


DEFVERIFY: Do Hate Speech Models Reflect Their Dataset’s Definition?

Urja Khurana , Eric Nalisnick , Antske Fokkens 

 Computational Linguistics and Text Mining Lab, Vrije Universiteit Amsterdam

 Department of Computer Science, Johns Hopkins University

u.khurana@vu.nl, nalisnick@jhu.edu, antske.fokkens@vu.nl

Abstract

Warning: Due to the nature of the topic, this paper contains offensive content. When building a predictive model, it is often difficult to ensure that application-specific requirements are encoded by the model that will eventually be deployed. Consider researchers working on hate speech detection. They will have an idea of what is considered hate speech, but building a model that reflects their view accurately requires preserving those ideals throughout the workflow of data set construction and model training. Complications such as sampling bias, annotation bias, and model misspecification almost always arise, possibly resulting in a gap between the application specification and the model’s actual behavior upon deployment. To address this issue for hate speech detection, we propose **DEFVERIFY**: a 3-step procedure that (i) encodes a user-specified definition of hate speech, (ii) quantifies to what extent the model reflects the intended definition, and (iii) tries to identify the point of failure in the workflow. We use **DEFVERIFY** to find gaps between definition and model behavior when applied to six popular hate speech benchmark datasets.

1 Introduction

Hate speech is a prevalent problem on social media but tackling it is not straightforward for numerous reasons. What constitutes hate speech varies by country and individual (Al Kuwatly et al., 2020; Schmidt and Wiegand, 2017). There are many ways to address hate speech in a detection task, causing the definition of hate speech to change based on what the context demands. A law-based hate speech detection model in Belgium, e.g., may consider “language” as a protected group identity while other countries might not (Khurana et al., 2022). Researchers may decide to include stereotypes or, alternatively, restrict their definition to slurs. The groups that are considered targets can also differ.

The type of hate speech that needs to be addressed affects the choice of data to train the detection models on. For instance, a dataset that only focuses on racism and sexism would not be suitable for an application that aims to capture hateful language against people with a disability. An important first step is thus making a *definition* of hate speech for the project at hand, both for dataset creators and downstream users. When *creating* a dataset, a creator needs to verify if the dataset is ultimately constructed according to the intended specifications for hate speech. For a user, *finding* a dataset that aligns with the type of hate speech they want to address is important. Ideally, the dataset’s hate speech definition¹ would serve as an effective proxy for assessing whether the dataset is suitable to the project. However, the definition does not necessarily translate into model behavior. The potentially noisy process of creating datasets (Ross et al., 2016; Madukwe et al., 2020; Fortuna et al., 2020; Vidgen and Derczynski, 2020) can result in the dataset not covering and/or a model not learning the correct aspects for classification (see Figure 1). This has severe consequences when deploying such models in the real world, potentially inviting unexpected behavior when the model has to generalize to new data or domain. Carefully analyzing all steps of this process to ensure proper generalization is not always feasible. *How can one verify that a constructed dataset adheres to the intended task type of hate speech detection?*

To investigate if a hate speech model behaves according to their dataset’s definition, we propose a procedure to verify this: DEFVERIFY. Our proposed methodology consists of three core steps when applied to a new hate speech dataset: (1) identifying which hate speech aspects, fueled by Hate Speech Criteria (Khurana et al., 2022),

¹For brevity, we will refer to this as *dataset’s definition* throughout this paper.

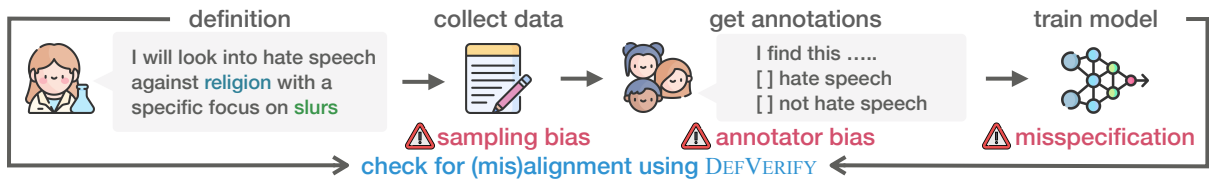


Figure 1: Potential failure points when creating a hate speech dataset. All biases can accumulate in the model.

models trained on it should capture, (2) investigating to what extent it does by looking at HateCheck (Röttger et al., 2021), an English diagnostic evaluation set to uncover the strengths and weakness of a system, and cross-dataset performance, and (3) conducting an early-stage analysis of where the model fails based on the evaluation. To facilitate these steps, we build on HateCheck by (a) matching different aspects of definitions to each instance and (b) adding test cases that should be deemed offensive (or neutral) but not *hate speech*. To demonstrate the utility of our approach, we apply it to *six* different English hate speech datasets and examine to what extent these datasets follow their dataset’s definition. We demonstrate gaps between dataset definitions and model capabilities for the datasets. Due to the task’s subjective nature, the verification and evaluation of model capabilities are still open problems. Drifting away from a one-approach-fits-all approach, we are the first to investigate responsible model behavior through the lens of alignment between a dataset’s hate speech definition and model capability. Our approach gives quick insights into what models can capture. The idea behind the approach can also be applied to other tasks.

2 Related Work

Variations in hate speech definitions. The complexity of *hate speech* is that there are various valid beliefs regarding what constitutes hate speech and what not (Röttger et al., 2022b). Moreover, several aspects influence what hate speech is (Schmidt and Wiegand, 2017), and it also depends on the social context (Sap et al., 2019). Analyses of definitions on existing datasets confirm this complexity. Fortuna et al. (2020) examine different hate speech datasets and find that even when datasets use very generic categories, their divergent definitions or inconsistent annotation can lead to distinct classifier performance. Datasets can also contain annotator-introduced variations. Madukwe et al. (2020) highlight that varying definitions of hate speech can result in different labels to similar in-

stances. Differences can furthermore stem from the way annotators interpret the guidelines (Vidgen and Derczynski, 2020). Awal et al. (2020) find inconsistencies in the labels for the Talat and Hovy, Davidson, and Founta datasets. Isaksen and Gambäck (2020) confirm this for the Founta dataset. Similarly, van Aken et al. (2018) find doubtful labels in the Davidson dataset. Ross et al. (2016) conclude that for reliable annotations, better definitions and guidelines are needed for such a vague concept. Similarly, Fortuna et al. (2021) discuss the need for accurate and non-overlapping definitions. Röttger et al. (2022b) point out how creators should think carefully about what kind of definition would suit their task: descriptive or prescriptive.

Generalization in hate speech detection. The safety-critical nature of hate speech detection makes analyzing generalizability essential. Generalization behavior of hate speech detection models has been studied from different angles (Bourgeade et al., 2023; Yoder et al., 2022; Fortuna et al., 2021; Swamy et al., 2019; Antypas and Camacho-Collados, 2023; Markov et al., 2021; Markov and Daelemans, 2021). Swamy et al. (2019) find that balanced datasets lead to better generalization in cross-dataset studies. Fortuna et al. (2021) find models and the nature of categories to be decisive for generalization and intra-dataset performance to be indicative of generalization. Yoder et al. (2022) study variation of hate speech target identity, find that models struggle to generalize to other target identities. Antypas and Camacho-Collados (2023) find that combining different datasets is best for robustness for cross-dataset generalization. Closest to our work, Bourgeade et al. (2023) examine generalization between topic-generic and topic-specific datasets, topic referring to target group. Our investigation is more extensive, studying the relation with more aspects of hate speech definitions. Next to targets, we address dominance, explicit references to a group, and potential consequences. To our knowledge, we are the first to propose a methodology for such an investigation and conduct it.

