# *What Did You Learn To Hate?* A Topic-Oriented Analysis of Generalization in Hate Speech Detection

**Tom Bourgeade[1], Patricia Chiril[3], Farah Benamara[1,2], Véronique Moriceau[1]**
(1) IRIT, Université de Toulouse, *CNRS, Toulouse INP, UT3*, Toulouse, France
{firstname.lastname}@irit.fr
(2) IPAL, CNRS-NUS-ASTAR, Singapore
(3) University of Chicago, Chicago, IL, USA
pchiril@uchicago.edu

## Abstract

**Warning:** *This paper includes messages that may contain instances of vulgarity, degrading terms, or hate speech, which may be offensive or upsetting to some readers.*

Hate speech has unfortunately become a significant phenomenon on social media platforms, and it can cover various topics (*misogyny*, *sexism*, *racism*, *xenophobia*, etc.) and targets (e.g., *black people*, *women*). Various hate speech detection datasets have been proposed, some annotated for specific topics, and others for hateful speech in general. In either case, they often employ different annotation guidelines, which can lead to inconsistencies, even in datasets focusing on the same topics. This can cause issues in models trying to generalize across more data and more topics in order to improve detection accuracy. In this paper, we propose, for the first time, a topic-oriented approach to study generalization across popular hate speech datasets. We first perform a comparative analysis of the performances of Transformer-based models in capturing *topic-generic* and *topic-specific* knowledge when trained on different datasets. We then propose a novel, simple yet effective approach to study more precisely which topics are best captured in implicit manifestations of hate, showing that selecting combinations of datasets with better out-of-domain topical coverage improves the reliability of automatic hate speech detection.

## 1 Introduction

On social media and other online communication platforms, hate speech (HS hereafter) is found in many forms, from textual harassment to threats, targeting an individual or group (e.g., black people, women), based on some characteristics, such as race, color, ethnicity, gender, sexual orientation, nationality, religion, etc, which we refer to as *topics*. While the broad nature of these topics of HS is generally understood (Erjavec and Kovačič, 2012), determining whether a social media post is a manifestation of HS (and if so, to which topic(s) it belongs to) is not a trivial task for humans and automated machine-learning systems. Indeed, the latter often require large quantities of annotated data, which in the case of HS, is made difficult due to (1) the absence of a comprehensive definition of these topics; but also, (2) the potential internal biases and subjectivity present in annotators and/or annotation guidelines employed to construct annotated datasets (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018; Vidgen et al., 2019; Fortuna et al., 2020).

To palliate these issues, cross-dataset evaluation has become an active line of research aiming at studying the generalization capabilities of HS detection systems to unseen data during training (Talat et al., 2018; Ludwig et al., 2022; Toraman et al., 2022). As data collection and annotation is an expensive and time-consuming process, current approaches use *mixtures* of existing HS detection datasets to study generalization across different social media platforms (Swamy et al., 2019; Salminen et al., 2020; Nejadgholi and Kiritchenko, 2020) or across different manifestations of HS, often relying on one-to-many (train on one dataset/mixture, test on others) experimental settings (Fortuna et al., 2021; Talat et al., 2018; Chiril et al., 2022; Talat et al., 2018; Karan and Šnajder, 2018). In most of these, a unification scheme is proposed to adapt the original datasets' potentially fine-grained annotations to a set of binary labels, which can often fail to take into account the heterogeneity inherently present across different encodings of HS (Vidgen et al., 2019).

In addition, some datasets intentionally focus only on some specific kinds of manifestations of HS covering for example gender (e.g., misogyny, sexism), ethnicity, religion or race (e.g., xenophobia, anti-immigrants/refugees HS), i.e., they are *topic-specific*, whereas others attempt to cast a



*Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3495–3508

wider net without specific sets of topics in mind, i.e., they are *topic-generic*. In either case, variations will always exist, either due to the reasons mentioned above, or due to shifts in the content found on social media through time. All of these may cause significant differences in the topics of HS that classification models might learn to recognize, which ultimately might make them less reliable, particularly in out-domain contexts (Yin and Zubiaga, 2021).

Acknowledging these issues and being aware of the noise that will invariably be introduced by mixing together different datasets, we propose for the first time to empirically study what modern Transformer architectures can effectively learn to generalize from existing datasets, in a unified setting, with a *focus on their topical nature*. More precisely: *Do models learn similarly from topic-generic datasets as from mixtures of topic-specific datasets? Does this acquired knowledge generalize to implicit expressions of HS, and if so, which finer-grained HS topics are learned by these models?* Our contributions are:

(1) An in-depth analysis of the generalizability of generic HS datasets, in which we show how using mixtures of those could be effective to attain better generalization across more topics.

(2) A similar analysis for *topic-specific* datasets, for which we show successful generalization in-domain, especially when using models fine-tuned on mixtures of such corpora.

(3) A novel, simple yet effective approach to study which finer-grained topics are best captured when dealing with *implicit* expressions of HS. We show that selecting mixtures of datasets with a better topical coverage can improve the reliability of models for out-of-domain applications.

## 2 Related Work

A number of previous works have studied generalizability of HS detection datasets and models, with different focuses: for example, Fortuna et al. (2020) and Fortuna et al. (2021) have analyzed the compatibility of many HS datasets (including some used in this study), both in terms of their properties (origin of data, annotated phenomena, class definitions, etc.) and empirically with intra- and inter-dataset generalization experiments. They conclude that model choice, intra-dataset performance, and the type of phenomenon being classified, are the most important factor that determine generalization. In particular, they conclude that phenomena like *toxicity*, *abuse*, or *offensiveness* that are often lexicalized are easier to generalize than hate speech, something which was also confirmed and discussed in other works (Nejadgholi and Kiritchenko, 2020; Yin and Zubiaga, 2021). In this paper, we are interested in the topic-oriented nature of HS specifically, and thus discard these former classes from our study.

