuse and there have been several attempts for multimodal analysis (Singh et al., 2017; Yang et al., 2019; Gomez et al., 2020). In our work, however, we do not address the aspect of multimodal abuse simply because many datasets only include the textual component of a micropost and the reconstruction of non-textual components of posts can only be reconstructed with greater effort or even not be obtained at all.

4.8 Phenomena Requiring World Knowledge and Inferences

Of the subtypes we present as implicit abuse, the final subtypes present the most difficult kind of abusive language. We subsume all those phenomena which can effectively only be detected with the help of inferencing and additional world knowledge. Given some appropriate training data and (linguistic) feature design, automatic methods should be able to detect any of the previous subtypes to a certain degree. All of the following types of implicit abuse, however, are unlikely to be established on the basis of such approaches.

- **Jokes.** Jokes as (47) can be severely abusive.
  (47) What's better than winning gold in the paralympics? Walking.

  The computational modeling of humor remains a challenging task (Mihalcea and Strapparava, 2006). We are not aware of any research on the detection of abusive humor.
- **Sarcasm.** Sarcasm is largely defined as the activity of saying [...] the opposite of what you mean (Macmillan, 2007). The way in which is spoken is intended to make someone else feel stupid or show them that you are angry. This explains the strong connection towards abusive language as in (48):
  (48) It's always fun watching sports with a woman in the room.

  Although the automatic detection of sarcasm has been investigated (Tsur et al., 2010; Riloff et al., 2013), the classification performance is still fairly limited.
- **Rhetorical questions.** Rhetorical questions are asked not to elicit information but to make a statement (Bhattasali et al., 2015). They have been examined on social-media texts (Ranganath et al., 2016; Oraby et al., 2017). Future work needs to address what makes a rhetorical question abusive:
  (49) Did Stevie Wonder choose these "models"?
- **Other implicit abuse.** Our final category comprises all *further forms* of implicit abuse that require world knowledge and inferencing:
  (50) She still thinks she matters.
  (51) I live in Ethiopia. Happy new year 1219!
  (52) These girls know skinny sausages are no fun.
  (53) Welcome to the Hotel Islamfornia. You may check out any time but you can never leave.

	datasets					
subtype	**Kumar**	**SBFrames**	**Waseem**	**Warner**	**OffensEval**	***average***
other implicit abuse	9.8	28.4	12.8	30.4	2.4	16.8
perpetrator	18.2	2.4	22.0	17.1	15.2	15.0
stereotype	13.4	2.0	12.2	20.0	14.2	12.4
joke	0.0	40.8	0.2	2.5	0.0	8.7
call for action	3.8	1.6	1.0	4.6	2.8	2.8
dehumanization	2.2	0.6	1.0	2.5	3.0	1.9
euphemistic construction	1.4	0.6	2.0	1.3	3.8	1.8
rhetorical question	1.2	1.6	1.6	2.1	0.6	1.4
comparison	0.6	0.0	1.4	0.0	0.0	0.4
sarcasm	1.0	0.0	0.2	0.0	0.6	0.4
unknown	37.0	11.0	37.8	10.8	23.0	23.9
explicit abuse (abus. word missing in Wiegand et al. (2018))	11.4	10.0	7.8	8.8	34.4	14.5

Table 2: Percentage of different subtypes of implicit abuse (including overlooked explicit abuse) within a dataset. The numbers are obtained by manually inspecting 500 implicit texts from each of the datasets.

4.9 Distribution of Subtypes

Table 2 shows the distribution of the subtypes of implicit abuse in the examined samples of the datasets. It also includes cases of **explicit abuse missed** by the lexicon from Wiegand et al. (2018) and **unknown** cases of implicit abuse which we could not assign to any of the previous subtypes. We were surprised by the high number of unknown cases, most notably in *Kumar*, *Waseem* and *OffensEval*. Some of posts are pretty short, such as *RIP*, *Why so* or *Ouch!* A large part of those unknown microposts requires the inclusion of further context information (e.g. multi-media attachments or links) in order to comprehend their abusive nature.