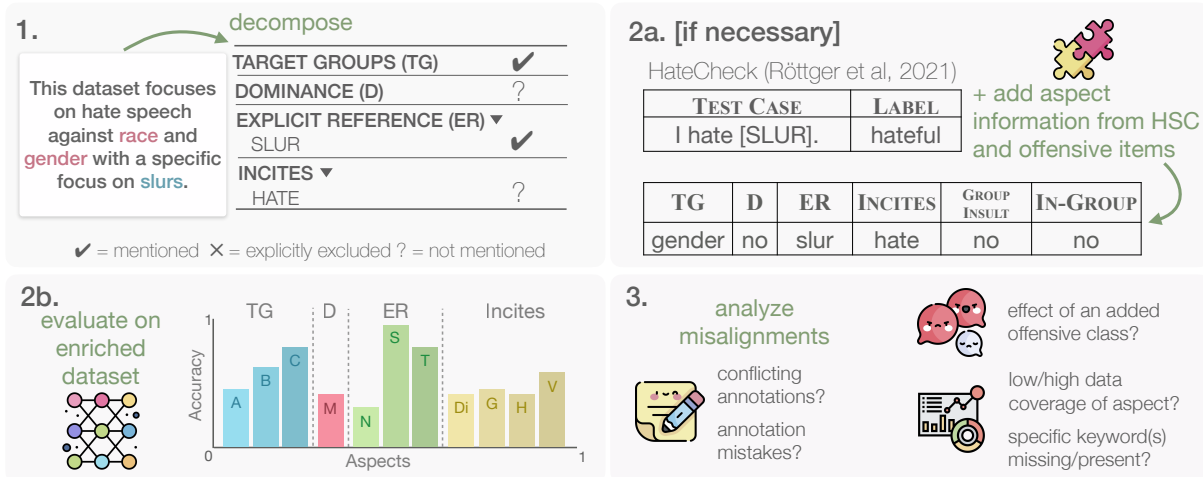


Figure 2: **DEFVERIFY**: Our proposed methodology to verify if model behavior is aligned with the dataset’s intended definition. The figure depicts each individual step in the context of hate speech detection in our paper.

3 DEFVERIFY

We propose **DEFVERIFY** a flexible framework that consists of the following steps:

1. Decompose definitions into hate speech aspects using Hate Speech Criteria (Section 3.1)
- 2a. Obtain (or create) a diagnostic set with aspect information (Section 3.2)
- 2b. Evaluate trained models according to alignment with its dataset’s definition (Section 3.3)
3. Investigate the source of misalignment in the dataset (Section 3.4)

3.1 Step 1: Decomposing Definitions

A typical hate speech dataset definition elaborates on what kind of hate speech its dataset intends to target. Hate Speech Criteria (HSC) (Khurana et al., 2022), a framework for creating definitions and annotation guidelines for hate speech tasks, identifies core aspects of hate speech that should be mentioned in a definition: either explicitly including or excluding the aspects. In practice, not many datasets specify all of them, either unintentionally or due to design choices. Mapping a definition to these aspects gives us a clear overview of what is intended to be included or excluded as hate speech and what remains underspecified.

Our first step is manually *decomposing definitions* (and annotation guidelines if publicly available) as provided by the creators of each dataset, using **HSC**. HSC identifies the following *four*² rel-

²We drop *perpetrator characteristics* as this information is not known for the diagnostic set nor the datasets we use.

evant aspects: (1) target groups, (2) whether dominant groups can be targets, (3) types of explicit references to the group, and (4) the possible consequences of a hateful statement (e.g., inciting *violence* or *discrimination*). For instance, whether the phrase “*white men are trash*” is seen as hate speech depends on aspect (2): *can a historically dominant group be considered a target or not?* Step 1 in Figure 2 illustrates this: ✓ for an aspect mentioned in the definition, ✗ if it is explicitly excluded, and ? if it is not mentioned.

3.2 Step 2a: Obtaining a Diagnostic Set

A diagnostic set provides instances representing specific phenomena. Such a set can reveal which aspects of hate speech are covered well by a model and verify if the model correctly excludes aspects considered out-of-scope, e.g. when it should ignore dominant groups as targets of hate. In *our case*, we start from **HateCheck** (Röttger et al., 2021). This diagnostic set for hate speech detection already provides a rich set of clear and straightforward examples that a model for this task should correctly recognize. It consists of 2,968 samples, of which 60.7% are hateful and 39.3% are not. HateCheck follows its own definition which is not consistent with all definitions in our set. We enrich HateCheck samples with labels specifying aspects of HSC, so we can easily adjust the labels of the diagnostic set to e.g. include or exclude a specific target group or form of hate speech. As several hate speech datasets contain an additional *offensive* class, we also extend HateCheck with *offensive* samples. We make all our code and enriched version of Hate-

Check available.³ This step is therefore only necessary when new phenomena are investigated in future work, e.g. perpetrator characteristics. We specify the additions we made in this step below.

Adding aspect information. For each aspect, we indicate the possible sub-aspects/categories that we label with. See Step 2a in Figure 2 for an example. **TARGET GROUP:** Target groups in HateCheck: *gender, sexual orientation, race, religion, nationality, and disability*.

DOMINANCE: yes or no, reflecting if it is a dominant group or not. All groups in the original HateCheck are non-dominant groups. We hence add two dominant groups: *men* and *white people*. It is critical to rigorously evaluate for these dominant groups, even when they are not considered targets of hate speech, to ensure that the model does not *erroneously* identify them as targets of hate. Please note that we do not prescribe, we leave it up to the researcher to assess how they want to treat dominant groups. **EXPLICIT REFERENCE:** how is the group mentioned: group characteristic, stereotype, or slur. Group Characteristic: assigned at the mention of the name. Other types of group characteristics are not present in the data. Stereotype: if something reads as a stereotype, it is marked as such, e.g. "typical of <GROUP>". We also factor in real-world stereotypes, e.g. how society views women in leadership positions. If an example contains a general stereotype (not group-specific), it is marked as only *stereotype, group characteristic* being absent, to prevent a high correlation between the two. Slur: mostly marked in HateCheck based on the type of capability. Some slurs are also marked as a stereotype, such as *c*m*lf*ck*r*.

INCITES: Hate: Most cases incite hate due to the nature of the diagnostic set. Violence: Marked when violence is a potential consequence, both explicit and implied. This also implies hate but for distinction we only consider these to be violent. Discrimination: Marked if the statement leads to exclusion. Cases that incite both violence and hate also incite discrimination but are not marked as such for distinction. **GROUP INSULT:** yes or no. Marked *yes* when a negative attribute is assigned to a group. If someone expresses their own negative sentiment with a group characteristic, e.g. *I dislike <GROUP>*, we do not consider it a group insult.

IN GROUP: yes or no. Only *yes* for reclaimed slurs present as we assume no exceptional context.

³<https://github.com/urjakh/defverify/>

Adding offensive cases. Hate speech datasets can also come with a third class; *offensive*. This clashes with HateCheck which only contains *hateful* and *non-hateful* classes. Upon analyzing definitions and samples for *offensiveness* of datasets used in this paper, this class mostly captures vulgar language in general (e.g. abusive words), violence toward individuals, and individual insults without any explicit link to group identity. We therefore treat two HateCheck functionalities as *offensive*: *profanity_nh* and *target_indiv_nh* (n=165). We also add 120 instances with vulgar or abusive words in offensive and non-offensive (+9 cases) contexts, violent threats, and individual insults. We use the *ableism* class from [Manerba and Tonelli \(2021\)](#) as inspiration for individual insult and threatening, where originally all samples included a form of slur that we generalize to "you".

3.3 Step 2b: Expectation vs. Reality

We now train a model on the dataset we are investigating. The model is then evaluated on the diagnostic set from Step 2a. We measure the model's accuracy on instances belonging to different sub-aspects. We can compare the model's performance to our expectations based on the definition and determine if the model's behavior reflects what was intended. In our case, HateCheck consists of very obvious and straightforward samples that we would expect a model to classify correctly. Hence, we expect performance to be quite good (e.g. at least 80% accuracy) on the desired aspects.

3.4 Step 3: Initial Root Cause Analysis

If the model performs unexpectedly on an aspect, we revisit the training data. Large datasets do curb the potential of conducting an extensive investigation for the failure. As a starting point, we use keywords related to the failing aspect to manually examine dataset coverage or inspect annotation consistency for similar training samples.

4 Retrospective Analysis

We now show the utility of DEFVERIFY and apply it to six widely used hate speech detection datasets.