Alternative solutions have recently been proposed to handle unseen topics by building new target-oriented datasets from scratch, such as the `HateXplain` dataset (Ludwig et al., 2022), with an additional labeling of the target topics of hate (Race, Religion, and Origin). While such linguistic resources are valuable for the research community, we believe existing larger datasets could be successfully exploited for generalization, as mixtures of datasets, as shown for example by Fortuna et al. (2018), Salminen et al. (2020), and Chiril et al. (2022), or, in a different fashion, by multitask HS detection systems or domain adaptation techniques (Talat et al., 2018; Kapil and Ekbal, 2020; Safi Samghabadi et al., 2020). In this paper, we continue this line of research by proposing for the first time, as far as we know, models able to generalize across topic generic vs. specific datasets as well as predict fine-grained manifestations in implicit hate messages. This is particularity challenging as these manifestations are more difficult to generalize due to their limited lexical features (ElSherief et al., 2021; Qian et al., 2019).

To this end, we rely on Transformer models like BERT that have been shown to be able to generalize better overall than previous architectures (Swamy et al., 2019). In particular, we experiment with both existing BERT-like models, some pre-adapted to the domain of HS, as well as T5 (Raffel et al., 2020), a text-to-text architecture that has never been used in a generalizability study.

## 3 Datasets

We experiment with six popular and freely available English tweets corpora (or English subsets thereof) from previous studies. We first present the datasets and how they were used in a common experimental setting. We then provide some discussions on the compatibility of these datasets, and how this may impact the generalization performances in our experiments.

### 3.1 Topic Generic vs. Topic Specific Datasets

The first two datasets are *topic-generic* while the other four are *topic-specific*, as follows.

**Davidson** (Davidson et al., 2017). It contains English tweets annotated into *hate speech*, *offensive*, and *neither*, with the intent of helping to distinguish between messages simply containing offensive terms, from those actually manifesting HS.

**Founta.** We make use of the dataset released by (Kallumadi et al., 2020) which is an updated version of the dataset initially proposed by Founta et al. (2018) and annotated for four types mutually exclusive of abusive behaviors: *abusive*, *hateful*, *spam* and *normal*.

**IberEval** (Fersini et al., 2018b) and **Evalita** (Fersini et al., 2018a) are part of the Automatic Misogyny Identification (AMI) shared task which aims at identifying tweets that convey hate or prejudice against women while categorizing different forms of misogynous behavior. We only use the main binary layer of annotation (i.e., presence vs. absence of misogyny).

**HatEval** (Basile et al., 2019) also known as SemEval-2019 Task 5, is a topic-specific HS detection dataset with tweets targeting *immigrants* and *women*, and annotated with three different binary layers of annotation: hateful/non-hateful, targeting a group/individual, aggressive/non-aggressive. Note that most of the tweets that target women in this dataset were derived from the AMI corpora (IberEval and Evalita).

**Waseem** (Talat and Hovy, 2016). It contains tweets targeting gender minorities, instances of racism, and tweets that were judged to be neither sexist nor racist.

We frame all the datasets used here as binary HS classification tasks, with the labels *"hateful"* (also referred to as the positive class) for instances containing some manifestation of hate speech, and *"normal"* for those containing none. Hence, for Founta and Davidson we filter out retweets/duplicates (keeping only the source tweets and their annotations) and only keep the *hateful/hate speech* and *normal* classes, and similarly, for the different topic-specific datasets, we unify the specific hate classes (*misogyny*, *sexism*, *racism*) with *hateful*, and the respective negative classes with *normal*.

To allow a more granular and topic-level analysis of results, we split all multi-targets topic-specific datasets into separate training, validation, and testing sets, according to the topics. Therefore, we split HatEval into its two topics subsets (i.e., $\texttt{HatEval}_{\texttt{women}}$ and $\texttt{HatEval}_{\texttt{immigrants}}$) and Waseem into $\texttt{Waseem}_{\texttt{sexism}}$ and $\texttt{Waseem}_{\texttt{racism}}$. However, in this last dataset, because only one negative class is provided, and corresponds to both the absence of sexism and racism, we choose to duplicate it and use it as the negative class for both subsets.

When not explicitly provided, we use a 75%-20%-5% train-test-validation split ratio. Table 1 further details how datasets have been mixed to train our models.

### 3.2 Issues with Dataset Compatibility

Because we also manipulate mixtures of these corpora, the effective hateful and non-hateful classes will contain instances annotated within different contexts and labelling guidelines. For example, the Evalita and IberEval datasets are annotated only for the presence of misogyny, and not other manifestations of HS. This makes their negative (i.e., non-misogyny) class inconsistent with, say, the negative class from a topic-generic dataset, which is not ideal.

Aside from re-annotating all these datasets under a unified and consistent annotation schema, the issues that may arise as part of these simplifications cannot be circumvented, and should thus be taken into account as a fixed parameter in our experiments. These issues have been broadly acknowledged in the relevant literature (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018; van Aken et al., 2018). As noted by Malmasi and Zampieri (2018) and Poletto et al. (2021), the distinctions between offense, abuse, and HS, are not always clear-cut, which can cause issues in generalization experiments, due to the former's in theory more lexical nature (Vidgen et al., 2019; Fortuna et al., 2020).

Furthermore, as noted by Madukwe et al. (2020), even when using relatively similar definitions for these phenomena, a number of other parameters may affect the compatibility, consistency, and comparability between HS detection datasets, from biases introduced in the annotations, differences in preprocessing steps (e.g., anonymization, emojis, URLs, etc.), to issues of class balance and unspecified train-test-validation splits.

We are aware that mixing the datasets may effectively introduce various kinds of noise in the training and evaluation data, but we consider it

