PeerJ Computer Science

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

Wenjie Yin and Arkaitz Zubiaga

School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom

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 and the reasons 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.

Subjects Artificial Intelligence, Computational Linguistics, Data Mining and Machine Learning, Natural Language and Speech, Social Computing
Keywords Hate speech, Text classification, Abusive language, Social media, Literature review, Generalisation

INTRODUCTION

The Internet saw a growing body of user-generated content as social media platforms flourished (*Schmidt & Wiegand, 2017*; *Chung et al., 2019*). While social media provides a platform for all users to freely express themselves, offensive and harmful contents 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 & Hovy, 2016*; *Davidson et al., 2017*; *Founta et al., 2018*; *Sharma, Agrawal & Shrivastava, 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, Kuppersmith & Huang, 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, McKeever & Delany, 2019*).

Submitted 3 November 2020
Accepted 26 May 2021
Published 17 June 2021

Corresponding author
Wenjie Yin, w.yin@qmul.ac.uk

Academic editor
Yilun Shang

Additional Information and Declarations can be found on page 27

DOI **10.7717/peerj-cs.598**

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 & 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 & 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 subtypes of offensive language. For example, personal attacks (*Wulczyn, Thain & Dixon, 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, Zhou & Mao, 2016*), which is carried out repeatedly and over time against vulnerable victims that cannot defend themselves.[1] This paper focuses on hate speech and hate speech datasets, although studies that cover both hate speech and other offensive language are also mentioned.

[1] For a more elaborate comparison between similar concepts, see *Fortuna & Nunes (2018)*; *Poletto et al. (2020)*; *Banko, MacKeen & Ray (2020)*

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, Bengio & Courville, 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, Jamatia & Gambäck, 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, Prez & Poblete, 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 (*Gröndahl et al., 2018*; *Swamy, Jamatia & Gambäck, 2019*; *Wiegand, Ruppenhofer & Kleinbauer, 2019*; *Fortuna, Soler-Company & Wanner, 2021*), apart from a handful of studies that made individual attempts at addressing aspects of it (*Karan & Šnajder, 2018*; *Waseem, Thorne & Bingel, 2018*; *Arango, Prez & Poblete, 2020*). Recent survey papers on hate speech and abusive language detection (*Schmidt & Wiegand, 2017*; *Fortuna & Nunes, 2018*; *Al-Hassan & Al-Dossari, 2019*; *Mishra, Yannakoudakis & Shutova, 2019*; *Vidgen et al., 2019*; *Poletto et al., 2020*; *Vidgen & Derczynski, 2020*) have focused on the general trends in this field, mainly by comparing features, algorithms and datasets. Among these, *Fortuna & 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 & Derczynski (2020)* created extensive lists of resources and benchmark corpora while *Al-Hassan & 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".

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 & Wiegand, 2017*; *Fortuna & Nunes, 2018*; *Al-Hassan & Al-Dossari, 2019*; *Mishra, Yannakoudakis & Shutova, 2019*; *Vidgen et al., 2019*; *Poletto et al., 2020*; *Vidgen & 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, Ruppenhofer & Kleinbauer, 2019*). This evaluation method is called cross-dataset testing (*Swamy, Jamatia & Gambäck, 2019*) or cross-application (*Gröndahl et al., 2018*), and sometimes cross-domain classification (*Wiegand, Ruppenhofer & Kleinbauer, 2019*) or detection (*Karan & 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

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 & Derczynski (2020)* and *Poletto et al. (2020)*.

Dataset name	Publication	Source	Positive labels	Annotator type
Waseem	*Waseem & 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	Expert and crowdsourcing
Kaggle	*Jigsaw (2018)*	Wikipedia	Toxic Severe toxic Obscene Threat Insult Identity hate	Crowdsourcing
Gao	*Gao & Huang (2017)*	Fox News	Hateful	? (Native speakers)
AMI	*Fersini, Nozza & Rosso (2018)* *Fersini, Rosso & Anzovino (2018)*	Twitter	Misogynous	Expert
Warner	*Warner & Hirschberg (2012)*	Yahoo! American Jewish Congress	Anti-semitic Anti-black Anti-asian Anti-woman Anti-muslim Anti-immigrant Other-hate	? (Volunteer)
Zhang	*Zhang, Robinson & Tepper (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, Thain & Dixon (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

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.