4.1 Datasets

We consider the following widely used hate speech datasets: *TalatHovy* ([Talat and Hovy, 2016](#)) ([TalatHovy](#)), *Davidson* ([Davidson et al., 2017](#)), *Measuring Hate Speech Corpus* ([Kennedy et al., 2020](#)) ([MHSC](#)), *Dynamically Generated Hate*

Speech Dataset (Vidgen et al., 2021) (**DGHS**), *HateXplain* (Mathew et al., 2021) (**HX**), and *Founta* (Founta et al., 2018). The datasets are chosen based on their wide usage and diversity in hate speech definition. Note that **Davidson** and **HX** datasets have an *offensive* class and **Founta** an *abusive* class.

4.2 Experimental Setup

We fine-tune fBERT (Sarkar et al., 2021) and RoBERTa (Liu et al., 2019) on the six datasets, both shown to be competitive models for hate speech detection with different strengths in Bourgeade et al. (2023). For each dataset, we train 5 random seeds. Model selection is based on the validation macro F1. See Appendix A for technical details. For consistency with the original definitions, we keep the original labels of the datasets.⁴ Hence, we preserve the *offensive/abusive* class for the datasets that contain it. This enables learning the nuances, especially when dealing with borderline cases. Such cases may not be *hate speech* per se, but still contain upsetting content that warrants flagging, e.g. for a law-based context. We evaluate with HateCheck, using accuracy (as in the original paper), precision, and recall. Due to the nature of the investigation, we filter out capabilities that test robustness toward spelling mistakes. We further do a separate analysis for dominant groups (which we discuss in Section 4.3.2) and based on the definition, we mark these as *hate speech* targets or not. **TalatHovy** dataset uses two specific sub-types of hate speech (*sexism* and *racism*). Thus, we consider instances where *women* or *trans people* are targeted as *sexism* and *black people*, *Muslims*, and *immigrants* as *racism*. All other instances are labeled non-racist or non-sexist.

4.3 Applying DEFVERIFY

4.3.1 Decomposing Definitions

For each dataset, we take the definition (and annotation guidelines for **DGHS**) from the original paper and provide the decomposed aspects in Table 1 and covered target groups in Figure 3.⁵ We describe this process for each dataset, including the used definitions, in detail in Appendix B.

Included aspects in dataset definitions. We see that the **DGHS**, **TalatHovy**, and **MHSC** datasets

⁴The class distribution of the datasets can be found in Table 1. For the HateXplain dataset, *undecided* is excluded.

⁵Since **Davidson** does not mention any target groups in their definition, we leave it out of the plot.

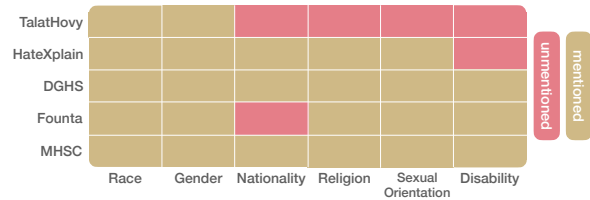


Figure 3: Overview of which target groups are mentioned in the definition for each dataset.

include the most aspects in their definition. Only one or two definitions specify whether they consider solely non-dominant groups or also include dominant groups as potential hate speech targets. The definition of **HX** is limited to a specification of the target groups. We can infer from this that dominant groups are included.

Expected model behavior Models trained on a respective dataset should have high performance identifying aspects indicated with a ✓ in Table 1 as hate speech and aspects marked with ✗ as non-hate speech. We do not construct expectations for aspects marked with a ? as its absence from the definition can introduce more space for annotator subjectivity or hint at limited to no data coverage.

We make certain assumptions for our expectations when decomposing the definitions of the datasets. Though most datasets do not clearly define that they require an *explicit reference* to the group being attacked, we do assume that this is intended to distinguish between hate speech and other forms of toxic language. We thus expect good performance on *group characteristics*, the aspect which also captures references to a group by their name, for all models. Additionally, when a hateful utterance can lead to discrimination, we automatically assume this to be a group insult as well.

Most expectations can be inferred directly from the table. For **TalatHovy**, we expect models to perform well on non-dominant groups based on their *race* and *gender*, as well as *religion* and *nationality* based on our observations in the dataset. Due to the focus on race and gender-based minorities, we do not expect the model to classify dominant groups from these categories (i.e. men and white people) or other target groups as hate speech. **Davidson** does not mention any target group or dominance. Thus, it is unclear on which target groups the model will do well or how it will respond to dominant groups.

	SIZE	LABEL COMPOSITION	TG	Do	ID	IV	IH	GI	St	GC	SI
DGHS	41,255	54% HS, 46% not	✓	✗	?	✓	✓	✓	✓	✓	✓
TalatHovy	16,914	20% sexist, 11.7% racist, 68% neither	✓	✗	?	✓	✓	✓	✓	✓	✓
MHSC	39,565	26.2% HS, 78.8% not	✓	✓	✓	✓	✓	?	?	✓	?
Davidson	24,802	5.8% HS, 77.4% offensive, 16.8% neither	?	?	?	✓	✓	✓	?	?	?
Founta	99,996	5.0% HS, 27.2% abusive, 14.0% spam, 53.8% normal	✓	?	?	?	✓	✓	?	?	?
HX	20,148	29.5% HS, 27.2% offensive, 38.8% normal, 4.5% undecided	✓	✓	?	?	?	?	?	?	?

Table 1: *Decomposing* the datasets by mapping their respective definitions to the different aspects from **HSC**. The results are accompanied with information about the dataset size and label distribution (where **HS** stands for hate speech). ✓ indicates that the aspect is mentioned in the definition and is considered, ✗ indicates that the aspect is explicitly not considered for the dataset, and ? means that it is unmentioned and hence we do not know what to expect. **TG**: target groups, **Do**: dominant groups, **ID**: incitement of discrimination, **IV**: incitement of violence, **IH**: incitement of hate, **GI**: group insult, **St**: stereotype, **GC**: group characteristics, **SI**: slur.

4.3.2 Expectation vs. Reality

We now evaluate the models of each dataset on our diagnostic evaluation set, the enriched version of HateCheck.⁶ Each dataset’s validation and test set results can be found in Appendix C. We see a large drop in performance from the original test to diagnostic tests, indicating that the models learned the data, but not the intended aspects.

Alignment of model behavior with definition.

Figure 4 showcases the results on all the tested aspects in HateCheck for each dataset. For **Davidson**, we expected good performance for incitement of violence and hate and group insults. We only see this for violence, the other two achieving rather average performance. Similarly, with **Founta**, we see average performance for the aspects we expected to perform well: incitement of hate and group insult. Although *gender*, *sexual orientation*, and *religion* get recognized as hateful, there are a lot of false positives. **MHSC** has low recall on identifying hate speech for target groups in general. We see the same for the dominant groups (explicitly covered). Low to average performance is achieved for all the other aspects it was expected to do well on except for *violence*. **HX** underperforms severely; regardless of whether an aspect is mentioned in the definition or not, it predominantly results in false negatives. **TalatHovy** models correctly identify excluded target groups (dominant groups, disability, and sexual orientation) as true negatives. Race (mostly classified as *sexist* by the model), gender,

and nationality (latter both classified as *neither* by the model) perform badly. Other mentioned aspects also underperform. **DGHS** performs well on most aspects, only misclassifying white people as false positives, while they were excluded as a target in the definition. **We find that many aspects, even when mentioned in the definition, do not get picked up by the model correctly, either overclassifying or failing to recognize hateful aspects.**

Removing offensive when training. Datasets with an extra *offensive* class have a disadvantage when the evaluation set only considers the other two classes. We hence remove the *offensive* class from **Davidson**, **Founta**,⁷ and **HX**. We show the results in Figure 5. For all three datasets, most aspects see an increase in accuracy, as expected. For **Davidson**, the impact is biggest for aspects related to *explicit references* with good results on *group characteristic* in general. When it comes to consequences, *hate* and *group insult* increase, just as **Founta**. Additionally, for **Founta**, we see that *discrimination* and instances covering *stereotype* or *group characteristic* have an increased accuracy, while *slurs* remains challenging. For **HX**, *target groups* and *dominance*’s accuracies increase in general, while all *explicit references* increase. It still struggles with any *discriminatory* aspects.