| Topic | Dataset | Abbrev. | Total Size | Pos. Size | Ratio | Train | Val. | Test |
|---|---|---|---|---|---|---|---|---|
| Topic-Generic | Davidson | Gene Davi | 5590 | 1430 | 25.58% | 4293 | 179 | 1118 |
| | Founta | Gene Fnt | 57355 | 4119 | 7.18% | 44048 | 1836 | 11471 |
| Topic-Specific (Gender) | Evalita | Gndr Evit | 5000 | 2245 | 44.90% | 3839 | 161 | 1000 |
| | $\text{HatEval}_{\text{women}}$ | Gndr HatE | 6472 | 2845 | 43.96% | 4500 | 500 | 1472 |
| | IberEval | Gndr Iber | 3977 | 1851 | 46.54% | 3120 | 131 | 726 |
| | $\text{Waseem}_{\text{sexism}}$ | Gndr Wasm | 14531 | 3216 | 22.13% | 11159 | 465 | 2907 |
| Topic-Specific (Race) | $\text{HatEval}_{\text{immigrants}}$ | Race HatE | 6499 | 2617 | 40.27% | 4500 | 500 | 1499 |
| | $\text{Waseem}_{\text{racism}}$ | Race Wasm | 13272 | 1957 | 14.75% | 10192 | 425 | 2655 |
| Mixtures | Topic-Generic Mixture | Gene Mixt | Davidson + Founta | | | | | |
| | Gender Topic Mixture | Gndr Mixt | $\text{Evalita} + \text{HatEval}_{\text{women}} + \text{IberEval} + \text{Waseem}_{\text{sexism}}$ | | | | | |
| | Race Topic Mixture | Race Mixt | $\text{HatEval}_{\text{immigrants}} + \text{Waseem}_{\text{racism}}$ | | | | | |
| | Topic-Specific Mixture | Spec Mixt | Topic Gender Mixture + Topic Race Mixture | | | | | |

Table 1: Overview of the datasets used to train models in this study.

to be part of the experimental settings, especially since these kinds of issues would likely as well be encountered in end-user applications dealing with raw unfiltered data.

## 4 Models

To study how different pre-trained models and architectures may differ in how they capture HS in various settings, we choose five pre-trained Transformers from the literature. Among them, the last three have been adapted to the domain of HS detection, either through pretraining or pre-finetuning, on data related directly or indirectly to HS detection. For the experiments, we relied on the Hugging Face transformers library (Wolf et al., 2020).

**RoBERTa-base** (Liu et al., 2019) is an optimized BERT-like (Devlin et al., 2019) encoder Transformer commonly used for various NLP classification tasks, including HS detection.

**T5-base** is the 220 million parameters pretrained variant of the T5 architecture, initially proposed by Raffel et al. (2020). It differs from BERT-like encoder Transformers in that it is a *text-to-text* model, for which classification tasks are reframed as text generation, with the output labels used in their textual form (in our case, *"normal"* and *"hateful"*), and as this particular variant was also pretrained on various supervised tasks (sentiment analysis, natural language inference, and question answering, etc.), a task prefix is traditionally prepended to the input text, which for our fine-tuning, we fix to *"hate speech: "*.

**fBERT** (Sarkar et al., 2021) and **HateBERT** (Caselli et al., 2021) are two models derived from BERT, retrained with a Masked Language Modelling (MLM) objective on over 1.4 million social media offensive posts from the SOLID dataset, and Reddit Abusive Language English dataset (RAL-E), respectively.

**ToxDectRoBERTa** was proposed by Zhou et al. (2021), and is a RoBERTa-large model, finetuned on Founta, with the *hateful* and *abusive* classes merged into a single *toxic* class. The authors use the methods proposed by Clark et al. (2019) and Swayamdipta et al. (2020) to attempt to avoid dataset bias issues, such as spurious correlations between particular lexical and dialectical markers (such as those found in African American English) with the toxic class.

## 5 Cross-Topics Generalization

We first study the differences in generalizability between topic-generic and topic-specific datasets and their potential mixtures, in a cross-dataset/mixture setting: in each individual experiment, one of the previously described models is trained on one dataset/mixture and tested on all individual test sets. For technical details on the experimental parameters used in this study, see Appendix A.

To measure improvements or deteriorations in performances in the generalization experiments, Table 2 presents an intra-dataset evaluation, in which models are simply trained and tested on the same datasets, for the sake of comparison. In the remainder of this section, we report macro F1 scores on all test sets (see Table 1 for abbreviations), with the best scores for each test set highlighted in **bold**. Scores are gradient-colored for legibility.

| Dataset \ Model | T5-base | RoBERTa-base | ToxDectRoBERTa | fBERT | HateBERT |
|---|---|---|---|---|---|
| Gene Davi | 92.70 | 93.30 | **93.78** | 92.46 | 93.14 |
| Gene Fnt | 81.08 | 80.84 | **85.92** | 80.98 | 79.41 |
| Gndr Evit | 67.00 | 69.58 | 67.37 | **73.12** | 71.02 |
| Gndr HatE | 46.29 | 56.88 | 51.65 | **64.28** | 59.11 |
| Gndr Iber | 84.86 | **88.76** | 83.89 | 84.55 | 86.41 |
| Gndr Wasm | 85.81 | **87.11** | 86.27 | 86.28 | 86.05 |
| Race HatE | 38.73 | 38.24 | 40.75 | **41.72** | 38.53 |
| Race Wasm | 87.03 | 86.27 | 86.29 | **87.53** | 86.37 |

Table 2: Results of intra-dataset evaluations (training and testing on the same dataset's train and test sets).

### 5.1 Learning from Topic-Generic Datasets

We investigate here how well knowledge can transfer from topic-generic to topic-specific datasets. We thus train each of the five chosen models on the Davidson (Gene Davi) and Founta (Gene Fnta) train sets, as well as on a mixture of the two (Gene Mixt), then evaluate those models on every individual test set.

Table 3 presents the results of these evaluations, in terms of macro F1 score. Observing the results, firstly, we can notice that Davidson and Founta generalize relatively well to each other, with relatively small deteriorations in F1 scores ($\sim -10$ F1) compared to the intra-dataset models (see Table 2), in the favor of Davidson $\rightarrow$ Founta for our 5 models. When used as a mixture of datasets (Gene Mixt), we observe very small improvements for 3 of our five models and very small deteriorations ($\sim \pm 1$ F1) for the rest (RoBERTa-base and ToxDectRoBERTa), which would tend to indicate a good compatibility between these two datasets, or at the very least, some high overlap in the topic-generic HS knowledge extracted by Transformer models across these two datasets.

Looking then at the evaluation results obtained on the topic-specific test sets, we can make a number of observations: first, we note quite significant deteriorations compared to the intra-dataset models for all datasets, except for both HatEval subsets, in particular $\texttt{HatEval}_{\texttt{immigrants}}$, for which all three topic-generic datasets/mixture yield significant improvements (from $\sim +3$ F1 up to $\sim +27$ F1). We believe this to be due to a significant distribution shift between the HatEval train and test set, which would explain why the models perform so poorly in the intra-dataset setting, while using different training sets appears to significantly improve performance.