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 & Kiritchenko (2020)* used different datasets for training and testing; *Fortuna, Soler & Wanner (2020)* compared datasets through class vector representations.

		Study											
Dataset name	**Type**	*Karan & Šnajder (2018)*	*Gröndahl et al. (2018)*	*Waseem (2016)*	*Wiegand, Ruppenhofer & Kleinbauer (2019)*	*Swamy, Jamatia & Gambäck (2019)*	*Pamungkas & Patti (2019)*, *Pamungkas, Basile & Patti (2020)*	*Arango, Prez & Poblete (2020)*	*Fortuna, Soler & Wanner (2020)*	*Caselli et al. (2020)*	*Nejadgholi & Kiritchenko (2020)*	*Glavaš, Karan & Vulić (2020)*	*Fortuna, Soler-Company & Wanner (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, ALBERT, fastText, SVM

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*).

For example, the CNN-GRU model by *Zhang, Robinson & Tepper (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, Prez & Poblete (2020)* showed a similar F1 drop of around 30 points 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, Prez & Poblete (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, Robinson & Tepper (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, Jamatia & Gambäck, 2019*) and on the leaderboards of shared tasks, such as *Zampieri et al. (2020)*; *Fersini, Nozza & Rosso (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, Jamatia & Gambäck, 2019*), which is less drastic compared to earlier state-of-the-art neural networks tested in other studies (*Gröndahl et al., 2018*; *Arango, Prez & Poblete, 2020*). *Pamungkas, Basile & Patti (2020)* and *Fortuna, Soler-Company & Wanner (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, Thorne & Bingel, 2018*), Support Vector Machines (*Karan & Šnajder, 2018*; *Fortuna, Soler-Company & Wanner, 2021*; *Pamungkas & Patti, 2019*; *Pamungkas, Basile & Patti, 2020*), and shallow networks with pre-trained embeddings, e.g., MLP with Byte-Pair Encoding (BPE)-based subword embeddings (*Heinzerling & Strube, 2018*; *Waseem, Thorne & Bingel, 2018*) and FastText (*Joulin et al., 2017a*; *Wiegand, Ruppenhofer & Kleinbauer, 2019*; *Fortuna, Soler-Company & Wanner, 2021*).

Generally, these simpler models do not perform as good as deep neural networks, such as LSTM (*Pamungkas & Patti, 2019*) and especially BERT and its variants (*Pamungkas, Basile & Patti, 2020*; *Fortuna, Soler-Company & Wanner, 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 & Patti, 2019*; *Pamungkas, Basile & Patti, 2020*), and FastText when comparing to BERT (*Fortuna, Soler-Company & Wanner, 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, Prez & Poblete, 2020*; *Pamungkas & Patti, 2019*; *Pamungkas, Basile & Patti, 2020*; *Fortuna, Soler-Company & Wanner, 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 & Šnajder, 2018*), yet some datasets are more similar to each other. Furthermore, certain attributes of a dataset could lead to more generalisable models.

Similarity between datasets varies, as there are groups of datasets that produce models that test much better on each other. For example, in *Wiegand, Ruppenhofer & Kleinbauer (2019)*'s study, FastText models (*Joulin et al., 2017a*) trained on three datasets *(Kaggle, Founta, Razavi)* achieved F1 scores above 70 when tested on one another, while models trained or tested on datasets outside this group achieved around 60 or less. In *Swamy, Jamatia & Gambäck (2019)*'s study with fine-tuned BERT models (*Devlin et al., 2019*), *Founta* and *OLID* produced models that performed well on each other. The source of such differences are usually traced back to search terms (*Swamy, Jamatia & Gambäck, 2019*), topics covered (*Nejadgholi & Kiritchenko, 2020*; *Pamungkas, Basile & Patti, 2020*), label definitions (*Pamungkas & Patti, 2019*; *Pamungkas, Basile & Patti, 2020*; *Fortuna,*

Soler-Company & Wanner, 2021), and data source platforms (*Glavaš, Karan & Vulić, 2020*; *Karan & Šnajder, 2018*).