Performance on offensive items. We evaluate performance on 265 added offensive statements and show the results in Figure 6. In general, the

⁶For brevity, we refer to a model’s performance trained on a specific dataset as *dataset’s* performance.

⁷As the original proportion of hateful samples is only 1/10, we improve the imbalance by increasing this to 1/3

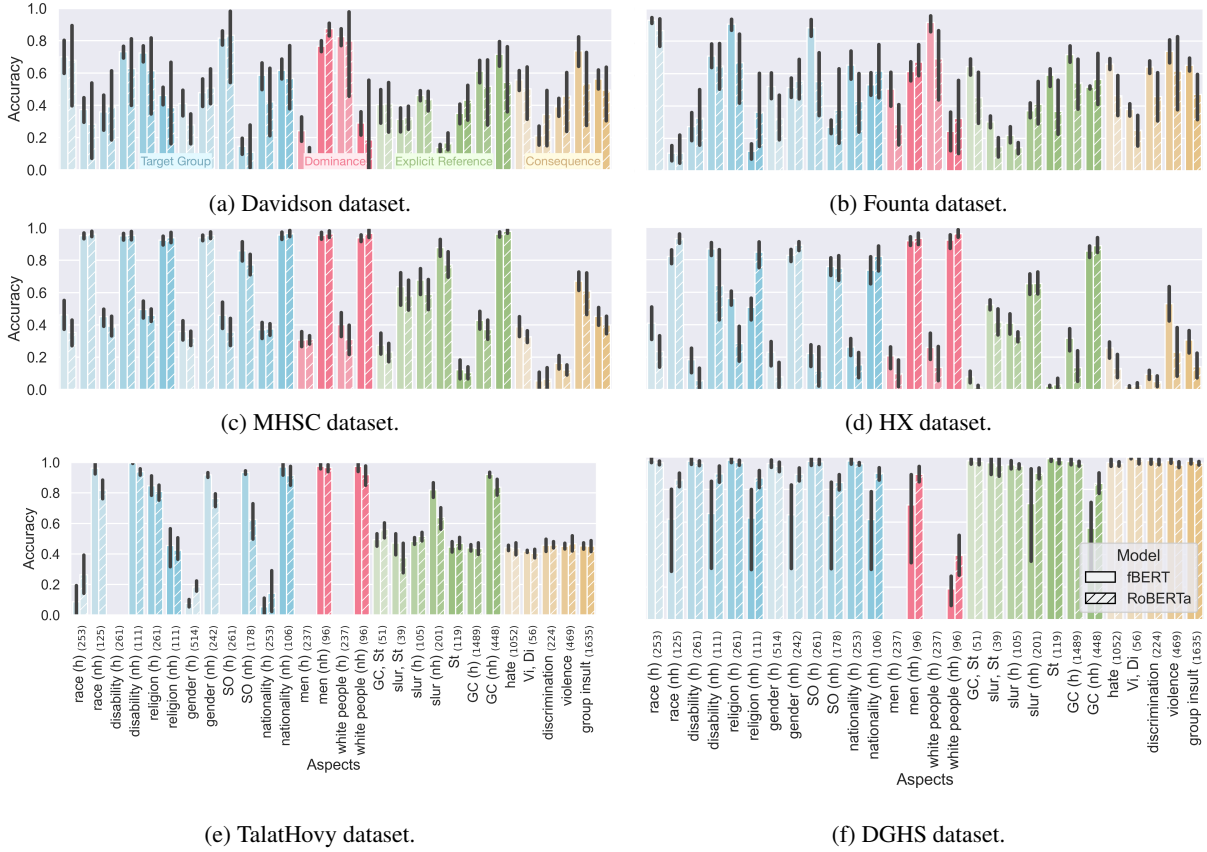


Figure 4: Accuracy for the different aspects on HateCheck, with the number of samples for each aspect in brackets. We separate the accuracy for aspects with both hate (h) and non-hate (nh) samples. **SO**: sexual orientation, **GC**: group characteristics, **St**: stereotype, **Vi**: incitement of violence, **Di**: incitement of discrimination.

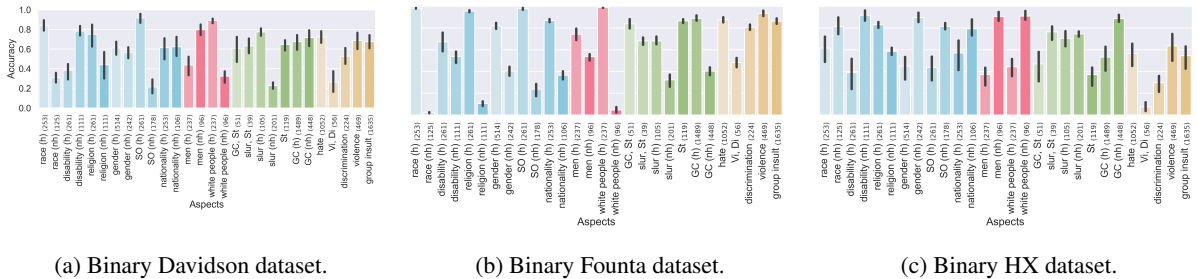


Figure 5: Accuracy for the different aspects on HateCheck, when removing the *offensive* classes from the **Davidson**, **Founta**, and **HX** datasets with fBERT.

two models perform similarly across most datasets. For **Davidson**, only roughly 50% is seen as offensive and for Founta this increases to 60%. This is against expectations. **HX** models (Figure 6d) primarily view the offensive statements as *non-offensive*, exposing the lack of generalization to unseen *offensive* data. **MHSC** (Figure 6c), **TalatHovy** (Figure 6e), and **DGHS** (Figure 6f) models correctly predict most samples to be *non-hateful*.

Cross-dataset performance. In Figure 7, we showcase the accuracies achieved when evaluat-

ing a model trained on dataset A on the test set of dataset B (where $A \neq B$). Due to different labels in different datasets, we only focus on how many of the hate speech instances of a dataset the model recognizes correctly as hate. For example, we expect **DGHS** to perform well on **TalatHovy** due to their similarity in terms of aspects. **We do not observe that datasets with similar definitions tend to yield better performance on each other's test sets compared to sets with dissimilar definitions.** Overall, we find that **DGHS** achieves the best performance on all the datasets, with great results on

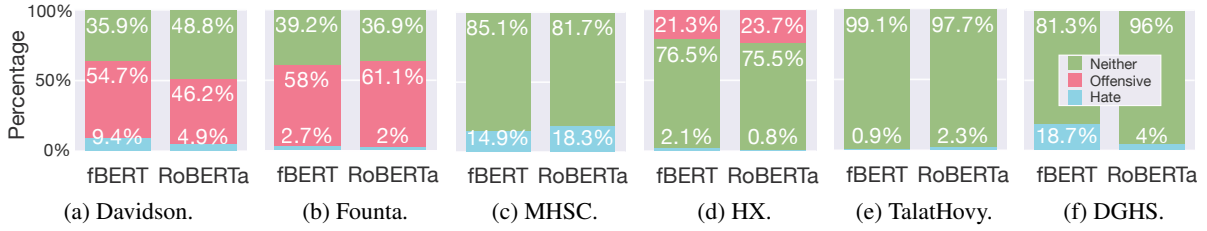


Figure 6: Percentage of predictions on 285 offensive samples for each dataset.

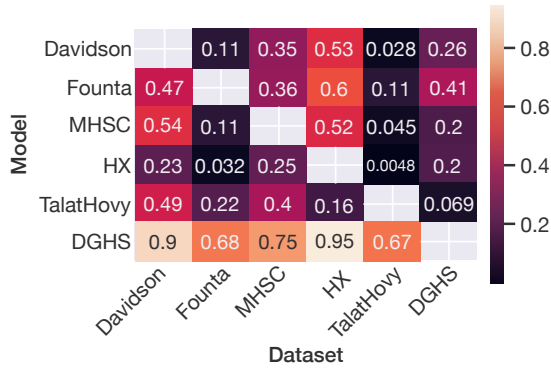


Figure 7: Accuracy (averaged over seeds) of identifying hate instances on cross-dataset test sets with fBERT.