For the other datasets, the least significant deteriorations can be found for Evalita, while the greatest ones are found for $\texttt{Waseem}_{\texttt{sexism}}$, which may indicate an overall low overlap in the types of manifestations of HS found in the topic-generic and topic-specific datasets explored here. However, in most cases, the use of the topic-generic mixture (Gene Mixt) appears to be beneficial, in that it tends to attenuate the worst deteriorations found in models trained on Davidson or Founta individually. This may indicate that using mixtures of topic-generic training datasets may be beneficial when trying to detect HS instances where the topic is not necessarily known.

### 5.2 Learning from Topic-Specific Datasets

We similarly trained HS classifiers, this time on the remaining 6 topic-specific datasets, separated into the Race and Gender topics (see Table 1). Table 4 presents the results (in a form similar to Table 3), however, for space reasons and as they are the focus of our study, we only present the results for the gender and race topics mixtures (Gndr Mixt, and Race Mixt respectively), as well as the mixture of both (Spec Mixt). See Table A in Appendices for more detailed results.

Observing the results, we can first see that, similarly to the previous experiments, training on these three mixtures of topic-specific datasets does not seem to generalize too well back onto the topic-generic datasets, with even greater deteriorations in F1 scores compared to the intra-dataset setting (cf. Table 2). This is the most pronounced for Race Mixt, which is the smaller of the two one-topic mixtures. Even more prominently than for topic-generic mixture discussed previously (Gene Mixt), we find that the topic-specific mixture (Spec Mixt), which combines both the Gender and Race topics, yields significantly lesser deteriorations than ei-

| Model | T5-base | | | RoBERTa-base | | | ToxDectRoBERTa | | | fBERT | | | HateBERT | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Test \ Train | Gene Davi | Gene Fnta | Gene Mixt | Gene Davi | Gene Fnta | Gene Mixt | Gene Davi | Gene Fnta | Gene Mixt | Gene Davi | Gene Fnta | Gene Mixt | Gene Davi | Gene Fnta | Gene Mixt |
| Gene Davi | 92.70 | 82.64 | **93.92** | 93.30 | 82.50 | 91.60 | 93.78 | 84.73 | 90.96 | 92.46 | 85.59 | 92.92 | 93.14 | 85.43 | 93.42 |
| Gene Fnt | 73.02 | 81.08 | 81.41 | 74.50 | 80.84 | 80.49 | 81.42 | **85.92** | 84.70 | 76.47 | 80.98 | 81.37 | 73.02 | 79.41 | 80.32 |
| Gndr Evit | 63.29 | 58.07 | 61.79 | 60.68 | 58.51 | 56.97 | 62.10 | 54.25 | 60.20 | 61.44 | 57.00 | 62.16 | **63.92** | 56.74 | 59.99 |
| Gndr HatE | 42.21 | 42.90 | 54.60 | 46.39 | 55.17 | 49.45 | 40.30 | 39.81 | 52.35 | 38.25 | 49.75 | 45.83 | 44.46 | 55.76 | **56.31** |
| Gndr Iber | 57.76 | **74.22** | 59.48 | 60.50 | 62.96 | 68.89 | 64.14 | 73.00 | 67.83 | 60.34 | 71.43 | 63.74 | 60.12 | 67.11 | 65.24 |
| Gndr Wasm | 54.06 | 59.07 | 52.21 | 51.55 | 51.11 | 51.63 | 51.13 | 56.44 | 52.72 | 53.84 | **61.47** | 51.73 | 52.36 | 58.24 | 54.52 |
| Race HatE | 43.75 | 64.54 | 50.97 | 47.79 | 56.76 | 63.71 | 52.23 | **68.58** | 66.92 | 50.65 | 66.49 | 45.70 | 44.13 | 66.08 | 65.62 |
| Race Wasm | 56.66 | 72.73 | 73.69 | 56.89 | **75.52** | 70.72 | 69.92 | 73.47 | 73.45 | 70.16 | 72.33 | 70.01 | 63.01 | 72.81 | 73.28 |

Table 3: Results of learning from the two topic-generic datasets used here, Davidson (Gene Davi) and Founta (Gene Fnta), as well as their mixture (Gene Mixt). Scores are gradient-colored for legibility.

| Model | T5-base | | | RoBERTa-base | | | ToxDectRoBERTa | | | fBERT | | | HateBERT | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Test \ Train | Gndr Mixt | Race Mixt | Spec Mixt | Gndr Mixt | Race Mixt | Spec Mixt | Gndr Mixt | Race Mixt | Spec Mixt | Gndr Mixt | Race Mixt | Spec Mixt | Gndr Mixt | Race Mixt | Spec Mixt |
| Gene Davi | 61.72 | 50.23 | 68.33 | 62.20 | 47.75 | 68.22 | 65.04 | 62.26 | 67.13 | 68.38 | 51.78 | **72.60** | 62.67 | 56.10 | 68.75 |
| Gene Fnt | 58.61 | 57.44 | 62.09 | 57.60 | 54.14 | 61.17 | 62.92 | 60.92 | 63.24 | 60.46 | 55.14 | 59.93 | 57.91 | 58.11 | **63.36** |
| Gndr Evit | 82.90 | 36.42 | 84.78 | 86.33 | 35.92 | 86.89 | 85.40 | 42.15 | 85.58 | **88.59** | 38.58 | 87.37 | 87.96 | 38.95 | 86.80 |
| Gndr HatE | 47.83 | 36.82 | 54.58 | 52.69 | 37.02 | 53.65 | 50.80 | 48.80 | **58.53** | 54.80 | 36.71 | 58.44 | 55.33 | 42.57 | 49.06 |
| Gndr Iber | **93.31** | 40.84 | 93.17 | 92.12 | 38.66 | 92.84 | 92.32 | 40.84 | 92.30 | 92.55 | 39.42 | 92.55 | 92.85 | 39.42 | 92.86 |
| Gndr Wasm | 85.14 | 44.92 | 87.85 | 85.66 | 43.96 | 88.53 | 85.72 | 44.76 | 87.39 | 86.92 | 43.59 | 87.93 | 84.85 | 44.28 | **88.87** |
| Race HatE | 42.98 | 40.49 | **44.29** | 37.69 | 40.44 | 38.06 | 38.59 | 41.23 | 44.03 | 39.47 | 43.68 | 40.40 | 37.56 | 37.01 | 35.80 |
| Race Wasm | 46.49 | 86.76 | 85.39 | 46.30 | 85.90 | 88.00 | 48.12 | 87.05 | 85.46 | 48.58 | 86.25 | **89.17** | 46.66 | 86.42 | 88.37 |