Another way of looking at generalisation and similarity is by comparing differences **between individual classes** across datasets (*Nejadgholi & Kiritchenko, 2020*; *Fortuna, Soler & Wanner, 2020*; *Fortuna, Soler-Company & Wanner, 2021*), as opposed to comparing datasets as a whole. In both *Nejadgholi & Kiritchenko (2020)* and *Fortuna, Soler-Company & Wanner (2021)*'s experiments, the best generalisation is achieved for more general labels such as "toxicity", "offensive", or "abusive". Generalisation is not as good for finer-grained hate speech labels. All in all, these findings are indicative of an imbalance of the finer-grained subclasses, particularly owing to disagreements in the definition of what constitutes hate speech, which proves more difficult than defining what constitutes offensive language.

Within the hate speech labels, the relative similarity also varies. *Fortuna, Soler & Wanner (2020)* used averaged word embeddings (*Bojanowski et al., 2017*; *Mikolov et al., 2018*) to compute the representations of classes from different datasets, and compared classes across datasets. One of their observations is that *Davidson*'s "hate speech" is very different from *Waseem*'s "hate speech", "racism", "sexism", while being relatively close to *HatEval*'s "hate speech" and *Kaggle*'s "identity hate". This echoes with experiments that showed poor generalisation of models from *Waseem* to *HatEval* (*Arango, Prez & Poblete, 2020*) and between *Davidson* and *Waseem* (*Waseem, Thorne & Bingel, 2018*; *Gröndahl et al., 2018*).

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

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

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

A larger training **data size** is generally believed to produce better and more generalisable models (*Halevy, Norvig & Pereira, 2009*). It is mentioned as one of the two biggest factors

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

Study	Data translation	Embedding and classifier models	Additional languages
Pamungkas & Patti (2019)	Training, automatic	MUSE embeddings, LSTM, "joint-learning"	Spanish, Italian, German
Pamungkas, Basile & Patti (2020)	Training, automatic	MUSE embeddings, mBERT, LSTM, SVM, "joint-learning"	Spanish, Italian
Glavaš, Karan & Vulić (2020)	Testing, manual	mBERT, XLM-R	Albanian, Croatian, German, Russian, Turkish
Arango, Prez & Poblete (2020)	Testing, automatic	MUSE embeddings, LSTM, GBDT	Spanish
Fortuna, Soler-Company & Wanner (2021)	None	mBERT	Italian, Spanish, Portuguese

contributing to cross-dataset performance in *Karan & Šnajder (2018)*'s study. *Caselli et al. (2020)* also found that, on *HatEval*, their dataset (*AbuseEval*) produced a model even better-performing than the one trained on *HatEval* end-to-end. They partially attributed this to a bigger data size, alongside **annotation quality**. However, the benefit of having more data is counterbalanced by data distribution differences (*Karan & Šnajder, 2018*), as discussed above. Moreover, its relative importance compared to other factors seems to be small, when the latter are carefully controlled (*Nejadgholi & Kiritchenko, 2020*; *Fortuna, Soler-Company & Wanner, 2021*).

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 & Derczynski, 2020*).

Cross-lingual generalisation can be considered a more "extreme" type of generalisation (*Arango, Prez & Poblete, 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, Basile & Patti, 2020*; *Glavaš, Karan & Vulić, 2020*; *Arango, Prez & Poblete, 2020*; *Fortuna, Soler-Company & Wanner, 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, Soler-Company & Wanner (2021)* hold that multilingual generalisation per se is likely to be worse than its monolingual counterpart, while *Arango, Prez & Poblete (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 displayed in the results of the cross-lingual generalisation studies mentioned above (*Pamungkas & Patti, 2019*; *Pamungkas, Basile & Patti, 2020*; *Glavaš, Karan & Vulić, 2020*; *Arango, Prez & Poblete, 2020*; *Fortuna, Soler-Company & Wanner, 2021*) and 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, Basile & Patti, 2020*) and GBDT (*Arango, Prez & Poblete, 2020*) compared to LSTM. Adding automatically translated training data alongside the original is beneficial (*Pamungkas & Patti, 2019*; *Pamungkas, Basile & Patti, 2020*).
- When it comes to the **data**, the most prominent additional factor compared to the monolingual setting is the similarity between the training (source) and testing (target) languages. For instance, Among the wide range of languages that *Glavaš, Karan & Vulić (2020)* have tested, the cross-lingual performance drop between English, the source language, and German, the most similar target language, was less than one third of that between English and Turkish, when using mBERT on *Wulczyn*.