Most subtypes of implicit abuse are rare in all datasets, so none of them is an appropriate source for learning to detect these subtypes. Stereotypes, perpetrators and other implicit abuse are frequent in most datasets, however. *SBFrames* has a large amount of jokes. We assume that the sampling process to produce this dataset notably distorted the distribution of subtypes. We discuss this in §5.1.

Though we only found very few comparisons in the samples of abusive microposts (Table 2), comparisons seem a fairly natural form of abuse. Indeed, by manually inspecting the general dataset for comparisons by Qadir et al. (2015), we found that 2/3 of the person-targeted negative comparisons are abusive comparisons. About 75% of those abusive comparisons are implicitly abusive.

5 What should(n't) the datasets for implicit abuse look like?

Driven by the requirements of data-hungry deep-learning methods, the most common strategy for abusive language detection is to create a single dataset and train a classifier on it. That dataset should be as large as possible. Unfortunately, most of the datasets that are created in this way are of little use to *really* learn implicit abuse.

For one thing, large datasets for abusive language detection that are produced by random sampling usually have an overwhelming proportion of explicit abuse among abusive instances (Wiegand et al., 2019). Currently, we do not know whether this is due to the predominance of explicit abuse on most social-media platforms or the fact that human annotators more readily detect explicit abuse.

5.1 Biases

Datasets that contain a higher proportion of implicit abuse mostly suffer from biases caused by the sampling of the underlying raw data. (Typically, one samples microposts containing certain keywords or topics that may coincide with abusive language.) As Wiegand et al. (2019) showed, classifiers trained on these datasets may correctly detect implicitly abusive instances on unseen test instances of the same datasets. However, these correct classifications are not produced by grasping the concept of implicit abuse but by exploiting some artifacts contained in the dataset. Such artifacts can be frequently occurring words, such as *women* and *football*, that, due to the sampling process, coincidentally only occur in abusive microposts.

Although additional datasets containing larger amounts of implicit abuse have been released since Wiegand et al. (2019) published their findings, we found that these new datasets also suffer from biases. We outline these biases on the most recent dataset that displays a high degree of implicit abuse and that is also fairly large (Table 1): the dataset by Sap et al. (2020) (*SBFrames*). Of the recent datasets, it is also the only dataset to cover a significant amount of abusive instances targeting common

identity groups (e.g. *Jews*, *Muslims*).

In order to get a larger amount of microposts, existing datasets (e.g. Founta et al. (2018)) were merged into *SBFrames*. In addition, further raw data was added, such as posts from the white-supremacist platform *stormfront.org* or subReddits on abusive jokes from *reddit.com*. While these additional data undoubtedly yield more abusive content, it is problematic to merge data from different domains into one corpus. The resulting dataset is bound to be fairly heterogeneous in terms of style.

For example, most jokes from *reddit.com* follow a specific syntactic pattern: a question is asked to which some (short) abusive answer is given. This is illustrated by (47) and (48).

(47) What's worse than an angry black woman? Nothing.
(48) How do you pick up a Jewish girl? With a shovel.

Since the dataset does not explicitly state the origin of each micropost, we approximated the set of jokes by mining for the above syntactic pattern. More than 80% of the jokes are abusive. Due to the recurring syntactic pattern of jokes, classifiers trained on the corpus from Sap et al. (2020) will find it easy to detect abusive utterances. They basically have to look for a joke, i.e. a question followed by an answer. They do not really have to understand the joke or the concept of abuse. This observation is particularly significant to the detection of implicit abuse since more than 40% of the implicitly abusive microposts that we randomly sampled from the dataset were jokes (Table 2).