Table 4: Results of learning from the mixtures of the gender topic-specific (Gndr Mixt) datasets (Evalita, $\text{HateEval}_{\text{women}}$, IberEval, and $\text{Waseem}_{\text{sexism}}$), race topic-specific (Race Mixt) datasets ($\text{HateEval}_{\text{immigrants}}$, and $\text{Waseem}_{\text{racism}}$), and a mixture of both topics (Spec Mixt).

ther of the two individually: this intuitively makes sense, as both separate topics can help cover different subsets of the topic-generic datasets, and should also help in learning manifestations of hate which exist at the intersection of both topics.

Looking at the scores obtained on the topic-specific test sets, we can observe that, unlike in the topic-generic generalization experiments, both the one-topic mixtures and Spec Mixt appear to yield improvements over the intra-dataset models, for most of the models. For Gndr Mixt and Race Mixt, the improvements are, as expected, mostly found for the test sets of the datasets making up the mixtures, but not always: for example, with the T5-base model, training on Gndr Mixt appears to yield improvements on $\text{Waseem}_{\text{racism}}$ ($\sim$ +4 F1). This seems to indicate a better ability to generalize in-domain from mixtures of topic-specific datasets, at least to other topic-specific datasets, but more experiments with more topic-specific datasets, covering a wider range of topics, would be necessary to determine whether this may also apply to topic-generic datasets.

## 6 Finer-Grained Topics Analysis

To better understand which manifestations of HS can be generalized to out-of-domain data, we then perform a finer-grained analysis of the topics learned by our models, by relying on an implicit hate speech dataset, for which fine-grained target annotations are available.

### 6.1 Implicit Hate Speech Dataset

IMPLICIT HATE CORPUS (which we refer to as ElSherief, for brevity) corresponds to the dataset proposed by ElSherief et al. (2021), which consists of 21,480 English tweets annotated for (in a first stage) the presence of implicit or explicit HS (or neither), as mutually exclusive classes. Further, for each of the tweets containing implicit HS, two an-

notators supplied the targeted demographic groups and the implied statement of the hateful messages, both as free-form texts. In our experiments, we rely on the 6,196 implicit hate instances and their target annotations, which we have manually grouped into different sets of finer-grained topics of HS (whose sizes are in Table 5): *Islam*, *Black People*, *Immigrants / Refugees*, *Beliefs / Religion*, *Gender*, *LGBTQIA+*, *Unspecified Minorities*, and *Nationality*.

### 6.2 Results

To better show the observations that can be mainly linked to training dataset/mixture selection, we choose to present, in Table 5, the results of the best performing non-domain adapted architecture used here, T5-base. To this end, we rely on the previously mentioned trained T5-base models to obtain binary HS predictions on the entire processed `ElSherief` dataset, and compute the accuracy (as the tested instances are by definition all from the positive class) of the models for each of the finer-grained topic-specific subsets described previously.

The first three results columns show the topics learned by the models trained on the topic-generic datasets, `Davidson`, `Founta`, and their mixture (`Gene Mixt`). We observe that the model trained on `Founta` alone yields the highest accuracy scores over a number of topics, except for the *Immigrants/Refugees*, *Unspecified Minorities*, and *Gender* topics. Due to the latter's small size, and higher degree of co-occurring kinds of HS manifestations (see Section 6.3), accuracies appear to be relatively lower across the board, even for the gender topic-specific datasets. Still, topic-generic datasets, and their mixture, display fairly decent generalization on average for most fine-grained topics analyzed here, and represent a promising avenue for future research aimed at constructing effective out-of-domain generalization mixtures of datasets.

The next group of three columns show the differences in variety of finer-grained topics, between the gender and the race topic, and this in spite of fewer number of instances in the Race topic mixture compared to Gender one. Training on both topics combined (`Spec Mixt`) only yields improvement in the *Gender* and *LGBTQIA* topics, while all the other accuracies drop. This could be a sign that the dual-topic mixture allows for the learning of intersectional manifestations of HS between these topics. Indeed, in the *Gender* group, we find that 22 out of the 58 instances also belong to one or more other groups, and among those, 16 are in the *Muslims*, *Immigrants / Refugees*, or both groups (10 for *Immigrants / Refugees* and/or *Black People* in *LGBTQIA*). As such, it appears that mixtures of topic-specific datasets may prove useful in the generalization of intersectional HS detection, though more research would be necessary to verify this for other topics, and with more data.

The next two columns demonstrate how two datasets annotated for similar topics can wildly vary in which fine-grained topics they allow models to capture. Due to the significant differences in data collection, annotation, and cultural/temporal contexts, the `HatEval` and `Waseem` datasets result in very different learned distributions of fine-grained topics, with the latter almost exclusively covering the *Muslims* topics, whereas the former yields a more varied, and overall more performant on the `ElSherief` dataset, out-of-domain generalization.

The final group of three columns further illustrates this point: `Waseem` yields significantly lower performance on the *Gender* topic, and, if removed from the Gender topic mixture (`Gndr Mixt`), improves its overall accuracy. While removing the entirety of a dataset might not be ideal, this shows the importance of data selection for both in and out-of-domain generalization, and we believe this kind of analysis could be used to guide data selection, for the purpose of constructing such mixtures of datasets.

### 6.3 Error Analysis

To further investigate how these different models behave on those out-of-domain instances, we perform and present an error analysis on a selection of instances[1] from some of the different finer-grained topics.

We find that misclassifications are often due to several factors, among which implicit expressions of HS such as stereotypes, irony, or humor (as in (1)), or the presence of multiple topics of HS used in potentially ambiguous ways (2): is this instance (ironically) Islamophobic, sexist, or both?

