

# Towards generalisable hate speech detection: a review on obstacles and solutions

Wenjie Yin <sup>Corresp., 1</sup>, Arkaitz Zubiaga <sup>1</sup>

<sup>1</sup> School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom

Corresponding Author: Wenjie Yin  
Email address: w.yin@qmul.ac.uk

Hate speech is one type of harmful online content which directly attacks or promotes hate towards a group or an individual member based on their actual or perceived aspects of identity, such as ethnicity, religion, and sexual orientation. With online hate speech on the rise, its automatic detection as a natural language processing task is gaining increasing interest. However, it is only recently that it has been shown that existing models generalise poorly to unseen data. This survey paper attempts to summarise how generalisable existing hate speech detection models are, reason why hate speech models struggle to generalise, sums up existing attempts at addressing the main obstacles, and then proposes directions of future research to improve generalisation in hate speech detection.

# Towards generalisable hate speech detection: a review on obstacles and solutions

Wenjie Yin<sup>1</sup> and Arkaitz Zubiaga<sup>1</sup>

<sup>1</sup>Queen Mary University of London, London, UK

Corresponding author:

Wenjie Yin<sup>1</sup>

Email address: w.yin@qmul.ac.uk

## ABSTRACT

Hate speech is one type of harmful online content which directly attacks or promotes hate towards a group or an individual member based on their actual or perceived aspects of identity, such as ethnicity, religion, and sexual orientation. With online hate speech on the rise, its automatic detection as a natural language processing task is gaining increasing interest. However, it is only recently that it has been shown that existing models generalise poorly to unseen data. This survey paper attempts to summarise how generalisable existing hate speech detection models are, reason why hate speech models struggle to generalise, sums up existing attempts at addressing the main obstacles, and then proposes directions of future research to improve generalisation in hate speech detection.

## INTRODUCTION

The Internet saw a growing body of user-generated content as social media platforms flourished (Schmidt and Wiegand, 2017; Chung et al., 2019). While social media provides a platform for all users to freely express themselves, offensive and harmful content are not rare and can severely impact user experience and even the civility of a community (Nobata et al., 2016). One type of such harmful content is **hate speech**, which is speech that **directly attacks** or **promotes hate** towards a group or an individual member based on their actual or perceived aspects of **identity**, such as ethnicity, religion, and sexual orientation (Waseem and Hovy, 2016; Davidson et al., 2017; Founta et al., 2018; Sharma et al., 2018).

Major social media companies are aware of the harmful nature of hate speech and have policies regarding the moderation of such posts. However, the most commonly used mechanisms are very limited. For example, keyword filters can deal with profanity, but not the nuance in the expression of hate (Gao et al., 2017). Crowd-sourcing methods (e.g. human moderators, user reporting), on the other hand, do not scale up. This means that by the time that a hateful post gets detected and taken down, it has already made negative impacts (Chen et al., 2019).

The automatic detection of hate speech is thus an urgent and important task. Since the automatic detection of hate speech was formulated as a task in the early 2010s (Warner and Hirschberg, 2012), the field has been constantly growing along the perceived importance of the task.

### Hate speech, offensive language, and abusive language

Although different types of abusive and offensive language are closely related, there are important distinctions to note. Offensive language and abusive language are both used as umbrella terms for harmful content in the context of automatic detection studies. However, while “strongly impolite, rude” and possible use of profanity are seen in the definitions of both (Fortuna and Nunes, 2018), abusive language has a strong component of intentionality (Caselli et al., 2020). Thus, offensive language has a broader scope, and hate speech falls in both categories.

Because of its definition mentioned above, hate speech is also different from other sub-types of offensive language. For example, personal attacks (Wulczyn et al., 2017) are characterised by being directed at an individual, which is not necessarily motivated by the target’s identity. Hate speech is also

different from cyberbullying (Zhao et al., 2016), which is carried out repeatedly and over time against vulnerable victims that cannot defend themselves<sup>1</sup>. This paper focuses on hate speech and hate speech datasets, although studies that cover both hate speech and other offensive language are also mentioned.

## Generalisation

Most if not all proposed hate speech detection models rely on supervised machine learning methods, where the ultimate purpose is for the model to learn the real relationship between features and predictions through training data, which generalises to previously unobserved inputs (Goodfellow et al., 2016). The **generalisation performance** of a model measures how well it fulfils this purpose.

To approximate a model's generalisation performance, it is usually evaluated on a set-aside test set, assuming that the training and test data, and future possible cases come from the same distribution. This is also the main way of evaluating a model's ability to generalise in the field of hate speech detection.

## Generalisability in hate speech detection

The ultimate purpose of studying automatic hate speech detection is to facilitate the alleviation of the harms brought by online hate speech. To fulfil this purpose, hate speech detection models need to be able to deal with the constant growth and evolution of hate speech, regardless of its form, target, and speaker.

Recent research has raised concerns on the generalisability of existing models (Swamy et al., 2019). Despite their impressive performance on their respective test sets, the performance significantly dropped when the models are applied to a different hate speech dataset. This means that the assumption that test data of existing datasets represent the distribution of future cases is not true, and that the generalisation performance of existing models have been severely overestimated (Arango et al., 2020). This lack of generalisability undermines the practical value of these hate speech detection models.

So far, existing research has mainly focused on demonstrating the lack of generalisability and (Gröndahl et al., 2018; Swamy et al., 2019; Wiegand et al., 2019; Fortuna et al., 2021), apart from a handful of studies that made individual attempts at addressing aspects of it (Karan and Šnajder, 2018; Waseem et al., 2018; Arango et al., 2020). Recent survey papers on hate speech and abusive language detection (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018; Al-Hassan and Al-Dossari, 2019; Mishra et al., 2019; Vidgen et al., 2019; Poletto et al., 2020; Vidgen and Derczynski, 2020) have focused on the general trends in this field, mainly by comparing features, algorithms and datasets. Among these, Fortuna and Nunes (2018) provided an in-depth review of definitions, Vidgen et al. (2019) concisely summarised various challenges for the detection of abusive language in general, Poletto et al. (2020) and Vidgen and Derczynski (2020) created extensive lists of dataset and corpora resources, while Al-Hassan and Al-Dossari (2019) focused on the special case of the Arabic language.

This survey paper thus contributes to the literature by providing (1) a comparative summary of existing research that demonstrated the lack of generalisability in hate speech detection models, (2) a systematic analysis of the main obstacles to generalisable hate speech detection and existing attempts to address them, and (3) suggestions for future research to address these obstacles.

This paper is most relevant to any researcher building datasets of, or models to detect, online hate speech, but can also be of use for those who work on other types of abusive or offensive language.

## SURVEY METHODOLOGY

For each of the three aims of this paper mentioned above, literature search was divided into stages.

### Sources of search

Across different stages, Google Scholar was the main search engine, and two main sets of keywords were used. References and citations were checked back-and-forth, with the number of iterations depending on how coarse or fine-grained the search of that stage was.

- General keywords: “hate speech”, “offensive”, “abusive”, “toxic”, “detection”, “classification”.
- Generalisation-related keywords: “generalisation” (“generalization”), “generalisability” (“generalizability”), “cross-dataset”, “cross-domain”, “bias”.

<sup>1</sup>for a more elaborate comparison between similar concepts, see Fortuna and Nunes (2018), Poletto et al. (2020), Banko et al. (2020)

We started with a pre-defined set of keywords. Then, titles of proceedings of the most relevant recent conferences and workshops (Workshop on Abusive Language Online, Workshop on Online Abuse and Harms) were skimmed, to refine the set of keywords. We also modified the keywords during the search stages as we encountered new phrasing of the terms. The above keywords shown are the final keywords.

### Main literature search stages

Before starting to address the aims of this paper, an initial coarse literature search involved searching for the general keywords, skimming the titles and abstracts. During this stage, peer-reviewed papers with high number of citations, published in high-impact venues were prioritised. Existing survey papers on hate speech and abusive language detection (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018; Al-Hassan and Al-Dossari, 2019; Mishra et al., 2019; Vidgen et al., 2019; Poletto et al., 2020; Vidgen and Derczynski, 2020) were also used as seed papers. The purpose of this stage was to establish a comprehensive high-level view of the current state of hate speech detection and closely related fields.

For the first aim of this paper – building a comparative summary of existing research on generalisability in hate speech detection – the search mainly involved different combinations of the general and generalisation-related keywords. As research on this topic is sparse, during this stage, all papers found and deemed relevant were included.

Building upon the first two stages, the main obstacles towards generalisable hate speech detection were then summarised: (1) presence of non-standard grammar and vocabulary, (2) paucity of and biases in datasets, and (3) implicit expressions of hate. This was done through extracting and analysing the error analysis of experimental studies found in the first stage, and comparing the results and discussions of the studies found in the second stage. Then, for each category of obstacles identified, another search was carried out, involving combinations of the description and paraphrases of the challenges and the general keywords. The search in this stage is the most fine-grained, in order to ensure coverage of both the obstacles and existing attempts to address them.

After the main search stages, the structure of the main findings in the literature was laid out. During writing, for each type of findings, the most representative studies were included in the writing up. We defined the relative representativeness within studies we have found, based on novelty, experiment design and error analysis, publishing venues, and influence. We also prioritised studies that addressed problems specific to hate speech, compared to better-known problems that are shared with other offensive language and social media tasks.

## GENERALISATION STUDIES IN HATE SPEECH DETECTION

Testing a model on a different dataset from the one which it was trained on is one way to more realistically estimate models' generalisability (Wiegand et al., 2019). This evaluation method is called cross-dataset testing (Swamy et al., 2019) or cross-application (Gröndahl et al., 2018), and sometimes cross-domain classification (Wiegand et al., 2019) or detection (Karan and Šnajder, 2018) if datasets of other forms of offensive language are also included.

As more hate speech and offensive language datasets emerged, a number of studies have touched upon cross-dataset generalisation since 2018, either studying generalisability per se, or as part of their dataset validation. The datasets they use (Table 1) to some extent reflect the best-known datasets in hate speech and other types of offensive language. These studies are further compared in Table 2 in terms of the models and datasets they used. As different datasets and models were investigated, instead of specific performance metrics, the remainder of this section will discuss the general findings of these studies, which can be roughly grouped into those on models and those on training and evaluation data.

### Models

First of all, **model performance had been severely over-estimated**. This includes existing “state-of-the-art” models and common baselines. Models used in the experiments ranged from neural networks – deep or shallow – to classical machine learning methods, including mixtures of both. When applied cross-dataset, all show a significant performance drop. Performance on a different dataset highlights that the test set of the same dataset does not realistically represent the distribution of unseen data.

Earlier (before 2019) state-of-the-art models often involved recurrent neural networks (Gröndahl et al., 2018).