Davidson (0.9) and **HX** (0.95) particularly. The rest of the models from other datasets perform notably low. Datasets with an *offensive* class do not stand out on each other performance-wise.

4.4 Initial Root Cause Analysis

Understanding why a model failed to capture an aspect it was expected to cover can lead to fixing those gaps. We thus conduct an early-stage analysis for some of the observed misalignments between model behavior and the original dataset’s definition. We do a simple keyword search in the dataset’s training set, using terms relevant to the failing hate speech aspects. These terms primarily stem from HateCheck. We then investigate the label distribution for matching samples and data coverage. If those do not show useful insights, we look at the consistency of annotations. While this analysis does not give a holistic view of the reasons for misalignment, it is a starting point to get more targeted insights from what we are training our model on. In case of sufficient relevant training data points, this could hint at model underfitting. A lack thereof could mean adding more relevant data points. We propose more thorough ways to analyze misalignments in the future in Section 5.

TalatHovy performs badly on sexism and racism. Since *sexism* in **HateCheck** consists of *trans peo-*

ple and *women*, we use those as keywords to find training samples. Only 0.8% of the training data mentions *trans*, with only 2 labeled sexist. For *women/woman* there are some mislabels, e.g. “A woman must be obedient” as *racism*. Several samples labeled as sexist contain the phrases “I’m not sexist”, “Call me sexist”, or “#sexist”. For *racism*, we find the term *black* in only 64 samples with only 11 marked as *racist*. Variations of the n-word and f-word are only found in 2 samples.

DGHS recognizes white people as hate speech.

Many of the cases in the training set that contain *white* express superiority, where 746 samples are deemed hate and 394 are not. In contrast, for *man* which it does not recognize as hate speech, 202 of 376 samples are marked as not hate. *Men*’s label distribution is around 50% for both classes, with a lot of intersectional cases when marked as hate.

MHSC does not recognize men as hate speech.

In total there are 528 samples in the training set that mention *men*. Out of these, 120 are hate, and 408 are labeled as not. We also see some inconsistencies in annotation where “Wow men are f*ck*ng trash and disgusting” is labeled as hate but two other examples calling *men trash* as not hate.

5 Discussion

Analyzing DGHS’ higher performance. We find that models trained on DGHS achieve the best results on HateCheck and cross-dataset evaluation experiments. The authors between HateCheck and DGHS overlap, which is addressed in the paper of DGHS: only 0.05% of the dataset matches HateCheck, but they still call for caution. The cross-evaluation results hint that the good performance on HateCheck does not solely stem from that. DGHS comes with a beneficial dataset composition (large training set with balanced class distribution) and is synthetically generated using an adversarially contrastive approach. The latter can reduce false positives and negatives. We ran additional experiments

adjusting the dataset to less ideal circumstances: a small training set and imbalanced class distribution. The exact results can be found in Appendix D, but performance only suffers significantly for the imbalanced cross-evaluation.

Root Cause Analysis. Our keyword-matching analysis is just the first step to locating the failure point for misaligned behavior. In the current scenario, it gives us a limited view of what is in the data, what the model relies on for predictions, and annotation bias at a large scale. More *data* and *model* interpretability techniques are needed, e.g. using influence functions (Koh and Liang, 2017) to understand what evidence a model used in the case of misclassifications. Important insights can come from analyzing which samples in the data have high annotator variation and which are straightforward. This could give insights into which datasets are more prone to misclassify edge cases. Many datasets do not provide individual annotator annotations nor annotation guidelines, which are essential elements for a full analysis of data quality. **In future work**, we intend to aim for a more holistic view to attack the cause of misalignment from different angles with DEFVERIFY, disentangling the biases shown Figure 1.

The complexity of statements in the diagnostic dataset. We use HateCheck due to its straightforward samples and the large number of different capabilities that a hate speech detection system should have. The simplicity of these statements in HateCheck is essential to establish whether the model can at least recognize the simplest forms of hate speech. However, this excludes more complex and implicit forms of hate speech, which are more representative of such language in the wild. It is even more so important that we should also be aware of how our model fares on those. **Future work** may focus on extending the diagnostic dataset to also cover rather complex hate speech samples.

6 Conclusion

We propose DEFVERIFY, a methodology to investigate if a hate speech dataset correctly captures the intentions of a dataset creator or user. We achieve this by (1) eliciting and formalizing the definitions underlying popular hate speech datasets through the lens of Hate Speech Criteria, (2a) enriching HateCheck with information based on this lens

and with extra offensive statements, (2b) evaluating potential misalignments between dataset definitions and model capabilities on six widely-used hate speech datasets, and (3) an early-stage analysis to understand observed gaps. Our findings indicate that for most datasets, their respective trained models failed to correctly recognize aspects, even when they were explicitly mentioned in the dataset’s definition.

Takeaways. Our method provides an approach to pinpoint if a dataset provides the type of hate speech a creator intended to address. The gaps we found using DEFVERIFY showcase that evaluating model behavior is important when *creating* datasets as well as *deploying* models. Our method can also help users select a dataset for their specific purposes. Our methodology is thus more than just a tool to investigate definition alignment in hate speech; it supports more informed dataset creation and usage to help ensure more reliable models in safety-critical contexts. For more reliable models in the wild we need (1) clarity in terms of what aspects are explicitly included and excluded in the definition, (2) clearly reflected in annotation guidelines that (3) should be made publicly available (4) along with individual annotations, and (5) rigorously evaluated before deployment.

Limitations

While we aim to rigorously analyze model behavior and capture performance on several aspects of hate speech that have not been tested before, there are a few limitations in our work.

Considering other relevant factors. Our paper is a first step in unraveling the disparity between the intended definition and what is in the data. For our methodology, we have chosen a method that can be applied with information that is publicly available for all datasets and is relatively easy and quick to carry out (since it is unlikely that non-research users will carry out an extensive analysis of annotation and data itself). There are many factors between the starting point of a general definition and model behavior that influence the final model: the annotation guidelines, the data itself (size, properties, selection method), the quality and number of the annotators, and the models used for training, to name a few. Each of these can and should ideally be studied at their own account to gain full insight into the suitability of a dataset. Our

approach merely provides a first step based on information and resources that are generally publicly available to potential users of the data, for off-the-shelf usage: the definition provided when the data was introduced, standard models, and the extended diagnostic test that we provide with this paper. For example, for our analysis, we would ideally have taken into account the annotation guidelines more but we could only do this for the **DGHS** dataset. We made use of information that was publicly available and most datasets do not provide annotation guidelines. This is why we focus on definitions provided in the papers introducing the datasets.

Monolingual analysis. Due to the large availability of data and definition diversity, our experiments are conducted on English datasets. As Multilingual HateCheck (Röttger et al., 2022a) is a multi-lingual extension of HateCheck, in the future, those who have knowledge of other languages and can understand relevant nuances can extend this research to other languages if more datasets with diverse definitions are available.

Model usage. Our results are primarily showcased using two models: fBERT and RoBERTa. However, it is not our intention to make generic statements about model generalizability or compare performance between models. To reduce the chances of observations being due to model-specific behavior, we chose two competitive models that are widely used for hate speech detection and revealed different strengths in prior work. Further, the reasonable to high macro-F1 scores indicate that the models are learning the dataset, ruling out a sub-optimal training pipeline. Our results indicate similar performance between fBERT and RoBERTa but it is not guaranteed that this trend extends to other model types.

We also specifically do not experiment with autoregressive large language models in combination with in-context learning as we do not always know what the training data is. Not only does this lead to complexities due to data leakage but in-context learning is not well-understood and thus cannot be relied upon for a safety-critical setting like hate speech detection.

Taking law into account. Our analysis does not take the law into account. Where HateCheck gets its target groups from UK law, MHSC looks at the legal definitions from the US. However, the other datasets do not discuss the impact of law on their

definition or type of hate speech they are addressing. Thus, we leave it out of our discussion as this would be implicitly encapsulated by the definition (or annotation guidelines) itself. However, the flexibility of our framework can easily allow for legal interpretations to be taken into account, given that a diagnostic dataset includes this type of information and can be modular in its labels to facilitate different laws.