(1) *what pronoun should we use to refer to people who are transitioning into martians ?*

(2) *women are intellectually deficient, islam says*

[1]These are examples taken from the ElSherief public dataset which provides texts of the tweets (because of Twitter otherwise having deleted a lot of them).

| Model | | T5-base | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Train / Test | Size | Gene Davi | Gene Fnta | Gene Mixt | Gndr Mixt | Race Mixt | Spec Mixt | Race HatE | Race Wasm | Gndr Iber | Gndr Wasm | Gndr Mixt\Wasm | All | *ES* |
| **Muslims** | 801 | 35.58 | **81.90** | 56.67 | 1.13 | 63.17 | 41.57 | 70.04 | 39.95 | 2.50 | 0.37 | 5.36 | 52.06 | 94.51 |
| **Black People** | 1391 | 78.43 | **78.58** | 54.71 | 2.80 | 34.94 | 23.22 | 52.41 | 2.44 | 3.16 | 0.93 | 9.78 | 50.75 | 88.79 |
| **Immigrants / Refugees** | 1386 | 24.82 | 47.40 | 26.19 | 2.45 | 53.03 | 42.14 | **66.52** | 1.30 | 2.45 | 0.51 | 9.95 | 50.36 | 96.90 |
| **Beliefs / Religion** | 652 | 63.03 | **76.99** | 53.53 | 1.68 | 35.74 | 28.53 | 51.69 | 6.74 | 2.76 | 0.16 | 6.44 | 53.99 | 93.25 |
| **Gender** | 58 | 31.03 | 41.38 | 24.14 | 32.76 | 17.24 | 46.55 | 25.86 | 5.17 | 44.83 | 12.07 | 46.55 | **48.28** | 96.55 |
| **LGBTQIA** | 87 | 43.68 | **63.22** | 33.33 | 3.45 | 12.64 | 21.84 | 37.93 | 0.00 | 12.64 | 2.30 | 14.94 | 32.18 | 88.51 |
| **Unsp. Minorities** | 1147 | **79.42** | 66.61 | 40.45 | 1.22 | 23.45 | 12.55 | 44.90 | 2.01 | 1.48 | 0.17 | 6.02 | 38.10 | 85.44 |
| **Nationality** | 203 | 50.24 | **67.98** | 44.83 | 1.97 | 34.97 | 24.14 | 56.16 | 3.45 | 4.43 | 0.49 | 7.39 | 40.39 | 84.24 |

Table 5: Accuracy scores obtained by the T5-base models trained on the datasets and mixtures of datasets (columns), evaluated on each of finer-grained topic of HS grouped from the ElSherief dataset (rows), with the number of instances of each topic reported in the Size column. The best accuracies for each topic are highlighted in **bold**, and the scores are gradient-colored for legibility. `Gndr Mixt\Wasm` corresponds to the Gender topic mixture (`Gndr Mixt`), from which $\text{Waseem}_{\text{sexism}}$ has been removed. **All** corresponds to the mixture of all 8 individual datasets (see Table 1). ***ES*** corresponds to baseline accuracy scores obtained from training a similar model on the ElSherief dataset itself, for comparison.

*in sahih bukhari 1:6:301. disagree? you are islamophobe.*

Both (3) and (4) were misclassified due to the lack of context: neither of the two implied statements provided by the annotators (i.e., *Women are dumb*/*Muslims are repugnant* and *kill gays*/*immigrants should be shot like rats*) can be reliably automatically inferred without more information.

(3) *oh honey. justoh, honey. bless your heart.*

(4) *maybe time to oil up the deer rifles, the ones with the hd night scopes to keep the "vermin" population down.*

## 7 Conclusion and Perspectives

In this work, we have proposed a set of topic-oriented analyses of the generalizability of HS datasets. We have shown how *topic-generic* and *topic-specific* datasets yield different degrees and nature of generalization, both when used individually, or as mixtures of datasets. With the former, we found that, while not very successful at generalizing in-domain to topic-specific datasets, the use of mixtures allows smoothing out individual weaknesses. With topic-specific datasets, we found that generalization is possible, both for single-topic and multi-topic mixtures. Through a finer-grain out-of-domain generalization analysis, we showed how a priori somewhat similar datasets can vary wildly in the forms of *implicit* HS that can be learned from them. Implicit expressions of hate, with few lexical features, are more difficult to generalize, as models can struggle to capture underlying hateful intents in messages. Notably, one barrier to understanding more implicit manifestations of hate is the lack of context for individual social media posts: more conversational datasets could represent an interesting avenue of research in that regard. These are all important considerations for the purpose of constructing more reliable automatic HS detection systems, intended to function on raw, potentially noisy, and never-seen-before data. In particular, we found that topic-generic datasets like `Founta` (Founta et al., 2018) appear promising for future research on generalization-optimized mixtures of datasets. Finally, concerning HS detection architectures, we found that, for BERT-like models, pre-domain adapted variants generalize slightly better than a more generic RoBERTa classifier, but yield similar results to a generic text-to-text T5 architecture, which seems promising for future research in HS detection.

In future work, we will explore the use of these approaches to guide data selection, by for example employing the method proposed by Swayamdipta et al. (2020), in order to construct mixtures of HS datasets better suited to out-of-domain generalization. Alternatively, Active Learning methods have been shown to be successful in NLP (Ein-Dor

et al., 2020), and could be used to palliate annotated scarcity in HS detection: by using topic-oriented analysis methods to detect difficult-to-predict topics, better targeted additional annotations could be acquired, to help improve the topical coverage of automated systems.

## Limitations

In this work, we acknowledge a number of issues with the compatibility of HS datasets (cf. Sections 1, 2, and 3.2): namely, the phenomena annotated in these datasets, even those labelled using similar or equal terms, will, in theory and in practice, represent wildly different classes. While, unlike some other previous works (Fortuna et al., 2020, 2021), we do not merge initially distinct labels in datasets, but instead only keep whole classes that correspond to hate speech (excluding other forms of abuse, like *toxicity*, *offensiveness*, etc.), using them in a unified binarized setting is still bound to introduce significant amounts of noise in the training data, with regards to the HS detection task. Most of these issues are well documented, however, as complete re-annotation of all relevant data would be a prohibitively expensive enterprise, we believe there is value in exploring alternative solutions that may enable generalization despite these problems, using existing annotated datasets as they are currently available. Additionally, the analyses presented in this work are by no means comprehensive, both in quantity and variety of datasets and models experimented with. For example, some of the fined-grain topic groups found in the `ElSherief` dataset are not large enough to draw strong conclusions from (namely, *Gender* and *LGBTQIA*): supplementing these less represented topics with more data would enable better insights into generalizability for these kinds of HS manifestations.