Dataset name	Publication	Source	Positive labels	Annotator type
Waseem	Waseem and Hovy (2016); Waseem (2016)	Twitter	Racism Sexism	Expert; Expert and crowdsourcing
Davidson	Davidson et al. (2017)	Twitter	Hate speech Offensive	Crowdsourcing
Founta	Founta et al. (2018)	Twitter	Hate speech Offensive	Crowdsourcing
HatEval	Basile et al. (2019)	Twitter	Hateful	Crowdsourcing
Kaggle	Jigsaw (2018)	Wikipedia	Toxic Severe toxic Obscene Threat Insult Identity hate	Crowdsourcing
Gao	Gao and Huang (2017)	Fox News	Hateful	? (Native speakers)
AMI	Fersini et al. (2018a) Fersini et al. (2018b)	Twitter	Misogynous	Expert
Warner	Warner and Hirschberg (2012)	Yahoo! American Jewish Congress	Anti-semitic Anti-black Anti-asian Anti-woman Anti-muslim Anti-immigrant Other-hate	? (Volunteer)
Zhang	Zhang et al. (2018)	Twitter	Hate	Expert
Stromfront	de Gibert et al. (2018)	Stormfront	Hate	Expert
Kumar	Kumar et al. (2018b)	Facebook, Twitter	Overtly aggressive Covertly aggressive	Expert
Wulczyn	Wulczyn et al. (2017)	Wikipedia	Attacking	Crowdsourcing
OLID (OffensEval)	Zampieri et al. (2019a)	Twitter	Offensive	Crowdsourcing
AbuseEval	Caselli et al. (2020)	Twitter	Explicit (abuse) Implicit (abuse)	Expert
Kolhatkar	Kolhatkar et al. (2019)	The Globe and Mail	Very toxic Toxic Mildly toxic	Crowdsourcing
Razavi	Razavi et al. (2010)	Natural Semantic Module Usenet	Flame	Expert
Golbeck	Golbeck et al. (2017)	Twitter	Harassing	Expert

**Table 1.** English datasets used in cross-dataset generalisation studies. Positive labels are listed with their original wording. Expert annotation type include authors and experts in social science and related fields. ?: Type of annotations not available in original paper, the found descriptions are thus included. Note that only datasets used in generalisation studies are listed – for comprehensive lists of hate speech datasets, see Vidgen and Derczynski (2020) and Poletto et al. (2020).

Dataset name	Type	Study											
		Karan and Šnajder (2018)	Gröndahl et al. (2018)	Waseem (2016)	Wiegand et al. (2019)	Swamy et al. (2019)	Pamungkas and Patti (2019); Pamungkas et al. (2020)	Arango et al. (2020)	Fortuna et al. (2020)	Caselli et al. (2020)	Nejadgholi and Kiritchenko (2020)	Glavaš et al. (2020)	Fortuna et al. (2021)
Waseem	H*	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓
Davidson	H,O		✓	✓		✓			✓				✓
Founta	H,O				✓	✓					✓		✓
HatEval	H*						✓	✓	✓	✓			✓
Kaggle	H,O*	✓			✓				✓				✓
Gao	H	✓										✓	
AMI	H*						✓		✓				✓
Warner	H				✓								
Zhang	H		✓										
Stromfront	H												✓
TRAC	O	✓			✓				✓			✓	✓
Wulczyn	O	✓	✓								✓	✓	
OLID	O					✓	✓			✓			✓
AbuseEval	O*									✓			
Kolhatkar	O	✓											
Razavi	O				✓								
Golbeck	O						✓						
model		SVM	LR, MLP, LSTM, CNN-GRU	MLP	FastText	BERT	SVM, LSTM, GRU, BERT	LSTM, GBDT	N/A	BERT	BERT	BERT, RoBERTa	BERT, AL-BERT, fastText, SVM

**Table 2.** Comparison of studies that looked at cross-dataset (“cross-domain”) generalisation, by datasets and models used. Dataset types: H: hate speech, O: other offensive language, \*: contains subtypes. Most studies carried out cross-dataset experiments, training and testing each model on all datasets. The exceptions are: Gröndahl et al. (2018) and Nejadgholi and Kiritchenko (2020) used different datasets for training and testing; Fortuna et al. (2020) compared datasets through class vector representations.

For example, the CNN-GRU model by Zhang et al. (2018) first extracts 2 to 4-gram features using convolutional layers with varying kernel sizes on word embeddings, then captures the sequence orders of these features with a gated recurrent unit (GRU) layer. This model outperformed previous models on six datasets when tested in-dataset. However, when tested cross-dataset by Gröndahl et al. (2018), the model's performance dropped even more than an LSTM, by over 30 points in macro-averaged F1.

Similarly, Badjatiya et al. (2017)'s model was once considered state-of-the-art when trained and evaluated on *Waseem*. Their two-stage training first produces word embeddings using a Long Short-Term Memory (LSTM) network through the same hate speech classification task, based on which another Gradient-Boosted Decision Tree (GBDT) classifier was trained. Arango et al. (2020) showed a similar F1 drop of around 30 when applied on *HatEval*, and discussed a crucial methodological flaw – overfitting induced by extracting features on the combination of training and test set. Gröndahl et al. (2018) also reported that they failed to reproduce Badjatiya et al. (2017)'s results. Both Gröndahl et al. (2018) and Arango et al. (2020) also tested a Long Short-Term Memory (LSTM) network, which had been commonly used as a strong baseline. The performance drop was similar to the above two state-of-the-art models by Zhang et al. (2018) and Badjatiya et al. (2017).

Since the introduction of BERT (Devlin et al., 2019), itself and its variants have been established as the new state-of-the-art. This is seen through the comparison to other neural networks (Swamy et al., 2019) and on the leaderboards of shared tasks, such as Zampieri et al. (2020); Fersini et al. (2020). The general approach is to fine-tune a model, which had been pre-trained on domain-general data, on a target classification dataset. Yet, BERT and its variants are no exception to the lack of generalisation, although the cross-dataset performance drop is seemingly smaller. In cross-dataset experiments with four datasets, macro-averaged F1 scores decreased by 2 to 30 points (Swamy et al., 2019), which is less drastic compared to earlier state-of-the-art neural networks tested in other studies (Gröndahl et al., 2018; Arango et al., 2020). Pamungkas et al. (2020); Fortuna et al. (2021) also found that BERT and ALBERT tended to generalise the best across the models they experimented with.

Building upon BERT, a handful of recent studies suggest that additional hate-specific knowledge from outside the fine-tuning dataset might help with generalisation. Such knowledge can come from further masked language modelling pre-training on an abusive corpus (Caselli et al., 2021), or features from a hate speech lexicon (Koufakou et al., 2020).

Other models that have been studied include traditional machine learning models, such as character n-gram Logistic Regression (Gröndahl et al., 2018), character n-gram Multi-Layer Perceptron (MLP) (Gröndahl et al., 2018; Waseem et al., 2018), Support Vector Machines (Karan and Šnajder, 2018; Fortuna et al., 2021; Pamungkas and Patti, 2019; Pamungkas et al., 2020), and shallow networks with pre-trained embeddings, e.g. MLP with Byte-Pair Encoding (BPE)-based subword embeddings (Heinzerling and Strube, 2018; Waseem et al., 2018) and FastText (Joulin et al., 2017; Wiegand et al., 2019; Fortuna et al., 2021).

Generally, these simpler models do not perform as good as deep neural networks, such as LSTM (Pamungkas and Patti, 2019) and especially BERT and its variants (Pamungkas et al., 2020; Fortuna et al., 2021), in- or cross-dataset. However, exceptions exist in some dataset combinations, especially when it comes to generalising. For example, n-gram Logistic Regression when comparing to LSTM (Gröndahl et al., 2018), SVM when comparing to LSTM and BERT (Pamungkas et al., 2020), and FastText when comparing to BERT (Fortuna et al., 2021).

These cross-dataset studies only cover some of the more representative and/or recent hate speech detection models, but one can expect that the generalisation problem go beyond this small sample, and is far more ubiquitous in existing models than what these studies cover.

Despite the significance of the problem, systematic studies that compared a variety of models with datasets controlled are very limited (Arango et al., 2020; Pamungkas and Patti, 2019; Pamungkas et al., 2020; Fortuna et al., 2021); there is also limited overlap in the datasets used between different studies (Table 2). Thus, one should be careful when drawing conclusions on the relative generalisability of models.

## Data

**Training data has a pronounced influence on generalisation.** The performance drops in models highlight the differences in the distribution of posts between datasets (Karan and Šnajder, 2018), yet some datasets are more similar to each other. Furthermore, certain attributes of a dataset could lead to more

200 generalisable models.

201 **Similarity between datasets** varies, as there are groups of datasets that produce models that test  
202 much better on each other. For example, in Wiegand et al. (2019)’s study, FastText models (Joulin et al.,  
203 2017) trained on three datasets (*Kaggle*, *Founta*, *Razavi*) achieved F1 scores above 70 when tested on  
204 one another, while models trained or tested on datasets outside this group achieved around 60 or less.  
205 In Swamy et al. (2019)’s study with fine-tuned BERT models (Devlin et al., 2019), *Founta* and *OLID*  
206 produced models that performed well on each other. The source of such differences are usually traced  
207 back to search terms (Swamy et al., 2019), topics covered Nejadgholi and Kiritchenko (2020); Pamungkas  
208 et al. (2020), label definitions Pamungkas et al. (2020); Fortuna et al. (2021), and data source platforms  
209 (Glavaš et al., 2020; Karan and Šnajder, 2018).

210 Another way of looking at generalisation and similarity is by comparing differences **between individ-**  
211 **ual classes** across datasets (Nejadgholi and Kiritchenko, 2020; Fortuna et al., 2020, 2021), as opposed to  
212 comparing datasets as a whole. In both Nejadgholi and Kiritchenko (2020) and Fortuna et al. (2021)’s  
213 experiments, the best generalisation is achieved for more general labels such as “toxicity”, “offensive”,  
214 or “abusive”. Generalisation is not as good for finer-grained hate speech labels. All in all, these findings  
215 are indicative of an imbalance of the finer-grained subclasses, particularly owing to disagreements in the  
216 definition of what constitutes hate speech, which proves more difficult than defining what constitutes  
217 offensive language.

218 Within the hate speech labels, the relative similarity also varies. Fortuna et al. (2020) used averaged  
219 word embeddings (Bojanowski et al., 2017; Mikolov et al., 2018) to compute the representations of classes  
220 from different datasets, and compared classes across datasets. One of their observations is that *Davidson*’s  
221 “hate speech” is very different from *Waseem*’s “hate speech”, “racism”, “sexism”, while being relatively  
222 close to *HatEval*’s “hate speech” and *Kaggle*’s “identity hate”. This echoes with experiments that showed  
223 poor generalisation of models from *Waseem* to *HatEval* (Arango et al., 2020) and between *Davidson* and  
224 *Waseem* (Waseem et al., 2018; Gröndahl et al., 2018).

225 In terms of what properties of a dataset lead to more generalisable models, there are frequently  
226 mentioned factors, but also inconsistency across different studies. Interactions between factors, which  
227 contribute to the inconsistency, are also reported.

228 The **proportion of abusive posts** in a dataset, first of all, plays a part. Swamy et al. (2019) holds that  
229 a larger proportion of abusive posts (including hateful and offensive) leads to better generalisation to  
230 dissimilar datasets, such as *Davidson*. This is in line with Karan and Šnajder (2018)’s study where *Kumar*  
231 and *Kolhatkar* generalised best, and Waseem et al. (2018)’s study where models trained on *Davidson*  
232 generalised better to *Waseem* than the other way round. In contrast, in Wiegand et al. (2019)’s study, the  
233 datasets with the least abusive posts generalised the best (*Kaggle* and *Founta*). Similarly, Fortuna et al.  
234 (2021) could not confirm the impact of class proportions. Nejadgholi and Kiritchenko (2020) offered an  
235 explanation to this: there exists a trade-off between true positive and true negative rates dictated by the  
236 class proportions, which impacts the minority class performance the most but this is not always reflected  
237 in the overall F1 score.