Geographic bias. Based on the used datasets, we take a Western perspective for our choice of dominant and non-dominant groups. We include *men* and *white people* as dominant groups as these are the two most popular dominant groups along the identifier categories of gender and race. It is a very subjective take to include dominant groups as a target of hate speech, and our framework leaves it to the user to determine their stance. For instance, HateXplain explicitly considers both *men* and *white people* as targets of hate speech while *DGHS* explicitly excludes them. In no way do we propose or endorse any of the used definitions. When excluded, it becomes even more important to test for these groups to verify and ensure that a model does not accidentally recognize hate speech. This also counts for other aspects from HSC.

Ethics Statement

We believe our proposed methodology contributes to the safer deployment of NLP systems. Knowing what is in the data and what the model is capable of is essential for responsible usage. Our proposed methodology intends to verify if a model behaves as originally intended for safety-critical applications. We focus on the task of hate speech detection with an analysis of existing datasets. We show the potential risks of creating or using a dataset without rigorous and fine-grained analysis of dataset creation and model behavior evaluation. Our methodology serves as a tool to identify issues in the dataset if model behavior is not according to expectations thus building toward better-constructed datasets and lesser unexpected inappropriate model performance.

A note of caution, our methodology does not cover all test items needed when deploying a model for safe and responsible usage. Every project needs thorough testing rounds thus adding project-appropriate test cases too. For instance, edge cases or cases that introduce subjectivity are not covered,

which is necessary to understand the full picture of model behavior. As such, this is just a first step.

Acknowledgments

We thank Iliia Markov for his useful comments on the draft version of this camera-ready and Pia Sommerauer for her valuable feedback on an earlier draft. We also thank the anonymous reviewers for their comments that helped improve this paper. All remaining errors are our own. This research was (partially) funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research. The figures and affiliation emojis have been designed using resources from *flaticon.com*; the crab is created by *mihimihi*, the tulips by *SBTS2018*, analysis chart by *Eucalyp*, neural network by *flatart_icons*, and the rest of the emojis by *freepik*.

References

- Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. [Identifying and measuring annotator bias based on annotators' demographic characteristics](#). In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.
- Dimosthenis Antypas and Jose Camacho-Collados. 2023. [Robust hate speech detection in social media: A cross-dataset empirical evaluation](#). In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 231–242, Toronto, Canada. Association for Computational Linguistics.
- Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings 23*, pages 57–64. Springer.
- Md Rabiul Awal, Rui Cao, Roy Ka-Wei Lee, and Sandra Mitrović. 2020. On analyzing annotation consistency in online abusive behavior datasets. *arXiv preprint arXiv:2006.13507*.
- Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.
- Tom Bourgeade, Patricia Chiril, Farah Benamara, and Véronique Moriceau. 2023. [What did you learn to hate? a topic-oriented analysis of generalization in hate speech detection](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3495–3508, Dubrovnik, Croatia. Association for Computational Linguistics.
- Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.
- Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30.
- Paula Fortuna, Juan Soler, and Leo Wanner. 2020. [Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.
- Paula Fortuna, Juan Soler-Company, and Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524.
- Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12.
- Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, et al. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference*, pages 229–233.
- Hugo Lewi Hammer. 2014. Detecting threats of violence in online discussions using bigrams of important words. In *2014 IEEE Joint Intelligence and Security Informatics Conference*, pages 319–319. IEEE.
- Nick Haslam and Michelle Stratemeyer. 2016. Recent research on dehumanization. *Current Opinion in Psychology*, 11:25–29.
- Vejbørn Isaksen and Björn Gambäck. 2020. [Using transfer-based language models to detect hateful and offensive language online](#). In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 16–27, Online. Association for Computational Linguistics.

- Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multi-task deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.
- Urja Khurana, Ivar Vermeulen, Eric Nalisnick, Marloes Van Noorloos, and Antske Fokkens. 2022. [Hate speech criteria: A modular approach to task-specific hate speech definitions](#). In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 176–191, Seattle, Washington (Hybrid). Association for Computational Linguistics.
- Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR.
- Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. [Benchmarking aggression identification in social media](#). In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Guntan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. [Datasets: A community library for natural language processing](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- 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*.
- Kosisochukwu Madukwe, Xiaoying Gao, and Bing Xue. 2020. [In data we trust: A critical analysis of hate speech detection datasets](#). In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 150–161, Online. Association for Computational Linguistics.
- Marta Marchiori Manerba and Sara Tonelli. 2021. [Fine-grained fairness analysis of abusive language detection systems with CheckList](#). In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 81–91, Online. Association for Computational Linguistics.
- Iliia Markov and Walter Daelemans. 2021. [Improving cross-domain hate speech detection by reducing the false positive rate](#). In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 17–22, Online. Association for Computational Linguistics.
- Iliia Markov, Nikola Ljubešić, Darja Fišer, and Walter Daelemans. 2021. [Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection](#). In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 149–159, Online. Association for Computational Linguistics.
- Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.
- Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. *Frontiers in artificial intelligence*, 3:55.
- Bjorn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. In *3rd Workshop on Natural Language Processing for Computer-Mediated Communication/Social Media*, pages 6–9. Ruhr-Universität Bochum.
- Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022a. [Multilingual Hate-Check: Functional tests for multilingual hate speech detection models](#). In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, Washington (Hybrid). Association for Computational Linguistics.
- Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022b. Two contrasting data annotation paradigms for subjective nlp tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190.
- Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Talat, Helen Margetts, and Janet Pierrehumbert. 2021. Hatecheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58.
- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. [The risk of racial bias in hate speech detection](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

- Diptanu Sarkar, Marcos Zampieri, Tharindu Ranasinghe, and Alexander Ororbia. 2021. [fBERT: A neural transformer for identifying offensive content](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1792–1798, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Anna Schmidt and Michael Wiegand. 2017. [A survey on hate speech detection using natural language processing](#). In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.
- Andrew Sellars. 2016. Defining hate speech. *Berkman Klein Center Research Publication*, (2016-20):16–48.
- Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. 2019. [Studying generalisability across abusive language detection datasets](#). In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 940–950, Hong Kong, China. Association for Computational Linguistics.
- Zeerak Talat and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.
- 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 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42, Brussels, Belgium. Association for Computational Linguistics.
- Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.
- Bertie Vidgen, Tristan Thrush, Zeerak Talat, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682.
- Bertie Vidgen and Taha Yasseri. 2020. Detecting weak and strong islamophobic hate speech on social media. *Journal of Information Technology & Politics*, 17(1):66–78.
- Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2014. Cursing in english on twitter. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 415–425.
- William Warner and Julia Hirschberg. 2012. [Detecting hate speech on the world wide web](#). In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Michael Yoder, Lynnette Ng, David West Brown, and Kathleen Carley. 2022. [How hate speech varies by target identity: A computational analysis](#). In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 27–39, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- 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 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

A Technical Setup

We fine-tune our models with fBERT (Sarkar et al., 2021) (110M parameters),⁸ and RoBERTa-base (Liu et al., 2019) (125M parameters). We follow the pre-processing steps as mentioned in Bourgeade et al. (2023) and replace usernames, URLs, and links with placeholders. For fBERT, we use the original hyperparameters and the learning rate from Bourgeade et al. (2023) as it gave more stable results. We train with a batch size of 8, warmup on 10% of the total training steps, a weight decay of 0, and a linear learning rate of $5e^{-5}$ for 3 epochs. For RoBERTa-base we use the original hyperparameters used for GLUE tasks, sometimes using a batch size of 8 with gradient accumulation step 2 due to memory constraints. Warmup is done on 6% of the total training steps, weight decay of 0.1, and a linear learning rate of $5e^{-5}$ for 10 epochs. For each dataset, we train 5 random seeds. Our models are all trained on an RTX 2080 Ti. Each experiment took a maximum of 2 hours, meaning that the experiments took around 90 hours in total to finish.

⁸109.483.778 parameters to be exact

Model selection is based on the macro F1 on the validation set. In some cases, the validation loss is much higher (> 0.1) than a similar (but slightly less) performing checkpoint. We then choose the one with better loss but slightly worse macro F1.