Although these studies more or less consider the "language" and "domain" variables as separate, there exists evidence that the two types of generalisation interact with each other. Studies that control the language variable more carefully tend to show a smaller drop across languages—for example, by manually translating exactly the same data (*Glavaš, Karan & Vulić, 2020*), as opposed to using automatic translation (*Pamungkas & Patti, 2019*; *Pamungkas, Basile & Patti, 2020*; *Arango, Prez & Poblete, 2020*) or different language dataset from the same shared task (*Pamungkas & Patti, 2019*; *Pamungkas, Basile & Patti, 2020*; *Fortuna, Soler-Company & Wanner, 2021*). Furthermore, adding data from a different domain can act as a regulariser from overfitting to the training language (*Glavaš, Karan & Vulić, 2020*).

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

OBSTACLES TO GENERALISABLE HATE SPEECH DETECTION

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

Non-standard grammar and vocabulary

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

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

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

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

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

Existing solutions

From a domain-specific perspective, *Taylor, Peignon & Chen (2017)* and *Magu & Luo (2018)* attempted to **identify code words** for slurs used in hate communities. Both of them

used keyword search as part of their sourcing of Twitter data and word embedding models to model word relationships. *Taylor, Peignon & Chen (2017)* identified hate communities through Twitter connections of the authors of extremist articles and hate speech keyword searches. They trained their own dependency2vec (*Levy & 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 & 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, Peignon & Chen (2017)* and *Magu & 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 & Gambäck (2019)* found hybrid and character CNN to perform best respectively. Probably most surprisingly, *Lee, Yoon & Jung (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 & Huang, 2017*; *Lee, Yoon & Jung, 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, Guthrie & Eisenstein, 2017*)-based model displayed similar performances (*Mishra, Yannakoudakis & Shutova, 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, McKeever & Delany, 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, Li & Zou, 2019*; *Mishra & 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, Llanso & Loup, 2018*). In hate speech detection, word embeddings trained on relevant data (social media or news sites) had a clear advantage (*Chen, McKeever & Delany, 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, Bengio & Courville, 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, McKeever & Delany, 2018*; *Mishra, Yannakoudakis & Shutova, 2018*; *Bodapati et al., 2019*).

In *Swamy, Jamatia & Gambäck (2019)*'s study of model generalisability, **abusive language-specific pre-trained embeddings** were suggested as a possible solution to limited dataset sizes. *Alatawi, Alhothali & Moria (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, Socher & Manning, 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 generalisablity, 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., 2019*), *Wiedemann, Yimam & Biemann (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, Yimam & Biemann (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 & Dinu (2019)* pre-trained a classification model on a large sentiment dataset (https://help.sentiment140.com/), 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, Lee & Hoang (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 & Hutto, 2014*); their model was deeper and used general word embeddings (*Mikolov et al., 2013*; *Pennington, Socher & Manning, 2014*; *Wieting et al., 2015*) and topic representation computed through Latent Dirichlet Allocation (LDA) (*Blei, Ng & Jordan, 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.

Table 4 Boosted sampling methods of the most commonly studied hate speech datasets (*Waseem & Hovy, 2016*; *Davidson et al., 2017*; *Founta et al., 2018*; *Basile et al., 2019*). Description as appeared in the publications.

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 (https://www.hatebase.org/)	"Each user from lexicon search"	N/A
Founta	HateBase, NoSwearing (https://www.noswearing.com/dictionary/)	N/A	Negative sentiment
HatEval	"Neutral keywords and derogatory words against the targets, highly polarized hashtags"	"Identified haters"	"Potential victims of hate accounts"

Notes.
N/A, no relevant descriptions found.

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

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

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

Moreover, different studies are based on varying definitions of "hate speech", as seen in different **annotation guidelines** (Table 5). Despite all covering the same two main aspects (directly attack or promote hate towards), datasets vary by their wording, what they consider a target (any group, minority groups, specific minority groups), and their clarifications on edge cases. *Davidson* and *HatEval* both distinguished "hate speech" from "offensive language", while "uses a sexist or racist slur" is in *Waseem*'s guidelines to mark a case positive of hate, blurring the boundary of offensive and hateful. Additionally, as both *HatEval* and *Waseem* specified the types of hate (towards women and immigrants; racism and sexism), hate speech that fell outside of these specific types were not included in the positive classes, while *Founta* and *Davidson* included any type of hate speech. Guidelines also differ in how detailed they are: Apart from *Founta*, all other datasets started the annotation process with sets of labels pre-defined by the authors, among which *Waseem* gave the most specific description of actions. In contrast, *Founta* only provided annotators with short conceptual definitions of a range of possible labels, allowing more freedom for a