## Ethics Statement

The data that was used for conducting the experiments is composed of text from the public domain taken from datasets publicly available to the research community. These corpora also conform to the Twitter Developer Agreement and Policy that allows unlimited distribution of the numeric identification number of each tweet. The desire to combat online HS and prevent the widespreading of stereotypes cannot be done without automatic moderation tools, at the risk of increasing cases of algorithmic discrimination. However, the deployment of such algorithms should be done with care, as algorithmic discrimination results from the introduction of biases at the time of the design of the system. These biases consist in the transposition of general (often stereotyped) or statistical observations into systematic algorithmic conditions.

## Acknowledgements

This work has been carried out in the framework of the STERHEOTYPES project funded by the Compagnia San Paolo 'Challenges for Europe'. The research of Farah Benamara is also partially supported by DesCartes: The National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program.

## References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for Abusive Language Detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Patricia Chiril, Endang Wahyu Pamungkas, Farah Benamara, Véronique Moriceau, and Viviana Patti. 2022. Emotionally Informed Hate Speech Detection: A Multi-target Perspective. *Cognitive Computation*, 14(1):322–352.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4069–4082, Hong Kong, China. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

Deep Bidirectional Transformers for Language Understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online. Association for Computational Linguistics.

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

Karmen Erjavec and Melita Poler Kovačič. 2012. "You Don't Understand, This is a New War!" Analysis of Hate Speech in News Web Sites' Comments. *Mass Communication and Society*, 15(6):899–920.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018a. Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018b. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 214–228. CEUR-WS.org.

Paula Fortuna, Ilaria Bonavita, and Sérgio Nunes. 2018. Merging datasets for hate speech classification in Italian. In *EVALITA@ CLiC-it*.

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.

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 Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Surya Kallumadi, Srijan Kumar, and Diyi Yang. 2020. ICWSM Data Challenge. In *In Proceedings of the Fourteenth International Conference on Web and Social Media (ICWSM)*. AAAI Organization.

Prashant Kapil and Asif Ekbal. 2020. A deep neural network based multi-task learning approach to hate speech detection. *Knowledge-Based Systems*, 210:106458.

Mladen Karan and Jan Šnajder. 2018. Cross-domain detection of abusive language online. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137, Brussels, Belgium. 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:1907.11692 [cs]*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. *arXiv:1711.05101 [cs, math]*.

Florian Ludwig, Klara Dolos, Torsten Zesch, and Eleanor Hobley. 2022. Improving generalization of hate speech detection systems to novel target groups via domain adaptation. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 29–39, Seattle, Washington (Hybrid). Association for Computational Linguistics.

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.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2):187–202.

Isar Nejadgholi and Svetlana Kiritchenko. 2020. On Cross-Dataset Generalization in Automatic Detection

of Online Abuse. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 173–183, Online. Association for Computational Linguistics.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2019. Learning to Decipher Hate Symbols. 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 3006–3015, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Niloofar Safi Samghabadi, Parth Patwa, Srinivas PYKL, Prerana Mukherjee, Amitava Das, and Thamar Solorio. 2020. Aggression and Misogyny Detection using BERT: A Multi-Task Approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 126–131, Marseille, France. European Language Resources Association (ELRA).

Joni Salminen, Maximilian Hopf, Shammur A. Chowdhury, Soon-gyo Jung, Hind Almerekhi, and Bernard J. Jansen. 2020. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10(1):1.

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.

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.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. 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, San Diego, California. Association for Computational Linguistics.

Zeerak Talat, James Thorne, and Joachim Bingel. 2018. Bridging the Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection. In Jennifer Golbeck, editor, *Online Harassment*, Human–Computer Interaction Series, pages 29–55. Springer International Publishing, Cham.

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. Large-Scale Hate Speech Detection with Cross-Domain Transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France. European Language Resources Association.

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, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. 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.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: A review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah Smith. 2021. Challenges in Automated Debiasing for Toxic Language Detection. In *Proceedings of the 16th Conference of the European*

*Chapter of the Association for Computational Linguistics: Main Volume*, pages 3143–3155, Online. Association for Computational Linguistics.

## A Experimental Parameters

In each cross-topic/cross-dataset experiment, we train one of the classifiers models described in Section 4 on either a specific dataset, or on a mixture of datasets (see Table 1). For all models, we use the Hugging Face transformer library's implementation of the AdamW optimizer (Loshchilov and Hutter, 2019), with the default $5 \times 10^{-5}$ learning rate for all models, except for T5-base ($1 \times 10^{-4}$) and ToxDectRoBERTa ($1 \times 10^{-5}$), as recommended by their respective authors. When possible using our available hardware (multiple Nvidia GTX 1080 Ti and RTX 2080 GPUs), we use "effective" batch sizes of 64 instances (either as proper mini-batches, or by using gradient accumulation alongside smaller mini-batch sizes). When using mixtures of datasets, batches are constructed by sampling each component dataset proportionally to its relative size, to avoid oversampling the smaller datasets or subsampling the larger ones. Models are trained for a maximum of 8 epochs, with early stopping according to the validation loss, and only the best model checkpoints are retained, according to the validation macro F1 score. Average runtimes vary between architectures and datasets, with the BERT-like (fBERT, HateBERT, and RoBERTa-base) taking the least time (less than an hour on the largest mixture of datasets, `All`, combining all 8 datasets detailed in Table 1), and the longest being ToxDectRoBERTa, since it is based on the RoBERTa-large architecture (approximately 6 hours of training for the `All` mixture). The smallest architectures in number of parameters used in this study are those derived from BERT-base, with $\sim$ 110 million parameters, followed by RoBERTa-base ($\sim$ 125 million parameters), T5-base (220 million parameters), and finally ToxDectRoBERTa, based on RoBERTa-large, with 355 million parameters. For data preprocessing, we replace all emojis with their text form descriptions using the Python `emoji` library, and replace all "@" user mentions and URLs with the replacement strings "*[USER]*" and "*[URL]*" respectively.