Splits. Not all datasets have pre-defined splits, thus in those cases, we create our own random splits with 80% train samples, 10% validation, and 10% test.

Packages. We use HuggingFace Transformers (Wolf et al., 2020) and Datasets (Lhoest et al., 2021) for our experiments.

B Decomposition Process

For each dataset, we introduce it, decompose the definition (and annotation guidelines, if provided) using Hate Speech Criteria (HSC), and provide the decomposed definition. We conclude with our model behavior expectations. These datasets are chosen based on their wide usage and diversity in their hate speech definition. We mark the different aspects in the definitions in the following way: target group, [dominance], consequence, and {explicit reference}.

The TALATHOVY (Talat and Hovy, 2016) dataset focuses particularly on *racism* and *sexism* toward minorities. It consists of 16,914 tweets, of which 20% are *sexist*, 11.66% are *racist*, and 68.34% are neither.

TALATHOVY Definition

(Page 2, Section Data) “A tweet is offensive if it

1. uses a sexist or racial slur
2. attacks a minority
3. seeks to silence a minority
4. criticizes a minority (without a well founded argument)
5. promotes, but does not directly use, hate speech or violent crime
6. criticizes a minority and uses a straw man argument
7. blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims

8. shows support of problematic hash tags. E.g. “#BanIslam”, “#whoriental”, “#whitegenocide”
9. negatively stereotypes a minority
10. defends xenophobia or sexism
11. contains a screen name that is offensive, as per the previous criteria, the tweet is ambiguous (at best), and the tweet is on a topic that satisfies any of the above criteria.”

Decomposed: For this dataset, hate speech is defined as language targeted at a [non-dominant person or group] based on their race and gender and incites violence and hate or insults a group through the usage of {negative stereotypes, group characteristics, or slur}. **Expectations:** We expect models to perform well for non-dominant groups based on their race and gender, as well as religion and nationality due to their closeness to racism in the dataset. Furthermore, it should recognize incitement of violence, hate, and group insult. It should also capture all three types of explicit references to a group. As the dataset focuses on race and gender-based minorities, we do not expect the model to classify dominant groups from the same category (i.e. men and white people) or other target groups as hate speech. As incitement of discrimination is not explicitly mentioned, it is unknown if the model can capture this phenomenon and generalize well.

The DAVIDSON (Davidson et al., 2017) dataset targets hate speech in general. The dataset comprises 24,802 tweets, of which 5.77%⁹ are *hate speech*, 77.43% *offensive*, and 16.80% *neither*.

DAVIDSON Definition

(Page 1, Introduction) “We define hate speech as language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group. In extreme cases this may also be language

⁹The percentages were calculated from the dataset from <https://github.com/t-davidson/hate-speech-and-offensive-language>. These values might slightly deviate from the ones mentioned in the original paper.

that threatens or incites violence, but limiting our definition only to such cases would exclude a large proportion of hate speech. Importantly, our definition does not include all instances of offensive language because people often use terms that are highly offensive to certain groups but in a qualitatively different manner. For example some African Americans often use the term n*gga in everyday language online (Warner and Hirschberg, 2012), people use terms like h*e and b*tch when quoting rap lyrics, and teenagers use homophobic slurs like f*g as they play video games. Such language is prevalent on social media (Wang et al., 2014), making this boundary condition crucial for any usable hate speech detection system”

(Page 2, Section Data) “They [Annotators] were provided with our definition along with a paragraph explaining it in further detail. Users were asked to think not just about the words appearing in a given tweet but about the context in which they were used. They were instructed that the presence of a particular word, however offensive, did not necessarily indicate a tweet is hate speech.”

Decomposed: For this dataset, hate speech is defined as language targeted at a group to incite hate or a person of a group to insult or incite violence through the usage of {group characteristics}. **Expectations:** No target group or anything about dominance is mentioned. Thus, it is unclear on which target types the model will work well, nor if dominant groups are also covered. The same holds for *slurs* and *stereotypes*. However, we do expect the models to capture incitement of hate, violence, and group insult, except for incitement of discrimination, which is unknown.

The MEASURING HATE SPEECH CORPUS (Kennedy et al., 2020) dataset targets a variety of hate speech types and contains 39,565 comments. Of these comments, 26.2% is hate speech and 73.8% is not.¹⁰

¹⁰The instances come with a score which we translate to labels (i.e. hate speech or not) using the threshold from the original paper: <https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech>

MHSC Definition

(Page 7, Section 3.1) “We draw from the legal definition of hate crimes in the United States that protects against discriminatory actions targeting one of the following protected groups: race, religion, ethnicity, nationality, gender, sexual orientation, gender identity, and disability. In identifying groups within these broad categories, we include subjugated groups that have been discriminated against in the United States, as well as power-dominant groups who have not. Targeting of a group or an individual on the basis of their membership in a group is common to most definitions of hate speech (Sellars, 2016). Not only do we adopt this convention, but we allow for intersectional or overlapping identities to be selected for further analysis. We consider intersectional identities and the possibility of compounding hate speech directed at an individual who belongs to multiple groups.

Speech can also lead to individual acts of violence and when targeted against a group, genocide and extermination. The “dangerous speech” framework ties the effects of hateful speech to actions that it can incite (Benesch et al. 2018). Dehumanization, such as radio broadcasts in Rwanda referring to the Tutsi people as cockroaches, is directly linked to later genocidal killing of that group. Incitement towards violence is a narrowly defined concept under US law, and the dangerous speech framework that we use takes a broader view of the link between cause and effect. Sellars (2016) points out that the accumulated affects of anti-Semitic or racist speech can have multi-generational impacts on the well-being of individuals in a group born long after hateful speech was original created. Given the complexities of these concepts, we focus on calls to individual violence or collective extermination, with the idea that these are the final step after expression of hate and deeming a group inferior or inhuman.

Table 2 describes the eight levels of our theorized hate speech scale. The positive levels on the scale designate hate speech of increasing severity. Unlike many existing scales, our typology includes both neutral and positive identity speech, represented by 0 and negative values, respectively. Following Anti-Defamation League (2016) and Stanton (2013), we place speech supporting the systematic killing of a specific group as the most severe form of hate speech. Viewing other types of hate speech as pathways to genocide, we pay special attention to individuals threats of violence and dehumanization that may justify violence.”

Decomposed: For this dataset, hate speech is defined as language targeted at a person or group based on their race, religion, ethnicity, nationality, gender, sexual orientation, gender identity, age, and disability. It incites hate, violence, and discrimination for both [dominant and (non-)dominant groups through the usage of {group characteristics}]. **Expectations:** Recognizes hate speech against all target types, including dominant groups. While we expect good generalization for incitement of hate, violence, and discrimination, it is unclear if group insult will work well. Slurs and stereotypes should be captured.

The DYNAMICALLY GENERATED HATE SPEECH (Vidgen et al., 2021) dataset is dynamically created with human-in-the-loop. The annotators are provided with extensive annotator guidelines with many annotation rounds. It consists of 41255 examples, of which 54.0% is hate speech and 46.0% is not.¹¹

DGHS Definition

(Pages 3-4, Section 3; Annotation Guidelines^a) “‘Hate’ is defined as “abusive speech targeting specific group characteristics, such as ethnic origin, religion, gender, or sexual orientation.” (Warner and Hirschberg, 2012)

3.1 Types of hate:

Derogation: Content which explicitly

attacks, demonizes, demeans or insults a group. This resembles similar definitions from Davidson et al. (2017), who define hate as content that is ‘derogatory’, Talat and Hovy (2016) who include ‘attacks’ in their definition, and Zampieri et al. (2019) who include ‘insults’.

Animosity: Content which expresses abuse against a group in an implicit or subtle manner. It is similar to the ‘implicit’ and ‘covert’ categories used in other taxonomies (Waseem et al., 2017; Vidgen and Yasseri (2020); Kumar et al. (2018)).

Threatening language: Content which expresses intention to, support for, or encourages inflicting harm on a group, or identified members of the group. This category is used in datasets by Hammer (2014), Golbeck et al. (2017) and Anzovino et al. (2018). Support for hateful entities Content which explicitly glorifies, justifies or supports hateful actions, events, organizations, tropes and individuals (collectively, ‘entities’).