Table 5 **Annotation guidelines of the most commonly studied hate speech datasets.**

Dataset	Action	Target	Clarifications
Waseem	Attacks, seeks to silence, criticises, *negatively stereotypes, promotes hate speech or violent crime, blatantly misrepresents truth or seeks to distort views on*, 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	*Express hatred towards*, humiliate, insult *	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	*Express hatred towards*, humiliate, insult	Individual or group, on the basis of attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender	N/A
HatEval	*Spread, incite, promote, justify hatred or violence towards*, dehumanizing, hurting or intimidating **	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

Notes.

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).

first exploratory round of annotation. After that, labels were finalised, and another round of annotation was carried out. As a result, the labelling reflects how the general public, without much domain knowledge or extensive training, would classify offensive language. For example, the "abusive" and "offensive" classes were so similar that they were merged in the second stage. However, as discussed above, they differ by whether intentionality is present (*Caselli et al., 2020*). Such different annotation and labelling criteria result in essentially different tasks and different training objectives, despite their data having a lot in common.

As a result of the varying and sampling methods, definitions, and annotation schemes, what current models can learn on one dataset is specific to the examples in that dataset

and the task defined by the dataset, limiting the models' ability to generalise to new data. One type of possible resulting bias is **author bias**. For example, 65% of the hate speech in the *Waseem* dataset was produced by merely two users, and their tweets exist in both the training and the test set. Models trained on such data thus overfit to these users' language styles. This overfitting to authors was proven in two state-of-the-art models (*Badjatiya et al., 2017*; *Agrawal & Awekar, 2018*; *Arango, Prez & Poblete, 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, Ruppenhofer & Kleinbauer, 2019*). Such biases can also be measured through the semantic similarity between keywords used for building the datasets and topics present in the dataset (*Ousidhoum, Song & Yeung, 2020*).

Existing solutions

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

Waseem, Thorne & Bingel (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 & 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 & 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 & 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, Bhattacharya & Weber, 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, Shin & Fung, 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, Bhattacharya & Weber, 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, Shin & Fung, 2018*; *Nozza, Volpetti & Fersini, 2019*). Alternatively, bias-prone terms could be identified through looking at skewed distributions within a specific dataset (*Badjatiya, Gupta & Varma, 2019*; *Mozafari, Farahbakhsh & Crespi, 2020b*).

A few studies measured biases across directly **predicted language styles or demographic attributes** of authors. *Davidson, Bhattacharya & Weber (2019)* and *Kim et al. (2020)* both tested their hate speech detection models on *Blodgett, Green & OConnor (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 & 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, Shin & Fung (2018)* created examples containing the counterpart of gendered terms found in the data to address gender bias in the *Waseem* and *Founta* datasets. *Badjatiya, Gupta & Varma (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, Shin & Fung (2018)* also compared pre-trained debiased word embeddings (*Bolukbasi et al., 2016*) and transfer learning from a larger, less biased corpus. Similarly, *Nozza, Volpetti & Fersini (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, Farahbakhsh & Crespi (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, Mai & Ning (2019)* used a similar multitask learning framework to *Waseem, Thorne & Bingel (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, Field & Tsvetkov (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, Shin & Fung, 2018*; *Nozza, Volpetti & Fersini, 2019*), pinned AUC Equality Difference (*Dixon et al., 2018*; *Badjatiya, Gupta & Varma, 2019*), Pinned Bias (*Badjatiya, Gupta & Varma, 2019*), synthetic test set AUC (*Park, Shin & Fung, 2018*), and weighted average of subgroup AUCs (*Nozza, Volpetti & Fersini, 2019*; *Vaidya, Mai & Ning, 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, Mai & Ning (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

Implicit expressions are an obstacle to generalisability that comes from the nature of hate speech, and is arguably the trickiest to address. Compared to explicity, which is more transferable between datasets (*Nejadgholi & Kiritchenko, 2020*), implicity 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, Yannakoudakis & Shutova, 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 & Huang, 2017*) *directly attacks* a woman *based on* her female *identity* using stereotypes, fufilling 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 & Luo, 2018*; *Qian et al., 2018*; *Basile et al., 2019*; *Mozafari, Farahbakhsh & Crespi, 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, Llanso & Loup, 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, Llanso & Loup, 2018*; *Swamy, Jamatia & Gambäck, 2019*). It is especially necessary for explainability (*Mishra, Yannakoudakis & Shutova, 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 & 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, Alhothali & Moria (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, Kuppersmith & Huang (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 methods.