238 **Biases in the samples** are also frequently mentioned. Wiegand et al. (2019) hold that less biased  
239 sampling approaches produce more generalisable models. This was later reproduced by Razo and Kübler  
240 (2020) and also helps explain their results with the two datasets that have the least positive cases. Similarly,  
241 Pamungkas and Patti (2019) mentioned that a wider coverage of phenomena lead to more generalisable  
242 models. So do topics that are more general rather than platform-specific (Nejadgholi and Kiritchenko,  
243 2020).

244 A larger training **data size** is generally believed to produce better and more generalisable models  
245 (Halevy et al., 2009). It is mentioned as one of the two biggest factors contributing to cross-dataset  
246 performance in Karan and Šnajder (2018)’s study. Caselli et al. (2020) also found that, on *HatEval*, their  
247 dataset (*AbuseEval*) produced a model even better-performing than the one trained on *HatEval* end-to-end.  
248 They partially attributed this to a bigger data size, alongside **annotation quality**. However, the benefit  
249 of having more data is counterbalanced by data distribution differences (Karan and Šnajder, 2018), as  
250 discussed above. Moreover, its relative importance compared to other factors seems to be small, when the  
251 latter are carefully controlled (Nejadgholi and Kiritchenko, 2020; Fortuna et al., 2021).



Study	Data translation	Embedding and classifier models	Additional languages
Pamungkas and Patti (2019)	Training, automatic	MUSE embeddings, LSTM, "joint-learning"	Spanish, Italian, German
Pamungkas et al. (2020)	Training, automatic	MUSE embeddings, mBERT, LSTM, SVM, "joint-learning"	Spanish, Italian
Glavaš et al. (2020)	Testing, manual	mBERT, XLM-R	Albanian, Croatian, German, Russian, Turkish
Arango et al. (2020)	Testing, automatic	MUSE embeddings, LSTM, GBDT	Spanish
Fortuna et al. (2021)	None	mBERT	Italian, Spanish, Portuguese

**Table 3.** Cross-lingual generalisation studies. All studies included English as the main language, hence only additional languages are mentioned.

### The cross-lingual case

Most of these studies only worked with English data. Yet, it is worth stressing that hate speech is a universal problem that exists in many languages, and generalisation studies focused on languages other than English are to date very sparse, despite the importance of the problem. Thus, **research on cross-lingual generalisation is still in early stages.**

One way to look at generalisation in non-English hate speech detection is applying the same cross-dataset evaluation on multiple datasets in another language. However, such studies do not yet exist. This is related to the fact that the majority of datasets are in English, which reflects linguistic and cultural unevenness in this field of research (Poletto et al., 2020; Vidgen and Derczynski, 2020).

Cross-lingual generalisation can be considered a more "extreme" type of generalisation (Arango et al., 2020). The ideal case would be to be able to use data in one language for training and apply the model on data in another language, which would help address the challenge in low-resource languages. In a few studies (Pamungkas et al., 2020; Glavaš et al., 2020; Arango et al., 2020; Fortuna et al., 2021), language was included as a separate variable, alongside a "domain" variable independent to it, which is characterised by the source platform or the data collection method. These cross-lingual experiments are summarised in Table 3

Although these studies all touch on the same problem, how they evaluate cross-lingual performance differs. There are two main ways of enabling cross-lingual experiments: translating data and using multi-lingual models. These studies differ mainly by whether they perform translation on training or testing data and whether the translation is automatic or manual. Studies that use different evaluation methods also tend to look at the difficulty of the task differently. For example, Fortuna et al. (2021) hold that multilingual generalisation per se is likely to be worse than its monolingual counterpart, while Arango et al. (2020) consider the two types of generalisation similar.

The factors that contribute to cross-lingual generalisation are similar to those in the monolingual setting as discussed above, with a few additional challenges:

- In terms of **models**, pre-trained multilingual word embeddings (MUSE (Conneau et al., 2017)) and language models (mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020)) are frequently chosen as baselines. They are an intuitive and easily accessible starting point for cross-lingual experiments, but their limitations are also clear – the "curse of multilinguality" trades off single-language performance for its broad language coverage, as in the results of the cross-lingual generalisation studies mentioned above (Pamungkas and Patti, 2019; Pamungkas et al., 2020; Glavaš et al., 2020; Arango et al., 2020; Fortuna et al., 2021) as well as in other tasks (Conneau et al., 2020). Similarly to the monolingual case, there are cases where traditional machine learning models outperform deep learning ones, such as SVM (Pamungkas et al., 2020) and GBDT (Arango et al., 2020) compared to LSTM. Adding automatically translated training data alongside the original is beneficial (Pamungkas and Patti, 2019; Pamungkas et al., 2020).
- When it comes to the **data**, the most prominent additional factor compared to the monolingual

289 setting is the similarity between the training (source) and testing (target) languages. For instance,  
290 Among the wide range of languages that Glavaš et al. (2020) have tested, the cross-lingual perfor-  
291 mance drop between English, the source language, and German, the most similar target language,  
292 was less than one third of that between English and Turkish, when using mBERT on *Wulczyn*.

293 Although these studies more or less consider the “language” and “domain” variables as separate,  
294 there exists evidence that the two types of generalisation interact with each other. Studies that control  
295 the language variable more carefully tend to show a smaller drop across languages – for example, by  
296 manually translating exactly the same data (Glavaš et al., 2020), as opposed to using automatic translation  
297 (Pamungkas and Patti, 2019; Pamungkas et al., 2020; Arango et al., 2020) or different language dataset  
298 from the same shared task (Pamungkas and Patti, 2019; Pamungkas et al., 2020; Fortuna et al., 2021).  
299 Furthermore, adding data from a different domain can act as a regulariser from overfitting to the training  
300 language (Glavaš et al., 2020).

301 As more datasets emerge, we can expect more generalisation studies considering language as a  
302 parameter in the near future. For the remainder of this paper, we discuss issues that can apply to hate  
303 speech detection in any language.

## 304 OBSTACLES TO GENERALISABLE HATE SPEECH DETECTION

305 Demonstrating the lack of generalisability is only the first step in understanding this problem. This section  
306 delves into three key factors that contribute to it: (1) presence of non-standard grammar and vocabulary,  
307 (2) paucity of and biases in datasets, and (3) implicit expressions of hate.

### 308 Non-standard Grammar and Vocabulary

309 Hate speech detection, which is largely focused on social media, shares similar challenges to other social  
310 media tasks and has its specific ones, when it comes to the grammar and vocabulary used. Such user  
311 language style introduces challenges to generalisability at the data source, mainly by making it difficult to  
312 utilise common NLP pre-training approaches.

313 On social media, syntax use is generally more casual, such as the omission of punctuation (Blodgett  
314 and O’Connor, 2017). Alternative spelling and expressions are also used in dialects (Blodgett and  
315 O’Connor, 2017), to save space, and to provide emotional emphasis (Baziotis et al., 2017). Sanguinetti  
316 et al. (2020) provided extensive guidelines for studying such phenomena syntactically.

317 Commonly seen in hate speech, the offender adopts various approaches to evade content moderation.  
318 For example, the spelling of offensive words or phrases can be obfuscated (Nobata et al., 2016; Serrà  
319 et al., 2017), and common words such as “Skype”, “Google”, and “banana” may have a hateful meaning –  
320 sometimes known as euphemism or code words (Taylor et al., 2017; Magu and Luo, 2018).

321 When the spelling is obfuscated, a word is considered out-of-vocabulary and thus no useful information  
322 can be given by the pre-trained models. In the case of code words, pre-trained embeddings will not  
323 reflect its context-dependent hateful meaning. At the same time, simply using identified code words for a  
324 lexicon-based detection approach will result in low precision (Davidson et al., 2017). As there are infinite  
325 ways of combining the above alternative rules of spelling, code words, and syntax, hate speech detection  
326 models struggle with these rare expressions even with the aid of pre-trained word embeddings.

327 In practice, this difficulty is manifested in false negatives. Qian et al. (2018) found that rare words and  
328 implicit expressions are the two main causes of false negatives; van Aken et al. (2018) compared several  
329 models that used pre-trained word embeddings, and found that rare and unknown words were present  
330 in 30% of the false negatives of Wikipedia data and 43% of Twitter data. Others have also identified  
331 rare and unknown words as a challenge for hate speech detection (Nobata et al., 2016; Zhang and Luo,  
332 2018). More recently, Fortuna et al. (2021) drew a more direct line between out-of-vocabulary words  
333 and generalisation performance, by showing that the former is one of the top contributing features in a  
334 classifier for the latter. It has also been shown as an important factor in the cross-lingual case (Pamungkas  
335 et al., 2020).

### 336 Existing solutions

337 From a domain-specific perspective, Taylor et al. (2017) and Magu and Luo (2018) attempted to  
338 **identify code words** for slurs used in hate communities. Both of them used keyword search as part of  
339 their sourcing of Twitter data and word embedding models to model word relationships. Taylor et al.  
340 (2017) identified hate communities through Twitter connections of the authors of extremist articles and

hate speech keyword searches. They trained their own dependency2vec (Levy and Goldberg, 2014) and FastText (Bojanowski et al., 2017) embeddings on the hate community tweets and randomly sampled “clean” tweets, and used weighted graphs to measure similarity and relatedness of words. Strong and weak links were thus drawn from unknown words to hate speech words. In contrast, Magu and Luo (2018) collected potentially hateful tweets using a set of known code words. They then computed the cosine similarity between all words based on a word2vec model (Mikolov et al., 2013) pre-trained on news data. Code words, which have a neutral meaning in news context, were further apart from other words which fit in the hate speech context. Both Taylor et al. (2017) and Magu and Luo (2018) focused on the discovery of such code words and expanding relevant lexicons, but their methods could potentially complement existing hate lexicons as classifier features or for data collection.

Recently, an increasing body of research is approaching the problem by adapting character or sequence-level features to evade the challenge posed by words:

The benefit of **character-level features** has not been consistently observed. Three studies compared character-level, word-level, and hybrid (both character- and word-level) CNNs, but drew completely different conclusions. Park (2018) and Meyer and Gambäck (2019) found hybrid and character CNN to perform best respectively. Probably most surprisingly, Lee et al. (2018) observed that word and hybrid CNNs outperformed character CNN to similar extents, with all CNNs performing worse than character n-gram logistic regression. Small differences between these studies could have contributed to this inconsistency. More importantly, unlike the word components of the models, which were initialised with pre-trained word embeddings, the character embeddings were trained end-to-end on the very limited respective training datasets. It is thus likely that these character embeddings overfit on the training data.

In contrast, simple character n-gram logistic regression has shown results as good as sophisticated neural network models, including the above CNNs (van Aken et al., 2018; Gao and Huang, 2017; Lee et al., 2018). Indeed, models with fewer parameters are less likely to overfit. This suggests that character-level features themselves are very useful, when used appropriately. A few studies used word embeddings that were additionally enriched with subword information as part of the pre-training. For example, FastText (Bojanowski et al., 2017) models were consistently better than hybrid CNNs (Bodapati et al., 2019). In addition, a MIMICK (Pinter et al., 2017)-based model displayed similar performances (Mishra et al., 2018).

The use of **sentence embeddings** partially solves the out-of-vocabulary problem by using the information of the whole post instead of individual words. Universal Sentence Encoder (Cer et al., 2018), combined with shallow classifiers, helped one team (Indurthi et al., 2019) achieve first place at the HatEval 2019 shared task (Basile et al., 2019). Sentence embeddings, especially those trained with multiple tasks, also consistently outperformed traditional word embeddings (Chen et al., 2019).

**Large language models** with sub-word information have the benefits of both subword-level word embeddings and sentence embeddings. They produce the embedding of each word with its context and word form. Indeed, BERT (Devlin et al., 2019) and its variants have demonstrated top performances at hate or abusive speech detection challenges recently (Liu et al., 2019a; Mishra and Mishra, 2019).