The above *reddit-joke-bias* is just one example of that corpus. We also noticed that identity groups (i.e. *Jews*, *Muslims*, *blacks* etc.), which comprise the typical targets of the dataset, also highly correlate with abuse (Table 3). For instance, almost all mentions of *Jew(s)* are abusive. This property makes the detection of such abusive instances considerably easier since a classifier can predict all cases including mentions of these words as abuse and reaches a high classification performance.

Simply removing the mentions of identity groups is insufficient. Microposts addressing those particular identity groups would still be restricted to the abusive microposts. Supervised classifiers are likely to infer that a micropost refers to some identity group although it has been been removed. For instance, one can easily infer that (49) is about *Jews* and (50) is about *Muslims* due to further contextual clues (*Hitler* & *gas* (49); *ISIS* & *Al-Qaeda* (50)).

(49) I'm pretty sure Hitler just said "I wanna glass of juice" not I wanna gas the <IDENTITY_GROUP>.
(50) Being a <IDENTITY_GROUP> I have a confusion choosing my career. Either to go with ISIS or Al-Qaeda?

Moreover, we have to assume further biases in the dataset from Sap et al. (2020): The proportion of abuse across the different sources from which this dataset is created seems to vary considerably: Abusive utterances in Founta et al. (2018) (this is one source of the dataset) are rare (14%) while the majority of posts from the white-supremacy site *stormfront.org* (another source of the dataset) should be abusive. This is so since the major topic of this platform (i.e. *white supremacy*) is racist. Since these texts also vary much in style across the different sources (the former are tweets, while the latter are longer posts with fully grammatical sentences), a classifier that learns to detect the style of the different sources will already have a good prior as to whether a particular post is abusive.

5.2 Divide and Conquer

We argue that by creating one dataset to cover all phenomena of abusive language, the creators of those datasets lose sight of appropriate *negative data*. By negative data, we mean those instances that are not abusive and contrast the abusive instances so that a classifier can learn a good distinction between abusive and non-abusive instances. By using inappropriate negative data, biases as those described in §5.1 will notably distort classification performance. If datasets are created for individual subtypes of implicit abuse (§4.1-§4.8) we obtain a less heterogeneous set of abusive instances for which it is easier to produce suitable negative instances. In order to classify unrestricted text, it would simply take a final meta-classifier that collects the predictions of all the specialized classifiers for specific subtypes of abuse.

5.3 More training data does NOT necessarily mean better training data

As we outlined in §5.1, increasing the size of data by merging different corpora is highly problematic. Supervised classifiers may simply produce higher classification scores as a result of further biases introduced by the merging process.

Thinking about negative data is important. If there are certain artifacts that coincide with the abusive instances due to the sampling process (i.e. they are not representative of abusive language), then one can neutralize that bias by enforcing it

identity group	woman	lesbian	gay	black	muslim	jew
% abusive	67.3	71.7	75.2	87.2	87.8	93.8

Table 3: Abusive posts with identity group.

to also occur in the negative data. For supervised classifiers, this artifact will then be ignored as it will occur in all classes equally.

For example, the mentions of identity groups (Jews, Muslims, women, gay people etc.) are mostly limited to abusive instances (Table 3). A less biased dataset would enforce mentions of identity groups in the negative data. Although the resulting overall dataset may be smaller as a result of selecting specific negative data, the overall quality of the training data should rise. In general, the NLP community is increasingly aware of such biased constructions in datasets and measures, as we propose, are an approved means to produce datasets to evaluate classifiers under more realistic conditions (McCoy et al., 2019).

Another problem of randomly sampling data is that due to the fact that the frequency distribution of a language vocabulary is generally a power law distribution (Zipf, 1965), instances will always be dominated by a few frequently occurring words. Supervised classifiers may achieve high classification performance by just focusing on these particular words. However, a dataset would be much harder if we tried to represent words more equally.