DISCUSSION

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

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

Future research

Generalisability is a complex problem concerning every aspect of hate speech detection—dataset building, model training and evaluation, and application. Thus, obstacles to generalisabile hate speech detection are largely intertwined.

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

Datasets

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

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

Annotation quality. Related to clear label definitions, ensuring annotation quality would help improve generalisation by reducing the gaps between datasets and between annotations within each dataset. Guidelines range from brief descriptions of each class to long paragraphs of definitions and examples (Table 5). Yet, only about two thirds of the existing datasets report inter-anntotator agreement rates (*Poletto et al., 2020*). There exists a trade-off between creating a larger dataset with the help of external workers and having high-quality annotations that reflect a precise and informed understanding of hate speech. High-quality, expert-produced annotations can help produce better models (*Caselli et al., 2020*). At the same time, extra guidelines were shown to be effective in addressing some of the biases in crowd-sourced annotations (*Sap et al., 2019*). Future research can look into what type of, and how much, training or instruction is required 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 of the papers (*de Gibert et al., 2018*; *Mandl et al., 2019*), experts in linguistics (*Kumar et al., 2018a*), activists (*Waseem, 2016*; *Waseem & 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, Alhothali & Moria, 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*), so 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 abusive posts, compared to drawing a filtered sample using offensive keywords from the beginning (*Wiegand, Ruppenhofer & Kleinbauer, 2019*; *Razo & 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, Song & Yeung, 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, Thorne & Bingel, 2018*; *Karan & Šnajder, 2018*) or transfer learning from a larger dataset (*Uban & Dinu, 2019*; *Alatawi, Alhothali & Moria, 2020*) and/or a closely related task, such as sentiment analysis (*Uban & Dinu, 2019*; *Cao, Lee & Hoang, 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 would be helpful. 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, Ruppenhofer & Kleinbauer, 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, it is only with detailed annotation guidelines and model source code made public that 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 detection 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.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Wenjie Yin is funded by the School of Electronic Engineering and Computer Science, Queen Mary University of London. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
School of Electronic Engineering and Computer Science, Queen Mary University of London.

Competing Interests

Arkaitz Zubiaga is an Academic Editor for PeerJ Computer Science.

Author Contributions

- Wenjie Yin conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

- Arkaitz Zubiaga conceived and designed the experiments, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The paper is a literature review that did not involve raw data or code.

REFERENCES

Agrawal S, Awekar A. 2018. Deep learning for detecting cyberbullying across multiple social media platforms. In: *European conference on information retrieval*. Grenoble, France: Springer, 141–153.

Al-Hassan A, Al-Dossari H. 2019. Detection of hate speech in social networks: a survey on multilingual corpus. In: *Computer Science & Information Technology (CS & IT)*. Chennai, India: AIRCC Publishing Corporation, 83–100.

Alatawi HS, Alhothali AM, Moria KM. 2020. Detecting white supremacist hate speech using domain specific word embedding with deep learning and BERT. ArXiv preprint. arXiv:2010.00357.

Alorainy W, Burnap P, Liu H, Williams ML. 2019. The enemy among us: detecting cyber hate speech with threats-based othering language embeddings. *ACM Transactions on the Web* **13(3)**:1–26 DOI 10.1145/3324997.

Arango A, Pérez J, Poblete B. 2020. Hate speech detection is not as easy as you may think: a closer look at model validation (extended version). *Information Systems* Epub ahead of print 2020 30 June DOI 10.1016/j.is.2020.101584.

Badjatiya P, Gupta M, Varma V. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In: Liu L, White RW, Mantrach A, Silvestri F, McAuley JJ, Baeza-Yates R, Zia L, eds. *The World Wide Web conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*. ACM, 49–59 DOI 10.1145/3308558.3313504.