Nonetheless, these relatively good solutions to out-of-vocabulary words (subword- and context-enriched embeddings) all face the same short-coming: they have only seen the standard English retrieved from BookCorpus and Wikipedia. NLP tools perform best when trained and applied in specific domains (Duarte et al., 2018). In hate speech detection, word embeddings trained on relevant data (social media or news sites) had a clear advantage (Chen et al., 2018; Vidgen et al., 2020). The domain mismatch could have similarly impaired the subword- and context-enriched models’ performances. There is little work so far on adapting them to the abusive domain to increase model generalisability so far (Caselli et al., 2021).

## Limited, Biased Labelled Data

### Small data size

Obstacles to generalisability also lie in dataset construction, and dataset size is the relatively most unequivocal one. When using machine learning models, especially deep learning models with millions of parameters, small dataset size can lead to overfitting and in turn harm generalisability (Goodfellow et al., 2016).

It is particularly challenging to acquire labelled data for hate speech detection as knowledge or relevant training is required of the annotators. As a high-level and abstract concept, the judgement of “hate speech” is subjective, needing extra care when processing annotations. Hence, datasets are usually not big in size.

### Existing solutions

The use of **pre-trained embeddings** (discussed earlier) and parameter dropout (Srivastava et al., 2014) have been accepted as standard practice in the field of NLP to prevent over-fitting, and are common in hate speech detection as well. Nonetheless, the effectiveness of domain-general embedding models is questionable, and there has been only a limited number of studies that looked into the *relative* suitability of different pre-trained embeddings on hate speech detection tasks (Chen et al., 2018; Mishra et al., 2018; Bodapati et al., 2019).

In Swamy et al. (2019)'s study of model generalisability, **abusive language-specific pre-trained embeddings** were suggested as a possible solution to limited dataset sizes. Alatawi et al. (2020) proposed White Supremacy Word2Vec (WSW2V), which was trained on one million tweets sourced through white supremacy-related hashtags and users. Compared to general word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) models trained on news, Wikipedia, and Twitter data, WSW2V captured meaning more suitable in the hate speech context – e.g. ambiguous words like “race” and “black” have higher similarity to words related to ethnicity than sports or colours. Nonetheless, their WSW2V-based LSTM model did not consistently outperform Twitter GloVe-based LSTM model or BERT (Devlin et al., 2019). They did not consider cross-dataset testing for generalisability, either.

The pre-training for BERT (and its variants) is both data and computationally-heavy, which limits the feasibility of training the hate speech equivalent of BERT from scratch. A reasonable compromise to that is performing further Masked Language-Modelling pre-training before the fine-tuning stage. By further pre-training RoBERTa (Liu et al., 2019b), Wiedemann et al. (2020) achieved first place at the Offenseval 2020 shared task (Zampieri et al., 2020). Caselli et al. (2021) pre-trained BERT further on a larger-scale dataset of banned abusive subreddits and observed improvement over standard BERT on three Twitter datasets (*OLID*, *AbuseEval*, *HatEval*), in-dataset for all cases and cross-dataset for most cases. Both studies show that abusive language-specific pre-training, built upon generic pre-training, can be beneficial for both in-dataset performance and cross-dataset generalisation. The main downside is that the improvement gains, ranging from less than 1% to 4% in macro F1, seem disproportionate to the computational cost – Wiedemann et al. (2020) only did the training on a small sample due to hardware limitations; it took Caselli et al. (2021) 18 days to complete 2 million training steps on one Nvidia V100 GPU. There also exists a trade-off between precision and recall for the positive class due to the domain shift (Caselli et al., 2021).

Research on **transfer learning from other tasks**, such as sentiment analysis, also lacks consistency. Uban and Dinu (2019) pre-trained a classification model on a large sentiment dataset<sup>2</sup>, and performed transfer learning on the *OLID* and *Kumar* datasets. They took pre-training further than the embedding layer, comparing word2vec (Mikolov et al., 2013) to sentiment embeddings and entire-model transfer learning. Entire-model transfer learning was found to be always better than using the baseline word2vec (Mikolov et al., 2013) model, but the transfer learning performances with only the sentiment embeddings were not consistent.

More recently, Cao et al. (2020) also trained sentiment embeddings through classification as part of their proposed model. The main differences are: the training data was much smaller, containing only *Davidson* and *Founta* datasets; the sentiment labels were produced by VADER (Gilbert and Hutto, 2014); their model was deeper and used general word embeddings (Mikolov et al., 2013; Pennington et al., 2014; Wieting et al., 2015) and topic representation computed through Latent Dirichlet Allocation (LDA) (Blei et al., 2003) in parallel. Through ablation studies, they showed that sentiment embeddings were beneficial for both *Davidson* and *Founta* datasets.

Use of existing knowledge from a more mature research field like that of sentiment analysis has the potential to be used to jumpstart the relatively newer field of hate speech detection. It also offers a compromise between hate speech models, which might not be generalisable enough, and completely domain-general models, which lack knowledge specific to hate speech detection. Nonetheless, more investigation into the conditions in which transfer learning works best to increase generalisability in particular still needs to be done.

### Sampling bias

In addition to a limited size, datasets are also prone to biases. Non-random sampling and subjective annotations introduce individual biases, and the different sampling and annotation processes across

<sup>2</sup><https://help.sentiment140.com/>

448 datasets further increase the difficulty of training models that can generalise across heterogeneous data.

449 Hate speech and, more generally, offensive language generally represent less than 3% of social media  
450 content (Zampieri et al., 2019b; Founta et al., 2018). To alleviate the effect of scarce positive cases  
451 on model training, all existing social media hate speech or offensive content datasets used boosted (or  
452 focused) sampling with simple heuristics.

453 Table 4 compares the **sampling methods** of hate speech datasets studied the most in cross-dataset  
454 generalisation. Consistently, keyword search and identifying potential hateful users are the most common  
455 methods. However, what is used as the keywords (slurs, neutral words, profanity, hashtags), which users  
456 are included (any user from keyword search, identified haters), and the use of other sampling methods  
457 (identifying victims, sentiment classification) all vary a lot.

Dataset	Keywords	Haters	Other
Waseem	“Common slurs and terms used pertaining to religious, sexual, gender, and ethnic minorities”	“A small number of prolific users”	N/A
Davidson	HateBase <sup>3</sup>	“Each user from lexicon search”	N/A
Founta	HateBase, NoSwearing <sup>4</sup>	N/A	Negative sentiment
HatEval	“Neutral keywords and derogatory words against the targets, highly polarized hashtags”	“Identified haters”	“Potential victims of hate accounts”

**Table 4.** Boosted sampling methods of the most commonly studied hate speech datasets (Waseem and Hovy, 2016; Davidson et al., 2017; Founta et al., 2018; Basile et al., 2019). Description as appeared in the publications. N/A: no relevant descriptions found.

458 Moreover, different studies are based on varying definitions of “hate speech”, as seen in different  
459 **annotation guidelines** (Table 5). Despite all covering the same two main aspects (directly attack or  
460 promote hate towards), datasets vary by their wording, what they consider a target (any group, minority  
461 groups, specific minority groups), and their clarifications on edge cases. *Davidson* and *HatEval* both  
462 distinguished “hate speech” from “offensive language”, while “uses a sexist or racist slur” is in *Waseem*’s  
463 guidelines to mark a case positive of hate, blurring the boundary of offensive and hateful. Additionally,  
464 as both *HatEval* and *Waseem* specified the types of hate (towards women and immigrants; racism and  
465 sexism), hate speech that fell outside of these specific types were not included in the positive classes,  
466 while *Founta* and *Davidson* included any type of hate speech. Guidelines also differ in how detailed they  
467 are: Apart from *Founta*, all other datasets started the annotation process with sets of labels pre-defined  
468 by the authors, among which *Waseem* gave the most specific description of actions. In contrast, *Founta*  
469 only provided annotators with short conceptual definitions of a range of possible labels, allowing more  
470 freedom for a first exploratory round of annotation. After that, labels were finalised, and another round  
471 of annotation was carried out. As a result, the labelling reflects how the general public, without much  
472 domain knowledge or extensive training, would classify offensive language. For example, the “abusive”  
473 and “offensive” classes were so similar that they were merged in the second stage. However, as discussed  
474 above, they differ by whether intentionality is present (Caselli et al., 2020). Such different annotation and  
475 labelling criteria result in essentially different tasks and different training objectives, despite their data  
476 having a lot in common.

477 As a result of the varying and sampling methods, definitions, and annotation schemes, what current  
478 models can learn on one dataset is specific to the examples in that dataset and the task defined by the  
479 dataset, limiting the models’ ability to generalise to new data. One type of possible resulting bias is **author**  
480 **bias**. For example, 65% of the hate speech in the *Waseem* dataset was produced by merely two users, and  
481 their tweets exist in both the training and the test set. Models trained on such data thus overfit to these  
482 users’ language styles. This overfitting to authors was proven in two state-of-the-art models (Badjatiya

<sup>1</sup><https://www.hatebase.org/>

<sup>2</sup><https://www.noswearing.com/dictionary/>

<sup>3</sup><https://www.hatebase.org/>

<sup>4</sup><https://www.noswearing.com/dictionary/>

Dataset	Action	Target	Clarifications
Waseem	<u>Attacks, seeks to silence, criticises, negatively stereotypes, promotes hate speech or violent crime, blatantly misrepresents truth or seeks to distort views on,</u> uses a sexist or racial slur, defends xenophobia or sexism	A minority	(Inclusion) 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
Davidson	<i>Express hatred towards, humiliate, insult*</i>	A group or members of the group	(Exclusion) Think not just about the words appearing in a given tweet but about the context in which they were used; the presence of a particular word, however offensive, did not necessarily indicate a tweet is hate speech
Founta	<i>Express hatred towards, humiliate, insult</i>	Individual or group, on the basis of attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender	N/A
HatEval	<i>Spread, incite, promote, justify hatred or violence towards, dehumanizing, hurting or intimidating**</i>	Women or immigrants	(Exclusion) Hate speech against other targets, offensive language, blasphemy, historical denial, overt incitement to terrorism, offense towards public servants and police officers, defamation

**Table 5.** Annotation guidelines of the most commonly studied hate speech datasets. Original wording from the publications or supplementary materials; action verbs grouped for easier comparison: underlined: directly attack or attempt to hurt, *italic*: promote hate towards. N/A: no relevant descriptions found. \*Davidson et al. (2017) also gave annotators “a paragraph explaining it (the definition) in further detail”, which was not provided in their publication. \*\*Basile et al. (2019) also gave annotators some examples in their introduction of the task (rather than the main guidelines, thus not included).

et al., 2017; Agrawal and Awekar, 2018) (Arango et al., 2020). **Topic bias** is another concern. With words such as “football” and “announcer” among the ones with the highest Pointwise Mutual Information (PMI) with hate speech posts, a topic bias towards sports was demonstrated in the *Waseem* dataset (Wiegand et al., 2019). Such biases can also be measured through the semantic similarity between keywords used to build the datasets and topics present in the dataset (Ousidhoum et al., 2020).

### Existing solutions

A few recent studies have attempted to go beyond one dataset when training a model.