For example, if we were to produce a dataset for learning to detect identity groups depicted as perpetrators (§4.2), the best way would be to sample microposts with mentions of co-occurrences of an identity group and some negative polar expression (e.g. *Muslims rape*, *Muslims criticize*). In order to build a dataset that captures the long tail of rare constructions, we would need to ensure that we do not only include the frequently occurring negative polar expressions (e.g. *kill*, *murder*, *rape*) but also the infrequent ones (e.g. *calumniate*, *concoct*, *racketeer*). As a consequence, a dataset with 10k microposts that focuses on the frequent polar expressions may be less suitable for training a classifier on than a dataset that comprises 1k microposts but includes a wide set of polar expressions with each expression only occurring a few times.

Our call for smaller datasets that do not contain similar non-informative instances but a sample of the task that allows for sharper decision boundaries echoes ideas from the field from active learning (Settles, 2012) and the recent proposal for NLP evaluation in terms of contrast sets (Gardner et al., 2020).

5.4 Classification Below the Micropost-Level

Previous research considered *entire* microposts as instances from which to learn abusive language. However, there may be good reason to focus on smaller meaningful units, such as sentences or even clauses. This view is also shared by parts of the community. SemEval 2021 includes a shared task that addresses the detection of abusive text spans within a micropost.[6] In the following, we describe how such classification schemes would facilitate learning implicit abuse.

Given that social-media platforms commonly used for obtaining natural language data, such as Twitter, increasingly ban abusive language on their sites[7], the amount of data available in which abusive language is actually *used* is decreasing.[8] However, there are still many *mentions* of abuse available, such as *reported* cases (Chiril et al., 2020), including implicit abuse (51)-(52).

(51) @USER exposes the hypocrisy of claims that [Muslims want to suppress free speech]$_{abusive\ clause}$.

(52) The Texas GOP thinks that [gay people need a cure]$_{abusive\ clause}$.

For example, we randomly sampled 50 tweets from Twitter containing the abusive clause *homosexuality is unnatural*. After manual inspection we found that 76% of the tweets just reported this claim and the author clearly opposed that view.

Sometimes, the presence of emojis (53) or interjections (54) also suggests that the author of the tweet does not share the stated proposition.

(53) [Black people are aliens]$_{abusive\ clause}$ now 😂😂😂

(54) **Wow**, [Jews control everything]$_{abusive\ clause}$, **cool lol**

Given the above observations, we suspect that there are many abusive clauses that are only available as embedded abuse (51)-(54). In order to use them as training data for genuine abuse (such clauses may occur as genuine abuse, i.e. abuse that is not embedded, in unseen test data), we think it would suffice to isolate the actual abusive clauses and train on them instead of the entire microposts.

[6] https://sites.google.com/view/toxicspans

[7] https://techcrunch.com/2020/03/05/twitter-bans-hate-speech-around-age-disability-and-in-the-wake-of-the-coronavirus-outbreak-disease/

[8] Alternative social-media platforms which are known to contain a higher proportion of abusive language, such as *gab.com*, are considerably more difficult to process, as technical support equivalent to *Twitter.API* is typically not available.

Recent research on the helpfulness of context may also support our view to restrict the context for training data. In an in-depth study, Pavlopoulos et al. (2020) found that increasing the context for abusive language detection by considering microposts neighbouring the post to be classified actually harms classification performance. Microposts, such as tweets from Twitter, themselves can already be fairly long (up to 280 characters) representing a paragraph of sentences. Future research should investigate whether the non-abusive sentences of a longer abusive micropost already negatively affect learning abusive language.

Apart from that, an abusive micropost often contains more than one predictive clue. For such microposts, a supervised classifier may not need to detect all of these clues. Typically, the classifier is more effective in spotting the easier clues, which, in the case of abusive language detection, are (explicitly) abusive words. (55) is a micropost that includes both explicit abuse (i.e. the word *sneaky*) and implicit abuse (i.e. an abusive clause expressing some anti-Semitic stereotype). If we want to effectively learn the more difficult implicit clues, it may be useful to focus only on the implicitly abusive clauses by removing the explicit clues from microposts that also include implicit abuse.