Badjatiya P, Gupta S, Gupta M, Varma V. 2017. Deep learning for hate speech detection in tweets. In: *Proceedings of the 26th international conference on World Wide Web companion*. 759–760.

Banko M, MacKeen B, Ray L. 2020. A unified taxonomy of harmful content. In: *Proceedings of the fourth workshop on online abuse and harms*. Association for Computational Linguistics, 125–137 DOI 10.18653/v1/2020.alw-1.16.

Basile V. 2020. It's the end of the gold standard as we know itOn the impact of pre-aggregation on the evaluation of highly subjective tasks. In: *CEUR workshop proceedings*. 10.

Basile V, Bosco C, Fersini E, Nozza D, Patti V, Rangel Pardo FM, Rosso P, 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*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 54–63 DOI 10.18653/v1/S19-2007.

Baziotis C, Pelekis N, Doulkeridis 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)*. Vancouver, Canada: Association for Computational Linguistics, 747–754 DOI 10.18653/v1/S17-2126.

Blei DM, Ng AY, Jordan MI. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* **3(Jan)**:993–1022.

Blodgett SL, Green L, OConnor 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*. 1119–1130.

Blodgett SL, O'Connor B. 2017. Racial disparity in natural language processing: a case study of social media African-American English. ArXiv preprint. arXiv:1707.00061.

Bodapati S, Gella S, Bhattacharjee K, Al-Onaizan Y. 2019. Neural word decomposition models for abusive language detection. In: *Proceedings of the third workshop on abusive language online*. Florence, Italy: Association for Computational Linguistics, 135–145 DOI 10.18653/v1/W19-3515.

Bojanowski P, Grave E, Joulin A, Mikolov T. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* **5**:135–146 DOI 10.1162/tacl_a_00051.

Bolukbasi T, Chang K, Zou JY, Saligrama V, Kalai AT. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Lee DD, Sugiyama M, von Luxburg U, Guyon I, Garnett R, eds. *Advances in neural information processing systems 29: Annual conference on neural information processing systems 2016, December 5-10, 2016, Barcelona, Spain*. 4349–4357.

Breitfeller L, Ahn E, Jurgens D, Tsvetkov Y. 2019. Finding microaggressions in the wild: a case for locating elusive phenomena in social media posts. In: *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 1664–1674 DOI 10.18653/v1/D19-1176.

Buolamwini J, Gebru T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: *Conference on fairness, accountability and transparency*. 77–91.

Cao R, Lee RK.-W, Hoang T.-A. 2020. DeepHate: hate speech detection via multi-faceted text representations. In: *12th ACM conference on web science*. New York, NY, USA: Association for Computing Machinery, 11–20.

Caruana R. 1997. Multitask learning. *Machine Learning* **28(1)**:41–75 DOI 10.1023/A:1007379606734.

Caselli T, Basile V, Mitrović J, Kartoziya I, Granitzer M. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In: *Proceedings of the 12th language resources and evaluation conference*. Marseille, France: European Language Resources Association, 6193–6202.

Caselli T, Basile V, Mitrovi J, Granitzer M. 2021. HateBERT: retraining BERT for abusive language detection in english. ArXiv preprint. arXiv:2010.12472.

Cer D, Yang Y, Kong S-y, Hua N, Limtiaco N, StJohn R, Constant N, Guajardo-Cespedes M, Yuan S, Tar C, Strope B, Kurzweil R. 2018. Universal sentence encoder for english. In: *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations*. Brussels, Belgium: Association for Computational Linguistics, 169–174 DOI 10.18653/v1/D18-2029.

Chen H, McKeever S, Delany SJ. 2018. A comparison of classical versus deep learning techniques for abusive content detection on social media sites. In: Staab S, Koltsova O, Ignatov DI, eds. *Social informatics. Lecture notes in computer science*, Cham: Springer International Publishing, 117–133.

Chen H, McKeever S, Delany SJ. 2019. The use of deep learning distributed representations in the identification of abusive text. *Proceedings of the International AAAI Conference on Web and Social Media* **13**:125–133.

Chung Y-L, Kuzmenko E, Tekiroglu SS, Guerini M. 2019. CONAN - Counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In: *Proceedings of the 57th annual meeting of the association for computational linguistics*. Florence, Italy: Association for Computational Linguistics, 2819–2829 DOI 10.18653/v1/P19-1271.