Waseem et al. (2018) used **multitask training** (Caruana, 1997) with hard parameter sharing up to the final classification components, which were each tuned to one hate speech dataset. The shared shallower layers, intuitively, extract features useful for both datasets, with the two classification tasks as regularisation against overfitting to either one. Their multitask-trained models matched the performances of models trained end-to-end to single datasets and had clear advantage over simple dataset concatenation, whilst allowing generalisation to another dataset. Karan and Šnajder (2018) presented a similar study. Frustratingly Easy **Domain Adaptation** (Daumé III, 2007) had similar beneficial effects but was much simpler and more efficient. These two studies showed the potential of combining datasets to increase generalisability, but further investigation into this approach is lacking.

# Representation bias

A different kind of bias is representation bias. To put simply, models trained on “norms” will fail to generalise to data far from the “norms”. This also harms model generalisability in a much broader sense, mainly through application practicality.

Natural language is a proxy of human behaviour, thus the biases of our society are reflected in the datasets and models we build. With increasing real-life applications of NLP systems, these biases can be translated into wider social impacts (Hovy and Spruit, 2016). Minority groups are underrepresented in available data and/or data annotators, thus causing biases against them when models are trained from this data. This phenomenon is also seen in audio transcribing (Tatman, 2017), sentiment analysis (Kiritchenko and Mohammad, 2018), etc.

Hate speech detection models not only have higher tendency to classify African-American English posts as offensive or hate than “white” English (Davidson et al., 2019), but also more often predict false negatives on “white” than African-American English (Sap et al., 2020). Certain words and phrases, including neutral identity terms such as “gay” (Dixon et al., 2018) and “woman” (Park et al., 2018) can also easily lead to a false positive judgement. Moreover, just like biases in real life, racial, gender, and party identification biases in hate speech datasets were found to be intersectional (Kim et al., 2020).

The prevalence of such biases mean that existing hate speech detection models are likely to struggle at generalising to unseen data that contain expressions related to these demographic groups. Furthermore, compared to the other types of biases mentioned above, they do more harm to the practical value of the automatic hate speech detection models. These biases may cause automatic models to amplify the harm against minority groups instead of mitigating such harm as intended (Davidson et al., 2019). For example, with higher false positive rates for minority groups, their already under-represented voice will be more often falsely censored.

## Existing solutions

Systematic studies of representation biases and their mitigation are relatively recent. Since Dixon et al. (2018) first quantified unintended biases in abusive language detection on the *Wulczyn* dataset using a synthetic test set, an increasing number of studies have been carried out on hate speech and other offensive language. These attempts to address biases against minority social groups differ by how they measure biases and their approaches to mitigate them.

Similar to Dixon et al. (2018), a number of studies measured bias as certain words and phrases being associated with the hateful or offensive class, which were mostly identity phrases. Attempts to mitigate biases identified this way focus on decoupling this association between features and classes. Model performance on a **synthetic test set** with classes and identity terms balanced, compared to the original test data, were used a measure for model bias. Well-known identity terms and synonyms are usually used as starting points (Dixon et al., 2018; Park et al., 2018; Nozza et al., 2019). Alternatively, bias-prone terms could be identified through looking at skewed distributions within a specific dataset (Badjatiya et al., 2019; Mozafari et al., 2020b).

A few studies measured biases across directly **predicted language styles or demographic attributes** of authors. Davidson et al. (2019) and Kim et al. (2020) both tested their hate speech detection models on Blodgett et al. (2016)’s distantly supervised dataset of African-American vs white-aligned English tweets, revealing higher tendencies of labelling an African-American-aligned tweet offensive or hateful. Kim et al. (2020) further extended this observation to gender and party identification. As the testing datasets do not have hateful or offensive ground truth labels, one caveat is that, using this as a metric of model bias assumes that all language styles have equal chances of being hateful or offensive, which might not be true.

Huang et al. (2020) approached author demographics from a different angle, and instead predicted author demographics on available hate speech datasets using user profile descriptions, names, and photos. They built and released a multilingual corpus for model bias evaluation. Although now with ground truth hate speech labels, this introduces additional possible bias existing in the tools they used into the bias evaluation process. For example, they used a computer vision API on the profile pictures to predict race, age, and gender, which displayed racial and gender biases (Buolamwini and Gebru, 2018).

One mitigation approach that stemmed from the first approach of measuring biases is “debiasing” training data through **data augmentation**. Dixon et al. (2018) retrieved non-toxic examples containing a range of identity terms following a template, which were added to *Wulczyn*. Following a similar logic, Park et al. (2018) created examples containing the counterpart of gendered terms found in the data to address gender bias in the *Waseem* and *Founta* datasets. Badjatiya et al. (2019) extended this word

replacement method by experimenting with various strategies including named entity tags, part of speech tags, hypernyms, and similar words from word embeddings, which were then applied on the *Wulczyn* and *Davidson* datasets.

Less biased **external corpora and pre-trained models** could also be used. To reduce gender bias, Park et al. (2018) also compared pre-trained debiased word embeddings (Bolukbasi et al., 2016) and transfer learning from a larger, less biased corpus. Similarly, Nozza et al. (2019) added samples from the *Waseem* dataset to their training dataset (*AMI*), to keep classes and gender identity terms balanced.

From the perspective of model training, biases could also be understood through **model explanation** and “debiasing” could be accordingly integrated into the **model training objective**. Based on 2-grams’ Local Mutual Information with a label, Mozafari et al. (2020b) gave each training example in the *Davidson* and *Waseem* datasets a positive weight, producing a new weighted loss function to optimise. Kennedy et al. (2020) built upon a recent study of post-hoc BERT feature importance (Jin et al., 2020). A regularisation term to encourage the importance of a set of identity terms to be close to zero was added to the loss function. This changed the ranks of importance beyond the curated set of identity terms in the final model trained on two datasets (de Gibert et al., 2018; Kennedy et al., 2018), with that of most identity terms decreasing, and some aggressive words increasing, such as “destroys”, “poisoned”. Vaidya et al. (2019) used a similar multitask learning framework to Waseem et al. (2018) on *Kaggle*, but with the classification of author’s identity as the auxiliary task to mitigate the confusion between identity keywords and hateful reference. Similarly, Xia et al. (2020) incorporated the prediction of African-American English dialect in their loss term, but this was done after an initial pre-training of the hate speech classification alone.

There is little consensus in how bias and the effect of bias mitigation should be measured, with different studies adopting varying “debiased” metrics, including Error Rate Equality Difference (Dixon et al., 2018; Park et al., 2018; Nozza et al., 2019), pinned AUC Equality Difference (Dixon et al., 2018; Badjatiya et al., 2019), Pinned Bias (Badjatiya et al., 2019), synthetic test set AUC (Park et al., 2018), and weighted average of subgroup AUCs (Nozza et al., 2019; Vaidya et al., 2019). More importantly, such metrics are all defined based on how the subgroups are defined – which datasets are used, which social groups are compared, which keywords or predictive models are chosen to categorise those groups. As a consequence, although such metrics provide quantitative comparison between different mitigation strategies within a study, the results are hard to compare horizontally. Nonetheless, a common pattern is found across the studies: the standard metric, such as raw F1 or AUC, and the “debiased” metrics seldom improve at the same time. This raises the question on the relative importance that should be put on “debiased” metrics and widely accepted raw metrics: How much practical value do such debiased metrics have if they contradict raw metrics? Or do we need to rethink the widely accepted AUC and F1 scores on benchmark datasets because they do not reflect the toll on minority groups?

In comparison, Sap et al. (2019) proposed to address the biases of human annotators during dataset building, rather than debiasing already annotated data or regularising models. By including each tweet’s dialect and providing **extra annotation instructions** to think of tweet dialect as a proxy of the author’s ethnic identity, they managed to significantly reduce the likelihood of the largely white annotator group (75%) to rate an African-American English tweet offensive to anyone or to themselves. This approach bears similarity to Vaidya et al. (2019)’s, which also sought to distinguish identity judgement from offensiveness spotting, although in automatic models. Although on a small scale, this study demonstrated that more care can be put into annotator instructions than existing datasets have.

## Hate Expression Can Be Implicit

In contrast, implicit expressions are an obstacle to generalisability that comes from the nature of hate speech, and is arguably the trickiest to address. Compared to explicit, which is more transferable between datasets (Nejadgholi and Kiritchenko, 2020), implicit poses challenges to generalisation through interacting with the aforementioned two obstacles: In implicit expressions, there are fewer lexical features to be learnt, and limited, biased data further magnify the challenge of learning generalisable features; implicit hate expressions diverge from standard language use even further than social media or explicit hate speech.

Slurs and profanity are common in hate speech. This is partly why keywords are widely used as a proxy to identify hate speech in existing datasets. However, hate can also be expressed through stereotypes (Sap et al., 2020), sarcasm, irony, humour, and metaphor (Mishra et al., 2019; Vidgen et al., 2019). For example, a post that reads “Hey Brianne - get in the kitchen and make me a samich. Chop Chop” (Gao



and Huang, 2017) *directly attacks* a woman *based on* her female *identity* using stereotypes, and thus certainly fulfills the definition of hate speech, without any distinctive keyword.

Implicit hate speech conveys the same desire to distance such social groups as explicit hate speech (Alorainy et al., 2019) and are no less harmful (Breitfeller et al., 2019). Implicit expressions are the most commonly mentioned cause of false negatives in error analysis (Zhang and Luo, 2018; Qian et al., 2018; Basile et al., 2019; Mozafari et al., 2020a). Inability to detect nuanced, implicit expressions of hate means the models do not go beyond lexical features and cannot capture the underlying hateful intent, let alone generalise to hate speech cases where there are no recurring hate-related words and phrases. Because of the reliance on lexical features, automatic detection models fall far short of human’s ability to detect hate and are thus far from being applicable in the real world as a moderation tool (Duarte et al., 2018).

It has been proposed that abusive language should be systematically classified into explicit and implicit, as well as generalised and directed (Waseem et al., 2017). Several subsequent studies have also identified nuanced, implicit expression as a particularly important challenge in hate speech detection for future research to address (van Aken et al., 2018; Duarte et al., 2018; Swamy et al., 2019). It is especially necessary for explainability (Mishra et al., 2019). Despite the wide recognition of the problem, there has been much fewer attempts at addressing it.

### Existing solutions

Implicit cases of hate speech are hard to identify because they can be understood only within their specific context or with the help of relevant real-world knowledge such as stereotypes. Some have thus **included context in datasets**. For example, Gao and Huang (2017) included the original news articles as the context of the comments. de Gibert et al. (2018)’s hate speech forum dataset organised sentences in the same post together, and has a “relation” label separate from “hate”/“no hate” to set apart cases which can only be correctly understood with its neighbours.

Offensive or abusive language datasets that include implicitness in annotation schemes have appeared only recently. The *AbuseEval* dataset (Caselli et al., 2020) is so far the only **dataset with a standalone “implicit” label**. They re-annotated the *OLID* dataset (Zampieri et al., 2019a), splitting the offensive class into implicitly abusive, explicitly abusive, and non-abusive. Their dataset thus offered a clearer distinction between abusiveness and offensiveness, and between implicit and explicit abuse. Sap et al. (2020) asked annotators to explicitly **paraphrase the implied statements** of intentionally offensive posts. The task defined by this dataset is thus very different from previously existing ones – it is a sequence-to-sequence task to generate implied statements on top of the classification task to identify hateful intent.

Both of their experiments reveal that predicting implicit abuse or biases remains a major challenge. Sap et al. (2020)’s model tended to output the most generic bias of each social group, rather than the implied bias in each post. Caselli et al. (2020)’s best model achieved only a precision of around .234 and a recall of 0.098 for the implicit class, in contrast to .864 and .936 for non-abusive and .640 and .509 for explicit.