(55) **Sneaky** [Jews are controlling the world through their banking]$_{abusive\ clause}$.

6 The Role of Machine Learning

Despite the continuing success of machine learning in many areas of NLP, particularly fairly generic methods, we should be careful in considering this the magic bullet for every problem including the detection of implicitly abusive language.

Already in some subtasks of (explicitly) abusive language detection, machine learning has not produced the anticipated results. For example, supervised learning still produces fairly poor classification performance on the cross-domain detection of abusive language, with lexicon-based approaches performing much stronger (Wiegand et al., 2018). Further, statistical debiasing methods for abusive language detection have also been reported to yield very limited success (Zhou et al., 2021). The authors of that research argue that spending more efforts in ensuring a high quality of the datasets during their creation is more worthwhile than applying sophisticated machine learning.

We anticipate that there are also some subtasks in the realm of implicit abuse that may not be solved with the help of supervised learning approaches. One such example may be the task of detecting novel or unknown stereotypes. If we compare the two stereotypes (56) and (57), we find that these sentences differ in meaning, sentiment and also in terms of syntactic structure.

(56) Asian children are intelligent.
(57) All Asian people lie.

If we train a classifier on (56) it is unlikely to identify (57) as an instance of the same category due to the lack of similar features. As a consequence, learning-based approaches are unlikely to succeed in this task.

Although generic supervised methods may always represent a good baseline, the community should also be open that other more linguistically informed approaches can be more effective for particular subtasks in the detection of implicitly abusive language. Riloff et al. (2013) demonstrated that mining for a particular linguistic construction is an effective means to recognize a specific type of sarcasm. We envisage that similar approaches may be effective for the detection of implicit abuse.

Due to the susceptibility of supervised learning to overfitting, we also recommend an experimental set-up in which a cross-domain evaluation is included in order to check whether the resulting classifiers generalize beyond the training data.

7 Conclusion

There are different subtypes of implicit abuse. Some of them are frequent in available datasets (e.g. jokes or stereotypes) while others are sparse (e.g. dehumanization or euphemisms). As far as frequent subtypes of implicit abuse (e.g. stereotypes and perpetrators) are concerned, unsuitable sampling causes biases that prevent classifiers from really learning these phenomena. Simply adding instances by merging datasets does not solve the problem. It may introduce further detrimental biases. Overall, our analysis supports the claim that the currently available datasets are not really suitable for effectively learning implicit abuse. We strongly argue for new datasets that focus on particular subtypes of implicit abuse. This will also facilitate thinking about appropriate negative data. Larger datasets are not necessarily the best datasets to train a classifier on, especially if they are dominated by frequently observed words. Finally, it may also make sense to learn on smaller units, such as clauses, rather than on entire microposts.

8 Ethical Considerations

This paper contains real-life examples of abusive language taken from actual web data. We are aware of the fact that some readers may feel offended by these examples, particularly since many of them address entire identity groups (e.g. Muslims, Jews etc.). We chose those examples deliberately in order to demonstrate that despite not being instances of explicit abuse, implicit abuse can still be extremely severe. Consequently, the automatic detection of implicit abuse should be considered equally pressing as the detection of explicit abuse.

The examples used in this paper in no way reflect the opinion of the authors. All mentions of specific user names were anonymized in order to comply with privacy principles.

Our work is critical of the design of existing datasets for abusive language detection. We would like to clarify that we do not generally challenge the usefulness of these datasets per se. Our criticism only relates to using these datasets for learning implicit abuse.

References

Shohini Bhattasali, Jeremy Cytryn, Elana Feldman, and Joonsuk Park. 2015. Automatic Identification of Rhetorical Questions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 743–749, Beijing, China.

Pete Burnap, Omer F. Rana, Nick Avis, Matthew Williams, William Housley, Adam Edwards, Jeffrey Morgan, and Luke Sloan. 2015. Detecting tension in online communities with computational Twitter analysis. *Technological Forecasting*, 95:96–108.