| Test \ Train | Gene Davi | Gene Fnta | Gndr Evit | Gndr HatE | Gndr Iber | Gndr Wasm | Gndr HatE | Gndr Wasm |
|---|---|---|---|---|---|---|---|---|
| **T5-base** | | | | | | | | |
| Gene Davi | 92.70 | 82.64 | 69.32 | 68.79 | 63.34 | 61.80 | 59.51 | 43.28 |
| Gene Fnt | 73.02 | 81.08 | 59.82 | 59.52 | 59.16 | 54.58 | 60.22 | 51.77 |
| Gndr Evit | 63.29 | 58.07 | 67.00 | 78.40 | 66.89 | 58.00 | 52.97 | 34.62 |
| Gndr HatE | 42.21 | 42.90 | 44.55 | 46.29 | 34.09 | 49.62 | 54.28 | 36.58 |
| Gndr Iber | 57.76 | 74.22 | **93.46** | 92.02 | 84.86 | 74.15 | 79.59 | 37.90 |
| Gndr Wasm | 54.06 | 59.07 | 63.65 | 68.92 | 67.90 | 85.81 | 51.00 | 42.46 |
| Race HatE | 43.75 | 64.54 | 42.14 | 43.07 | 37.59 | 41.29 | 38.73 | 41.09 |
| Race Wasm | 56.66 | 72.73 | 49.59 | 48.05 | 46.68 | 47.67 | 66.64 | 87.03 |
| **ToxDectRoBERTa** | | | | | | | | |
| Gene Davi | **93.78** | 84.73 | 73.74 | 70.36 | 74.33 | 69.57 | **81.34** | 60.76 |
| Gene Fnt | 81.42 | **85.92** | 72.81 | 67.81 | 74.96 | 60.89 | 78.97 | 61.90 |
| Gndr Evit | 62.10 | 54.25 | 67.37 | 82.08 | 63.08 | 64.69 | 59.10 | 45.69 |
| Gndr HatE | 40.30 | 39.81 | 45.63 | 51.65 | 32.88 | 47.04 | 41.59 | 50.67 |
| Gndr Iber | 64.14 | 73.00 | 91.25 | 92.07 | 83.89 | 74.35 | 68.65 | 38.93 |
| Gndr Wasm | 51.13 | 56.44 | 67.74 | 68.82 | 67.69 | 86.27 | 49.40 | 43.95 |
| Race HatE | 52.23 | 68.58 | 57.57 | 53.00 | 47.86 | 37.04 | 40.75 | 63.48 |
| Race Wasm | 69.92 | 73.47 | 68.07 | 62.16 | 64.55 | 46.50 | 73.59 | 86.29 |
| **fBERT** | | | | | | | | |
| Gene Davi | 92.46 | 85.59 | 64.64 | 66.18 | 67.33 | 73.69 | 64.46 | 42.88 |
| Gene Fnt | 76.47 | 80.98 | 58.23 | 59.32 | 61.04 | 60.43 | 62.45 | 51.57 |
| Gndr Evit | 61.44 | 57.00 | 73.12 | 88.66 | 69.40 | 63.19 | 52.79 | 34.48 |
| Gndr HatE | 38.25 | 49.75 | 64.34 | 64.28 | 43.62 | 50.30 | 55.19 | 36.58 |
| Gndr Iber | 60.34 | 71.43 | 93.01 | **93.02** | 84.55 | 76.99 | 52.80 | 38.66 |
| Gndr Wasm | 53.84 | 61.47 | 67.03 | 64.38 | 72.60 | 86.28 | 47.52 | 42.99 |
| Race HatE | 50.65 | 66.49 | 39.31 | 39.74 | 37.56 | 38.95 | 41.72 | 39.73 |
| Race Wasm | 70.16 | 72.33 | 48.06 | 46.64 | 47.80 | 47.53 | 74.03 | **87.53** |
| **HateBERT** | | | | | | | | |
| Gene Davi | 93.14 | 85.43 | 63.89 | 65.02 | 60.67 | 70.70 | 61.63 | 45.68 |
| Gene Fnt | 73.02 | 79.41 | 57.88 | 58.19 | 57.47 | 59.74 | 59.25 | 52.59 |
| Gndr Evit | 63.92 | 56.74 | 71.02 | 87.98 | 68.99 | 64.06 | 47.37 | 34.71 |
| Gndr HatE | 44.46 | 55.76 | 53.70 | 59.11 | 50.95 | 59.43 | 53.59 | 36.58 |
| Gndr Iber | 60.12 | 67.11 | 92.70 | 92.71 | 86.41 | 68.08 | 63.74 | 37.90 |
| Gndr Wasm | 52.36 | 58.24 | 74.26 | 68.54 | 71.20 | 86.05 | 50.83 | 43.01 |
| Race HatE | 44.13 | 66.08 | 43.79 | 43.60 | 37.11 | 36.89 | 38.53 | 39.26 |
| Race Wasm | 63.01 | 72.81 | 49.96 | 47.77 | 48.03 | 45.92 | 71.60 | 86.37 |
| **RoBERTa-base** | | | | | | | | |
| Gene Davi | 93.30 | 82.50 | 66.12 | 64.43 | 63.00 | 68.02 | 67.35 | 46.57 |
| Gene Fnt | 74.50 | 80.84 | 59.21 | 56.56 | 58.06 | 55.65 | 61.92 | 53.28 |
| Gndr Evit | 60.68 | 58.51 | 69.58 | 86.77 | 60.53 | 63.18 | 45.80 | 35.92 |
| Gndr HatE | 46.39 | 55.17 | 50.27 | 56.88 | 42.64 | 47.22 | 53.31 | 36.96 |
| Gndr Iber | 60.50 | 62.96 | 92.43 | 92.61 | **88.76** | 65.55 | 74.30 | 37.90 |
| Gndr Wasm | 51.55 | 51.11 | 71.14 | 68.11 | 74.93 | **87.11** | 53.06 | 43.70 |
| Race HatE | 47.79 | 56.76 | 43.30 | 40.56 | 37.86 | 37.10 | 38.24 | 48.65 |
| Race Wasm | 56.89 | 75.52 | 49.67 | 47.58 | 49.08 | 46.87 | 70.09 | 86.27 |

Table A: Detailed results of learning from individual datasets (in terms of Macro F1-scores).