To the best of our knowledge, so far there has only been one attempt at annotating the implicitness of hate speech specifically. Alatawi et al. (2020) crowd-sourced annotation on a small set of tweets collected through white supremacist hashtags and user names, dividing them into implicit white supremacism, explicit white supremacism, other hate, and neutral. Unfortunately, the inter-annotator agreement was so low (0.11 Cohen’s kappa (Cohen, 1960)) that they reduced the labels into binary (hateful vs non-hateful) in the end. The main disagreements are between neutral and implicit labels. Compared to Sap et al. (2020) and Caselli et al. (2020)’s studies, their result highlights the difficulty of annotating implicit hate speech and, more fundamentally, the perception of hate speech largely depends on the reader, as posited by Waseem (2016).

Fewer studies proposed **model design motivated by implicit hate speech**. Gao et al. (2017) designed a novel two-path model, aiming to capture both explicit hate speech with a “slur learner” path and implicit hate speech with an LSTM path. However, it is doubtful whether the LSTM path really learns to identify implicit hate speech, as it is also trained on hate speech cases acquired through initial slur-matching and the slur learner.

Targeting specific types of implicit hate speech seems more effective. Alorainy et al. (2019) developed a feature set using dependency trees, part-of-speech tags, and pronouns, to capture the us vs them sentiment in implicit hate speech. This improved classification performance on a range of classifiers including CNN-GRU and LSTM. The main shortcoming is that the performance gain was relative to unprocessed training data, so it is not clear how effective this feature set is compared to common pre-processing

663 methods.

## 664 DISCUSSION

665 While cross-dataset testing can be a useful tool for measuring generalisability, it is important not to reduce  
666 the study of generalisability in hate speech detection to cross-dataset performance or “debiased” metrics.  
667 Ultimately, we want generalisability to the real world. Why we are developing these models and datasets,  
668 how we intend to use them, and what potential impacts they may have on the users and the wider society  
669 are all worth keeping in mind. While mathematical metrics offer quantification, our focus should always  
670 be on what we plan to address and its context. Furthermore, hate speech datasets and models should be  
671 representative of what hate speech is with no prioritising of any facets of it (Swamy et al., 2019), and  
672 should not discriminate against minority groups that they are intended to protect (Davidson et al., 2019).

673 Hate speech detection as a sub-field of NLP is rather new. Despite the help of established NLP  
674 methods, achieving consensus in the formulation of the problem is still ongoing work – whether it is  
675 binary, multi-class, hierarchical, how to source representative data, what metadata should be included,  
676 and where we draw the line between offensive and hateful content. Thus, no existing dataset qualifies  
677 as a “benchmark dataset” yet (Swamy et al., 2019). In the near future, it is likely that new datasets will  
678 continue to emerge and shape our understanding of how to study hate speech computationally. Thus,  
679 while it is important to try to solve the problems defined by existing datasets, more emphasis should be  
680 put on generalisability.

## 681 Future Research

682 Generalisability is a complex problem concerning every aspect of hate speech detection – dataset building,  
683 model training and evaluation, and application. Thus, obstacles to generalisable hate speech detection are  
684 largely intertwined.

685 In the “obstacles” section above, we analysed the problem of generalisability and discussed existing  
686 research, organised by obstacles and their causes. Here, we suggest what can practically be done moving  
687 forward, from the specific perspectives of dataset and models, as well as other general challenges. These  
688 suggestions vary by problem complexity and generality. Nonetheless, they are all, in our opinion, critical  
689 things to keep in mind for any researcher working on hate speech detection to evaluate and improve  
690 generalisability.

## 691 Datasets

### 692 Clear label definitions

693 Unclear and different definitions surrounding hate speech lead to inconsistencies in the literature,  
694 and creates sampling and annotation biases and disparity between datasets, which in turn harms the  
695 generalisability of models trained on such data. Thus, a prerequisite is to have clear label definitions.

696 Hate speech should be separated from other types of offensive language (Davidson et al., 2017; Founta  
697 et al., 2018), and abusive language from offensive language (Caselli et al., 2020). In addition to this, to  
698 address the ambiguity between types of abusive language, future datasets can cover a wider spectrum  
699 of abusive language such as personal attacks, trolling, and cyberbullying. This could be done either in a  
700 hierarchical manner like what Basile et al. (2019) and Kumar et al. (2018b) did with subtypes of hate  
701 speech and aggression respectively, or in a multi-label manner, as there might be cases where more than  
702 one can apply, as seen in Waseem and Hovy (2016)’s racism and sexism labels. At the same time, the  
703 definitions of labels should have as little overlap as possible.

### 704 Annotation quality

705 Related to clear label definitions, ensuring annotation quality would help improve generalisation  
706 by reducing the gaps between datasets and between annotations within each dataset. Guidelines range  
707 from brief descriptions of each class to long paragraphs of definitions and examples (Table 5). Yet, only  
708 about two thirds of the existing datasets report inter-annotator agreement rates (Poletto et al., 2020).  
709 There exists a trade-off between creating a larger dataset with the help of external workers and having  
710 high-quality annotations that reflect a precise and informed understanding of hate speech. High-quality,  
711 expert-produced annotations can help produce better models (Caselli et al., 2020). At the same time, extra  
712 guidelines were shown to be effective in addressing some of the biases in crowd-sourced annotations (Sap  
713 et al., 2019). Future research can look into what type of, and how much, training or instruction is required  
714 to match the annotations of crowdworkers and experts.

## Understanding perception

With annotation quality, another very different approach can be taken – understanding why the perception of hate diverges across annotators. This can not only improve generalisability through addressing disparity in annotations, but also help evaluate potential representation biases and disentangle implicit expressions of hate.

While clear definitions and guidelines are worth pursuing, how each individual perceives hate speech is bound to be different depending on their background (Waseem, 2016). Thus, annotator disagreement will be inevitable even with the same guidelines and training. Instead of aggregating labels into a gold standard, an alternative way of looking at such disagreement is that it reflects an actual divergence of opinions and are all valid (Basile, 2020).

More research can be done to understand why and when disagreement arises, quantitatively or qualitatively. This can be done through building datasets with annotator attributes and their judgements. Existing datasets mostly reported the number of annotators and whether they are crowdworkers, but seldom the demographics of annotators. Furthermore, within the range of “expert” annotators, there are also many possibilities, such as the authors themselves (de Gibert et al., 2018; Mandl et al., 2019), experts in linguistics (Kumar et al., 2018a), activists (Waseem, 2016; Waseem and Hovy, 2016), experts in politics (Vidgen et al., 2020). By training models on different sets of annotations, unintended biases in models can also be better understood. Annotating implicit hate speech is especially challenging (Alatawi et al., 2020). Through improved understanding of hate speech perception, an implicit hate speech dataset could be made possible.

## Drawing representative samples

Before the annotation process, sampling approaches can introduce bias into the dataset and affect the proportion of implicit cases, both affecting the practical value of a detection model. Drawing more representative samples can help with generalisation through alleviating these two problems.

Abusive content represent less than 3% of social media (Zampieri et al., 2019b; Founta et al., 2018), all datasets use simple heuristics to boost the proportion of the positive label. It is a better approach to start with an initial sample and then apply boosting techniques to increase the proportion of abuse posts, compared to drawing a filtered sample using offensive keywords from the beginning (Wiegand et al., 2019; Razo and Kübler, 2020). Boosting techniques can also be improved, by shifting away from keywords towards other less lexical proxies of possible hate, to reduce the emphasis on explicit hate in the dataset. Future datasets should also actively address different types of possible biases, such as regularising each user’s contribution to one dataset, analysis of the topics present in the dataset, limiting the association between certain terms or language styles and a label. It will also help to measure sampling bias quantitatively (Ousidhoum et al., 2020).

## Models

### Reducing overfitting

Overfitting harms model generalisability in any task, but the small and biased hate speech datasets magnify this problem. In addition to the dataset building process, it can be addressed through reducing model overfitting.

Overfitting can be reduced through training on more than one dataset (Waseem et al., 2018; Karan and Šnajder, 2018) or transfer learning from a larger dataset (Uban and Dinu, 2019; Alatawi et al., 2020) and/or a closely related task, such as sentiment analysis (Uban and Dinu, 2019; Cao et al., 2020), yet synthesis in the literature is lacking. More work can be done on comparing different training approaches, and what characteristics of the datasets interact with the effectiveness. For example, when performing transfer learning, the trade-off between domain-specificity and dataset size and representativeness is worth investigating.

Reducing the reliance on lexical features can also help alleviate overfitting to the training dataset. Domain knowledge such as linguistic patterns and underlying sentiment of hate speech can inform model design, feature extraction or preprocessing (Alorainy et al., 2019). Future studies can look into how features of different nature can be effectively combined.

### Debiasing models

Unintended representation biases threaten the practicality of applying automatic hate speech detection on unseen real-world data. Model debiasing can be carried out in conjunction with the improvement and understanding of data collection and annotation.

A range of approaches could be used to make the model less biased against certain terms or language styles, from the perspectives of training data or objective. Each study shows that their approach takes some effect, yet comparison across studies is still difficult. More systematic comparisons between debiasing approaches is favourable. This can be done by applying a range of existing approaches on a number of datasets, with a set of consistent definitions of attributes. There could also be an interaction between debiasing approaches and the types of biases. When experimenting with “debiasing”, it is important to always stay critical of any metrics used.

### Model application and impact

Also related to real-world application, extra care needs to be taken with model evaluation, when addressing any of the obstacles mentioned above.

To realistically evaluate model performance, dataset-wise mathematical metrics like F1/AUC should not be the only measurement. It is also important to evaluate models also on datasets not seen during training (Wiegand et al., 2019), and carry out in-depth error analysis relevant to any specific challenge that the model claims to address. Evaluation methods that are aware of different possible perceptions of hate are also desirable (Basile, 2020).

Furthermore, machine learning models should be considered as part of a sociotechnical system, instead of an algorithm which only exists in relation to the input and outcomes (Selbst et al., 2019). Thus, more future work can be put into studying hate speech detection models in a wider context of application. For example, can automatic models practically aid human moderators in content moderation? In that case, how can human moderators make use of the outputs or post-hoc feature analysis (e.g. Kennedy et al. (2020)) most effectively? Would that introduce more bias or reduce bias in content moderation? Would such effects differ across different hate expressions? What would the impact be on the users of the platform? To answer these questions, interdisciplinary collaboration is needed.

### Other general challenges

Finally, in addition to the specific challenges regarding data and models mentioned above, these general efforts should be made in parallel:

- **Open-sourcing.** Experimental studies on generalisation require access to a variety of resources, data and models as a prerequisite. Furthermore, only with detailed annotation guidelines and model source code made public, detailed inspection into factors that affect generalisability can be enabled. Even without a focus on generalisation per se, easier access to evaluation data and models to compare to can help shift hate speech detection research, as a whole, towards more generalisable outputs. Thus, a joint effort on open-sourcing should be made.
- **Multilingual research.** English has a disproportionate representation in available hate speech data and existing hate speech detection research. The ubiquity of hate speech in any language and culture calls for more work on lower-resource languages in hate speech research. So far, all generalisation studies that mentioned language consider it as a dimension for generalisation. Such an approach can help address the challenge the scarcity of non-English data, if, for example, models trained on English annotated data only can work well on another language. Cross-lingual generalisation is thus practically valuable. On the other hand, there exists a limit to such an “extreme” type of generalisation, determined by language and culture dissimilarity and varying social events. Thus, future contribution to cross-lingual generalisation can be two-folds: increasing cross-lingual performance through model and dataset development, probing the limit of cross-lingual performance through in-depth analysis.