Pete Burnap and Matthew L. Williams. 2015. Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decition Making. *Policy & Internet*, 7(2):223–242.

Mike Cardwell. 1999. *Dictionary of Psychology*. Fitzroy Dearborn.

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.

Jenna Cryan, Shiliang Tang, Xinyi Zhang, Miriam Metzger, Haitao Zheng, and Ben Y. Zhao. 2020. Detecting Gender Stereotypes: Lexicon vs. Supervised Learning Methods. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pages 1–11, Honolulu, HI, USA.

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

Christian Felt and Ellen Riloff. 2020. Recognizing Euphemisms and Dysphemisms Using Sentiment Analysis. In *Proceedings of the Workshop on Figurative Language Processing*, 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.

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.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring Hate Speech Detection in Multimodal Publications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1470–1478.

Nick Haslam. 2006. Dehumanization: An integrative review. *Personality and social psychology review*, 10:252–264.

Nick Haslam and Steve Loughnan. 2016. Recent research on dehumanization. *Current Opinion in Psychology*, 11:25–29.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings of*

the Workshop on Trolling, Aggression and Cyberbullying (TRAC), pages 1–11, Santa Fe, NM, USA.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2020. Evaluating Aggression Identification in Social Media. In *Proceedings of the Workshop on Trolling, Aggression and Cyberbullying*, pages 1–5.

E. D. Macmillan. 2007. *Macmillan English Dictionary*. Michael Rundell.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3428–3448, Florence, Italy.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A Framework for the Computational Linguistic Analysis of Dehumanization. *Frontiers in Artificial Intelligence*.

Rada Mihalcea and Carlo Strapparava. 2006. Learning to laugh (automatically): Computational models for humor recognition. *Computational Intelligence*, 22(2):126–142.

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

Shereen Oraby, Vrindavan Harrison, Amita Misra, Ellen Riloff, and Marily Walker. 2017. Are you serious?: Rhetorical Questions and Sarcasm in Social Media Dialog. In *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue*, pages 310–319, Saarbrücken, Germany.

Alexis Palmer, Melissa Robinson, and Kristy K. Phillips. 2017. Illegal is not a Noun: Linguistic Form for Detection of Pejorative Nominalizations. In *Proceedings of the Workshop on Abusive Language Online (ALW)*, pages 91–100, Vancouver, BC, Canada.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxity Detection: Does Context Realler Matter? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4296–4305, Online.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*.

Ashequl Qadir, Ellen Riloff, and Marilyn A. Walker. 2015. Learning to Recognize Affective Polarity in Similes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 190–200, Lisbon, Portugal.

Suhas Ranganath, Xia Hu, Jiliang Tang, Suhang Wang, and Huan Liu. 2016. Identifying Rhetorical Questions in Social Media. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, Cologne, Germany.

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

Nils Reiter and Anette Frank. 2010. Identifying Generic Noun Phrases. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages Uppsala, Sweden, 40–49.

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.

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.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the EACL-Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 1–10, Valencia, Spain.

Burr Settles. 2012. Active Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114.

Vivek K. Singh, Souvick Ghosh, and Christin Jose. 2017. Toward Multimodal Cyberbullying Detection. In *Proceedings of the ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI)*, pages 2090–2099, Denver, CO, USA.

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.

Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. ICWSM - A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, Washington, DC, 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*. To appear.

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

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, 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.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 1391–1399, Perth, Australia.

Fan Yang, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore, and Goran Predovic. 2019. Exploring Deep Multimodal Fusion of Text and Photo for Hate Speech Classification. In *Proceedings of the Workshop on Abusive Language Online (ALW)*, pages 11–18, Florence, Italy.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 1415–1420, Minneapolis, MN, USA.

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*.

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.

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)*, Online.

George K. Zipf. 1965. *The Psycho-Biology of Language*. MIT Press.