Dehumanization: Content which ‘perceiv[es] or treat[s] people as less than human’ (Haslam and Stratemeyer, 2016). It often involves describing groups as leeches, cockroaches, insects, germs or rats (Mendelsohn et al., 2020).

3.2 Targets of hate: Hate can be targeted against any vulnerable, marginalized or discriminated-against group. We provided annotators with a non-exhaustive list of 29 identities to focus on (e.g., women, black people, Muslims, Jewish people and gay people), as well as a small number of intersectional variations (e.g., ‘Muslim women’). They are given in Appendix A. Some identities were considered out-of-scope for Hate, including men, white people, and heterosexuals.”

^a<https://github.com/bvidgen/Dynamically-Generated-Hate-Speech-Dataset/blob/main/Dynamically%20Generated%20Hate%20Dataset%20-%20annotation%20guidelines.pdf>

Decomposed: For this dataset, hate speech is defined as language targeted at a person or group based on religion, race, gender,

¹¹Note that this is the most balanced dataset out of all.

sexual orientation, nationality, and disability and insults the group or incites hate or violence for [especially non-dominant groups] through the usage of {group characteristics, slurs, and stereotypes}. **Expectations:** We expect good performance on all the target types in HateCheck. Dominant groups will not be seen as part of hate speech. Except for incitement of discrimination, which we do not know will be captured or not, group insult and incitement of hate and violence should have good performance. All types of explicit references should get good performance.

The HATEXPLAIN (Mathew et al., 2021) dataset comprises 20,148 examples and has classes for offensive language and undecided instances. The class distribution is 29.5% hateful, 27.2% offensive, 38.8% normal, and 4.5% undecided.

Target Groups	Categories
Race	African, Arabs, Asians, Caucasian, Hispanic
Religion	Buddhism, Christian, Hindu, Islam, Jewish
Gender	Men, Women
Sexual Orientation	Heterosexual, Gay
Miscellaneous	Indigenous, Refugee/Immigrant, None, Others

Table 2: (Page 3, Table 3) Target groups considered in HATEXPLAIN

Decomposed: For this dataset, hate speech is defined as language targeted at a person or group based on their race, religion, gender, or sexual orientation from both [dominant and non-dominant groups] through the usage of {group characteristics}. **Expectations:** We expect good performance on *race*, *religion*, *gender*, *nationality*, and *sexual orientation*, also for dominant groups. Other target types and aspects are unknown since they are unmentioned.

The FOUNTA (Founta et al., 2018) dataset consists of 4.97% hateful, 27.15% abusive, 14.03% spam, and 53.85% normal tweets.

FOUNTA Definition

(Page 5, Section Step 2: Exploratory Rounds) “**Hate Speech:** Language used to express hatred towards a targeted individual or group, or is intended to be derogatory, to humiliate, or to insult the members of the group, on the basis of attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender. (Davidson et al., 2017), (Badjatiya et al., 2017), (Warner and

Hirschberg, 2012), (Schmidt and Wiegand, 2017), (Djuric et al., 2015).”

Decomposed: For this dataset, hate speech is language targeted at a person or group based on their race, religion, ethnic origin, sexual orientation, disability, or gender etc. and insults a group or incites hate through the usage of {group characteristics}. **Expectations:** We expect good performance on all target types. We also expect the model to capture incitement of hate and group insult. The performance on dominance and the rest of the aspects is unknown.

C Results on Individual Datasets

In this section, we provide the macro F1 obtained on the respective validation and test sets of each dataset in Figure 8, for both fBERT and RoBERTa.

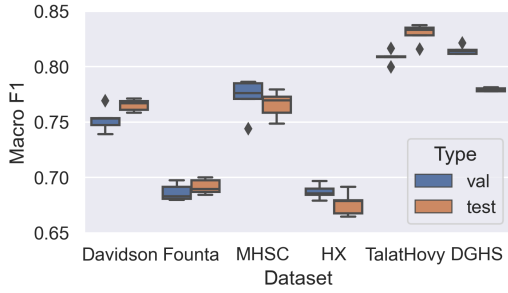
We observe that most datasets achieve average to good performance on their respective validation and test sets (Figure 8). Both fBERT (Figure 8a) and RoBERTa (Figure 8a) appear to yield similar results. Particularly, we see for **TalatHovy** and **DGHS** a high Macro F1, in the range of 0.8 to 0.825 with low variation in seeds, except for **DGHS** where one seed underperforms (0.35).¹² The **MHSC** dataset also achieves a higher Macro F1 (~ 0.775) but with slightly more variation across seeds. The rest of the datasets also have slightly more variation across seeds but lesser performance in comparison, yielding decent results (~ 0.675 – 0.75).

D Analysis of DGHS’ Composition

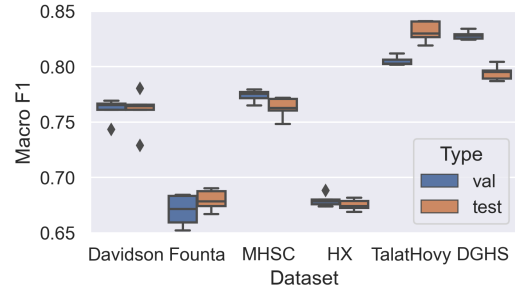
To identify the influence of DGHS’s dataset composition on its superior performance, we experiment with augmenting the training set size. In Table 3, we show the label distribution of the two different *DGHS* training sets we experiment with, using fBERT. **DGHS_{Small}**: a smaller training set size that approximates the size of the smallest dataset, the *TalatHovy* dataset. **DGHS_{Imbalanced}**: the imbalanced dataset, where we keep 10% of the instances hate and the rest of the 90% not hate, essentially also decreasing the size. This corresponds to many of the datasets where there are very few *hate speech* instances but many *non-hate speech* ones.

The results on the dataset’s respective test set, overall HateCheck, and cross-evaluation can be

¹²For clarity in the plot, we limit the y-axis.



(a) fBERT.



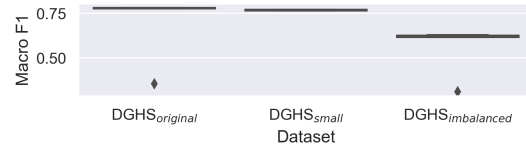
(b) RoBERTa.

Figure 8: Macro F1 scores on the respective validation and test set for all the *six* datasets.

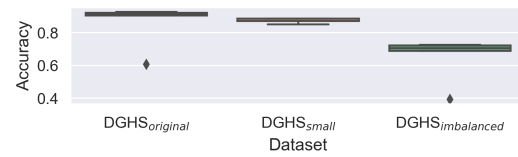
	#Hate	#Not Hate
DGHS _{Original}	17740	15184
DGHS _{Small}	7537	6463
DGHS _{Imbalanced}	1687	15184

Table 3: Data summary

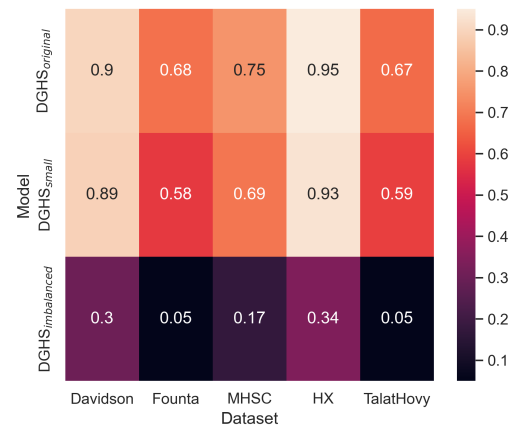
found in Figure 9. Sizing down the training set with DGHS_{Small} does not have as much impact on the results overall. The results are slightly lower than the results with the original training set. However, results of DGHS_{Imbalanced} clearly take a hit when cross-evaluating, indicating that it is essential to have a large amount of *hate speech* samples for good cross-evaluation and performance in general.



(a) Results on DGHS's test set.



(b) Results on HateCheck.



(c) Cross-evaluation results.

Figure 9: Results of testing different data compositions for DGHS's dataset.