## CONCLUSION

Existing hate speech detection models generalise poorly on new, unseen datasets. Cross-dataset testing is a useful tool to more realistically evaluate model generalisation performance, but the problem of generalisability does not stop there. Reasons why generalisable hate speech detection is hard come from limits of existing NLP methods, dataset building, and the nature of online hate speech, and are often intertwined. The behaviour of social media users and especially haters poses extra challenge to established NLP methods. Small datasets make deep learning models prone to overfitting, and biases in datasets transfer to models. While some biases come from different sampling methods or definitions, others merely reflect long-standing biases in our society. Hate speech evolves with time and context, and thus

has a lot of variation in expression. Existing attempts to address these challenges span across adapting state-of-the-art in other NLP tasks, refining data collection and annotation, and drawing inspirations from domain knowledge of hate speech. More work can be done in these directions to increase generalisability in two main directions: data and models. At the same time, wider context and impact should be carefully considered. Open-sourcing and multilingual research are also important.

## REFERENCES

- Agrawal, S. and Awekar, A. (2018). Deep learning for detecting cyberbullying across multiple social media platforms. In *European Conference on Information Retrieval*, pages 141–153. Springer.
- Al-Hassan, A. and Al-Dossari, H. (2019). Detection of Hate Speech in Social Networks: a Survey on Multilingual Corpus. In *Computer Science & Information Technology (CS & IT)*, pages 83–100. AIRCC Publishing Corporation.
- Alatawi, H. S., Alhothali, A. M., and Moria, K. M. (2020). Detecting White Supremacist Hate Speech using Domain Specific Word Embedding with Deep Learning and BERT. *arXiv:2010.00357 [cs]*. arXiv: 2010.00357.
- Alorainy, W., Burnap, P., Liu, H., and Williams, M. L. (2019). “The Enemy Among Us”: Detecting Cyber Hate Speech with Threats-based Othering Language Embeddings. *ACM Transactions on the Web*, 13(3):1–26.
- Arango, A., Pérez, J., and Poblete, B. (2020). Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, page 101584.
- Badjatiya, P., Gupta, M., and Varma, V. (2019). Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In Liu, L., White, R. W., Mantrach, A., Silvestri, F., McAuley, J. J., Baeza-Yates, R., and Zia, L., editors, *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 49–59. ACM.
- Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760.
- Banko, M., MacKeen, B., and Ray, L. (2020). A unified taxonomy of harmful content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 125–137, Online. Association for Computational Linguistics.
- Basile, V. (2020). It’s the End of the Gold Standard as we Know it. On the Impact of Pre-aggregation on the Evaluation of Highly Subjective Tasks. In *CEUR Workshop Proceedings*, page 10.
- Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., and Sanguinetti, M. (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.
- Baziotis, C., Pelekis, N., and Doukeridis, C. (2017). DataStories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada. Association for Computational Linguistics.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.
- Blodgett, S. L., Green, L., and O’Connor, B. (2016). Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130.
- Blodgett, S. L. and O’Connor, B. (2017). Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English. *arXiv:1707.00061 [cs]*. arXiv: 1707.00061.
- Bodapati, S., Gella, S., Bhattacharjee, K., and Al-Onaizan, Y. (2019). Neural word decomposition models for abusive language detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 135–145, Florence, Italy. Association for Computational Linguistics.
- Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Bolukbasi, T., Chang, K., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing*

- 875 *Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10,*  
876 *2016, Barcelona, Spain, pages 4349–4357.*
- 877 Breitfeller, L., Ahn, E., Jurgens, D., and Tsvetkov, Y. (2019). Finding microaggressions in the wild: A  
878 case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on*  
879 *Empirical Methods in Natural Language Processing and the 9th International Joint Conference on*  
880 *Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China. Association  
881 for Computational Linguistics.
- 882 Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial  
883 gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91.
- 884 Cao, R., Lee, R. K.-W., and Hoang, T.-A. (2020). DeepHate: Hate Speech Detection via Multi-Faceted  
885 Text Representations. In *12th ACM Conference on Web Science, WebSci '20*, pages 11–20, New York,  
886 NY, USA. Association for Computing Machinery.
- 887 Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75.
- 888 Caselli, T., Basile, V., Mitrović, J., Kartoziya, I., and Granitzer, M. (2020). I feel offended, don't be  
889 abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the 12th*  
890 *Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European  
891 Language Resources Association.
- 892 Caselli, T., Basile, V., Mitrović, J., and Granitzer, M. (2021). HateBERT: Retraining BERT for Abusive  
893 Language Detection in English. *arXiv:2010.12472 [cs]*. arXiv: 2010.12472.
- 894 Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes,  
895 M., Yuan, S., Tar, C., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder for English. In  
896 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System*  
897 *Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.
- 898 Chen, H., McKeever, S., and Delany, S. J. (2018). A Comparison of Classical Versus Deep Learning  
899 Techniques for Abusive Content Detection on Social Media Sites. In Staab, S., Koltsova, O., and  
900 Ignatov, D. I., editors, *Social Informatics*, Lecture Notes in Computer Science, pages 117–133, Cham.  
901 Springer International Publishing.
- 902 Chen, H., McKeever, S., and Delany, S. J. (2019). The Use of Deep Learning Distributed Representations  
903 in the Identification of Abusive Text. *Proceedings of the International AAAI Conference on Web and*  
904 *Social Media*, 13:125–133.
- 905 Chung, Y.-L., Kuzmenko, E., Tekiroglu, S. S., and Guerini, M. (2019). CONAN - COUNTER NARRATIVES  
906 through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of*  
907 *the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence,  
908 Italy. Association for Computational Linguistics.
- 909 Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological*  
910 *measurement*, 20(1):37–46.
- 911 Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M.,  
912 Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale.  
913 In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages  
914 8440–8451.
- 915 Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without  
916 parallel data. *arXiv preprint arXiv:1710.04087*.
- 917 Daumé III, H. (2007). Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Annual Meeting*  
918 *of the Association of Computational Linguistics*, pages 256–263.
- 919 Davidson, T., Bhattacharya, D., and Weber, I. (2019). Racial bias in hate speech and abusive language  
920 detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35,  
921 Florence, Italy. Association for Computational Linguistics.
- 922 Davidson, T., Warmusley, D., Macy, M., and Weber, I. (2017). Automated Hate Speech Detection and the  
923 Problem of Offensive Language. *arXiv:1703.04009 [cs]*. arXiv: 1703.04009.
- 924 de Gibert, O., Perez, N., García-Pablos, A., and Cuadros, M. (2018). Hate speech dataset from a white  
925 supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages  
926 11–20, Brussels, Belgium. Association for Computational Linguistics.
- 927 Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional  
928 Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North*  
929 *American Chapter of the Association for Computational Linguistics: Human Language Technologies*,

- Volume 1 (Long and Short Papers), pages 4171–4186.
- Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. (2018). Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, New Orleans LA USA. ACM.
- Duarte, N., Llanos, E., and Loup, A. (2018). Mixed Messages? The Limits of Automated Social Media Content Analysis. In *Conference on Fairness, Accountability and Transparency*, pages 106–106.
- Fersini, E., Nozza, D., and Rosso, P. (2018a). Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). In Caselli, T., Novielli, N., Patti, V., and Rosso, P., editors, *Evalita Evaluation of NLP and Speech Tools for Italian*, pages 59–66. Accademia University Press.
- Fersini, E., Nozza, D., and Rosso, P. (2020). AMI @ EVALITA2020: Automatic Misogyny Identification. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, page 8.
- Fersini, E., Rosso, P., and Anzovino, M. (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)*, pages 214–228.
- Fortuna, P. and Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, 51(4):1–30.
- Fortuna, P., Soler, J., and Wanner, L. (2020). Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.
- Fortuna, P., Soler-Company, J., and Wanner, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524.
- Founta, A.-M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., and Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of ICWSM*. AAAI Press.
- Gao, L. and Huang, R. (2017). Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266, Varna, Bulgaria. INCOMA Ltd.
- Gao, L., Kuppertschmidt, A., and Huang, R. (2017). Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 774–782, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Gilbert, C. and Hutto, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*. Available at (20/04/16) <http://comp.social.gatech.edu/papers/icwsml4.vader.hutto.pdf>, volume 81, page 82.
- Glavaš, G., Karan, M., and Vulić, I. (2020). XHate-999: Analyzing and detecting abusive language across domains and languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6350–6365, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Golbeck, J., Ashktorab, Z., Banjo, R. O., Berlinger, A., Bhagwan, S., Buntain, C., Cheakalos, P., Geller, A. A., Gergory, Q., Gnanasekaran, R. K., Gunasekaran, R. R., Hoffman, K. M., Hottle, J., Jienjilt, V., Khare, S., Lau, R., Martindale, M. J., Naik, S., Nixon, H. L., Ramachandran, P., Rogers, K. M., Rogers, L., Sarin, M. S., Shahane, G., Thanki, J., Vengataraman, P., Wan, Z., and Wu, D. M. (2017). A Large Labeled Corpus for Online Harassment Research. In *Proceedings of the 2017 ACM on Web Science Conference, WebSci '17*, pages 229–233, New York, NY, USA. Association for Computing Machinery.
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. <http://www.deeplearningbook.org>.
- Gröndahl, T., Pajola, L., Juuti, M., Conti, M., and Asokan, N. (2018). All you need is” love” evading hate speech detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, pages 2–12.
- Halevy, A., Norvig, P., and Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12.
- Heinzerling, B. and Strube, M. (2018). BPEmb: Tokenization-free pre-trained subword embeddings in

- 275 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2989–2993.
- Hovy, D. and Spruit, S. L. (2016). The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.
- Huang, X., Xing, L., Dernoncourt, F., and Paul, M. J. (2020). Multilingual Twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1440–1448, Marseille, France. European Language Resources Association.
- Indurthi, V., Syed, B., Shrivastava, M., Chakravartula, N., Gupta, M., and Varma, V. (2019). FERMI at SemEval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 70–74, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Jigsaw (2018). Toxic comment classification challenge. <https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge>.
- Jin, X., Wei, Z., Du, J., Xue, X., and Ren, X. (2020). Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Joulin, A., Grave, É., Bojanowski, P., and Mikolov, T. (2017). Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431.
- Karan, M. and Šnajder, J. (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.
- Kennedy, B., Atari, M., Davani, A. M., Yeh, L., Omrani, A., Kim, Y., Coombs, K., Havaldar, S., Portillo-Wightman, G., Gonzalez, E., Hoover, J., Azatian, A., Hussain, A., Lara, A., Olmos, G., Omary, A., Park, C., Wijaya, C., Wang, X., Zhang, Y., and Dehghani, M. (2018). The Gab Hate Corpus: A collection of 27k posts annotated for hate speech. Technical report, PsyArXiv.
- Kennedy, B., Jin, X., Mostafazadeh Davani, A., Dehghani, M., and Ren, X. (2020). Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.
- Kim, J. Y., Ortiz, C., Nam, S., Santiago, S., and Datta, V. (2020). Intersectional Bias in Hate Speech and Abusive Language Datasets. *arXiv:2005.05921 [cs]*. arXiv: 2005.05921.
- Kiritchenko, S. and Mohammad, S. M. (2018). Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of \*SEM*, pages 43–53.
- Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., and Taboada, M. (2019). The SFU Opinion and Comments Corpus: A Corpus for the Analysis of Online News Comments. *Corpus Pragmatics*, 4(2):155–190.
- Koufakou, A., Pamungkas, E. W., Basile, V., and Patti, V. (2020). HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, Online. Association for Computational Linguistics.
- Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2018a). 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.
- Kumar, R., Reganti, A. N., Bhatia, A., and Maheshwari, T. (2018b). Aggression-annotated corpus of Hindi-English code-mixed data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Lee, Y., Yoon, S., and Jung, K. (2018). Comparative studies of detecting abusive language on Twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 101–106, Brussels, Belgium. Association for Computational Linguistics.
- Levy, O. and Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308.



- 1040 Liu, P., Li, W., and Zou, L. (2019a). NULI at SemEval-2019 task 6: Transfer learning for offensive  
1041 language detection using bidirectional transformers. In *Proceedings of the 13th International Workshop*  
1042 *on Semantic Evaluation*, pages 87–91, Minneapolis, Minnesota, USA. Association for Computational  
1043 Linguistics.
- 1044 Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and  
1045 Stoyanov, V. (2019b). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint*  
1046 *arXiv:1907.11692*.
- 1047 Magu, R. and Luo, J. (2018). Determining code words in euphemistic hate speech using word embedding  
1048 networks. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 93–100,  
1049 Brussels, Belgium. Association for Computational Linguistics.
- 1050 Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., and Patel, A. (2019). Overview of  
1051 the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European  
1052 Languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation, FIRE '19*, pages  
1053 14–17, Kolkata, India. Association for Computing Machinery.
- 1054 Meyer, J. S. and Gambäck, B. (2019). A platform agnostic dual-strand hate speech detector. In *Proceedings*  
1055 *of the Third Workshop on Abusive Language Online*, pages 146–156, Florence, Italy. Association for  
1056 Computational Linguistics.
- 1057 Mikolov, T., Grave, E., Bojanowski, P., Puhersch, C., and Joulin, A. (2018). Advances in pre-training  
1058 distributed word representations. In *Proceedings of the Eleventh International Conference on Language*  
1059 *Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association  
1060 (ELRA).
- 1061 Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of  
1062 words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*,  
1063 pages 3111–3119.
- 1064 Mishra, P., Yannakoudakis, H., and Shutova, E. (2018). Neural character-based composition models for  
1065 abuse detection. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages  
1066 1–10, Brussels, Belgium. Association for Computational Linguistics.
- 1067 Mishra, P., Yannakoudakis, H., and Shutova, E. (2019). Tackling Online Abuse: A Survey of Automated  
1068 Abuse Detection Methods. *arXiv:1908.06024 [cs]*. arXiv: 1908.06024.
- 1069 Mishra, S. and Mishra, S. (2019). 3Idiots at HASOC 2019: Fine-tuning Transformer Neural Networks for  
1070 Hate Speech Identification in Indo-European Languages. *FIRE*, page 6.
- 1071 Mozafari, M., Farahbakhsh, R., and Crespi, N. (2020a). A BERT-Based Transfer Learning Approach  
1072 for Hate Speech Detection in Online Social Media. In Cherifi, H., Gaito, S., Mendes, J. F., Moro, E.,  
1073 and Rocha, L. M., editors, *Complex Networks and Their Applications VIII*, Studies in Computational  
1074 Intelligence, pages 928–940, Cham. Springer International Publishing.
- 1075 Mozafari, M., Farahbakhsh, R., and Crespi, N. (2020b). Hate speech detection and racial bias mitigation  
1076 in social media based on BERT model. *PLOS ONE*, 15(8):e0237861. Publisher: Public Library of  
1077 Science.
- 1078 Nejadgholi, I. and Kiritchenko, S. (2020). On cross-dataset generalization in automatic detection of online  
1079 abuse. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 173–183, Online.  
1080 Association for Computational Linguistics.
- 1081 Nobata, C., Tetreault, J. R., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection  
1082 in online user content. In Bourdeau, J., Hendler, J., Nkambou, R., Horrocks, I., and Zhao, B. Y., editors,  
1083 *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada,*  
1084 *April 11 - 15, 2016*, pages 145–153. ACM.
- 1085 Nozza, D., Volpetti, C., and Fersini, E. (2019). Unintended Bias in Misogyny Detection. In  
1086 *IEEE/WIC/ACM International Conference on Web Intelligence, WI '19*, pages 149–155, New York,  
1087 NY, USA. Association for Computing Machinery.
- 1088 Ousidhoum, N., Song, Y., and Yeung, D.-Y. (2020). Comparative evaluation of label-agnostic selection  
1089 bias in multilingual hate speech datasets. In *Proceedings of the 2020 Conference on Empirical Methods*  
1090 *in Natural Language Processing (EMNLP)*, pages 2532–2542, Online. Association for Computational  
1091 Linguistics.
- 1092 Pamungkas, E. W., Basile, V., and Patti, V. (2020). Misogyny Detection in Twitter: a Multilingual and  
1093 Cross-Domain Study. *Information Processing & Management*, 57(6):102360.
- 1094 Pamungkas, E. W. and Patti, V. (2019). Cross-domain and cross-lingual abusive language detection: A

- hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370, Florence, Italy. Association for Computational Linguistics.
- Park, J. H. (2018). Finding Good Representations of Emotions for Text Classification. *arXiv:1808.07235 [cs]*. arXiv: 1808.07235.
- Park, J. H., Shin, J., and Fung, P. (2018). Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.
- Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Pinter, Y., Guthrie, R., and Eisenstein, J. (2017). Mimicking Word Embeddings using Subword RNNs. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 102–112.
- Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2020). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*.
- Qian, J., ElSherief, M., Belding, E., and Wang, W. Y. (2018). Leveraging intra-user and inter-user representation learning for automated hate speech detection. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 118–123, New Orleans, Louisiana. Association for Computational Linguistics.
- Razavi, A. H., Inkpen, D., Uritsky, S., and Matwin, S. (2010). Offensive Language Detection Using Multi-level Classification. In Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Farzindar, A., and Kešelj, V., editors, *Advances in Artificial Intelligence*, volume 6085, pages 16–27. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.
- Razo, D. and Kübler, S. (2020). Investigating sampling bias in abusive language detection. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 70–78, Online. Association for Computational Linguistics.
- Sanguinetti, M., Cassidy, L., Bosco, C., Çetinoğlu, Ö., Cignarella, A. T., Lynn, T., Rehbein, I., Ruppenhofer, J., Seddah, D., and Zeldes, A. (2020). Treebanking User-Generated Content: a UD Based Overview of Guidelines, Corpora and Unified Recommendations. *arXiv:2011.02063 [cs]*. arXiv: 2011.02063.
- Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. (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.
- Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., and Choi, Y. (2020). Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.
- Schmidt, A. and Wiegand, M. (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.
- Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT\* ’19*, pages 59–68, New York, NY, USA. Association for Computing Machinery.
- Serrà, J., Leontiadis, I., Spathis, D., Stringhini, G., Blackburn, J., and Vakali, A. (2017). Class-based prediction errors to detect hate speech with out-of-vocabulary words. In *Proceedings of the First Workshop on Abusive Language Online*, pages 36–40, Vancouver, BC, Canada. Association for Computational Linguistics.
- Sharma, S., Agrawal, S., and Shrivastava, M. (2018). Degree based classification of harmful speech using Twitter data. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 106–112, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a

- simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.
- Swamy, S. D., Jamatia, A., and Gambäck, B. (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.
- Tatman, R. (2017). Gender and dialect bias in youtube’s automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59.
- Taylor, J., Peignon, M., and Chen, Y.-S. (2017). Surfacing contextual hate speech words within social media. *arXiv:1711.10093 [cs]*. arXiv: 1711.10093.
- Uban, A.-S. and Dinu, L. P. (2019). On Transfer Learning for Detecting Abusive Language Online. In Rojas, I., Joya, G., and Catala, A., editors, *Advances in Computational Intelligence*, Lecture Notes in Computer Science, pages 688–700, Cham. Springer International Publishing.
- Vaidya, A., Mai, F., and Ning, Y. (2019). Empirical Analysis of Multi-Task Learning for Reducing Model Bias in Toxic Comment Detection. *arXiv:1909.09758 [cs]*. arXiv: 1909.09758.
- van Aken, B., Risch, J., Krestel, R., and Löser, A. (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.
- Vidgen, B. and Derczynski, L. (2020). Directions in Abusive Language Training Data: Garbage In, Garbage Out. *arXiv:2004.01670 [cs]*. arXiv: 2004.01670.
- Vidgen, B., Hale, S., Guest, E., Margetts, H., Broniatowski, D., Waseem, Z., Botelho, A., Hall, M., and Tromble, R. (2020). Detecting East Asian prejudice on social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 162–172, Online. Association for Computational Linguistics.
- Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., and Margetts, H. (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.
- Warner, W. and Hirschberg, J. (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.
- Waseem, Z. (2016). Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.
- Waseem, Z., Davidson, T., Warmusley, D., and Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.
- Waseem, Z. and Hovy, D. (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.
- Waseem, Z., Thorne, J., and Bingel, J. (2018). Bridging the Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection. In Golbeck, J., editor, *Online Harassment*, Human–Computer Interaction Series, pages 29–55. Springer International Publishing, Cham.
- Wiedemann, G., Yimam, S. M., and Biemann, C. (2020). UHH-LT at SemEval-2020 task 12: Fine-tuning of pre-trained transformer networks for offensive language detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1638–1644, Barcelona (online). International Committee for Computational Linguistics.
- Wiegand, M., Ruppenhofer, J., and Kleinbauer, T. (2019). Detection of Abusive Language: the Problem of Biased Datasets. 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 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.
- Wieting, J., Bansal, M., Gimpel, K., and Livescu, K. (2015). From paraphrase database to compositional paraphrase model and back. *Transactions of the Association for Computational Linguistics*, 3:345–358.
- Wulczyn, E., Thain, N., and Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In Barrett, R., Cummings, R., Agichtein, E., and Gabrilovich, E., editors, *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 1391–1399. ACM.

- 1205 Xia, M., Field, A., and Tsvetkov, Y. (2020). Demoting racial bias in hate speech detection. In *Proceedings*  
1206 *of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14,  
1207 Online. Association for Computational Linguistics.
- 1208 Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019a). Predicting the  
1209 type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North*  
1210 *American Chapter of the Association for Computational Linguistics: Human Language Technologies,*  
1211 *Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for  
1212 Computational Linguistics.
- 1213 Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019b). SemEval-2019  
1214 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings*  
1215 *of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota,  
1216 USA. Association for Computational Linguistics.
- 1217 Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis,  
1218 Z., and Çöltekin, Ç. (2020). SemEval-2020 task 12: Multilingual offensive language identification in  
1219 social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*,  
1220 pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.
- 1221 Zhang, Z. and Luo, L. (2018). Hate Speech Detection: A Solved Problem? The Challenging Case of  
1222 Long Tail on Twitter. *arXiv:1803.03662 [cs]*. arXiv: 1803.03662.
- 1223 Zhang, Z., Robinson, D., and Tepper, J. (2018). Detecting Hate Speech on Twitter Using a Convolution-  
1224 GRU Based Deep Neural Network. In Gangemi, A., Navigli, R., Vidal, M.-E., Hitzler, P., Troncy, R.,  
1225 Hollink, L., Tordai, A., and Alam, M., editors, *The Semantic Web*, Lecture Notes in Computer Science,  
1226 pages 745–760, Cham. Springer International Publishing.
- 1227 Zhao, R., Zhou, A., and Mao, K. (2016). Automatic detection of cyberbullying on social networks based  
1228 on bullying features. In *Proceedings of the 17th international conference on distributed computing and*  
1229 *networking*, pages 1–6.