Cohen J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20(1)**:37–46 DOI 10.1177/001316446002000104.

Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave É, Ott M, Zettlemoyer L, Stoyanov V. 2020. Unsupervised cross-lingual representation learning at scale. In: *Proceedings of the 58th annual meeting of the association for computational linguistics*. 8440–8451.

Conneau A, Lample G, Ranzato M, Denoyer L, Jégou H. 2017. Word translation without parallel data. ArXiv preprint. arXiv:1710.04087.

Daumé III H. 2007. Frustratingly easy domain adaptation. In: *Proceedings of the 45th annual meeting of the association of computational linguistics*. 256–263.

Davidson T, Bhattacharya D, Weber I. 2019. Racial bias in hate speech and abusive language detection datasets. Florence: Association for Computational Linguistics, 25–35 DOI 10.18653/v1/W19-3504.

Davidson T, Warmsley D, Macy M, Weber I. 2017. Automated hate speech detection and the problem of offensive language. ArXiv preprint. arXiv:1703.04009.

De Gibert O, Perez N, García-Pablos A, Cuadros M. 2018. Hate speech dataset from a white supremacy forum. In: *Proceedings of the 2nd workshop on abusive language online (ALW2)*. Brussels, Belgium: Association for Computational Linguistics, 11–20 DOI 10.18653/v1/W18-5102.

Devlin J, Chang M-W, Lee K, Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: human language technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

Dixon L, Li J, Sorensen J, Thain N, Vasserman L. 2018. Measuring and mitigating unintended bias in text classification. In: *Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society*. New Orleans LA USA: ACM, 67–73.

Duarte N, Llanso E, Loup A. 2018. Mixed messages? The limits of automated social media content analysis. In: *Conference on fairness, accountability and transparency*. 106–106.

Fersini E, Nozza D, Rosso P. 2018. Overview of the Evalita 2018 task on automatic misogyny identification (AMI). In: Caselli T, Novielli N, Patti V, Rosso P, eds. *EVALITA evaluation of NLP and speech tools for Italian*. Accademia University Press, 59–66.

Fersini E, Nozza D, 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)*. 8.

Fersini E, Rosso P, Anzovino M. 2018. Overview of the Task on Automatic Misogyny Identication at IberEval 2018. In: *Proceedings of the third workshop on evaluation of human language technologies for iberian languages (IberEval 2018)*. Seville, Spain: 214–228.

Fortuna P, Nunes S. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys* **51(4)**:1–30 DOI 10.1145/3232676.

Fortuna P, Soler J, 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*. Marseille, France: European Language Resources Association, 6786–6794.

Fortuna P, Soler-Company J, Wanner L. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management* **58(3)**:102524 DOI 10.1016/j.ipm.2021.102524.

Founta A-M, Djouvas C, Chatzakou D, Leontiadis I, Blackburn J, Stringhini G, Vakali A, Sirivianos M, Kourtellis N. 2018. Large scale crowdsourcing and characterization of Twitter abusive behavior. In: *Proceedings of ICWSM*. AAAI Press,.

Gao L, 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*. Varna: INCOMA Ltd, 260–266 DOI 10.26615/978-954-452-049-6_036.

Gao L, Kuppersmith A, 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)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, 774–782.

Gilbert C, 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), volume 81*. 82. *Available at http:// comp. socialgatechedu/ papers/ icwsm14vaderhuttopdf* .

Glavaš G, Karan M, Vulić I. 2020. XHate-999: analyzing and detecting abusive language across domains and languages. In: *Proceedings of the 28th international conference on*

computational linguistics. Barcelona, Spain: International Committee on Computational Linguistics, 6350–6365 DOI 10.18653/v1/2020.coling-main.559.

Golbeck J, Ashktorab Z, Banjo RO, Berlinger A, Bhagwan S, Buntain C, Cheakalos P, Geller AA, Gergory Q, Gnanasekaran RK, Gunasekaran RR, Hoffman KM, Hottle J, Jienjitlert V, Khare S, Lau R, Martindale MJ, Naik S, Nixon HL, Ramachandran P, Rogers KM, Rogers L, Sarin MS, Shahane G, Thanki J, Vengataraman P, Wan Z, Wu DM. 2017. A large labeled corpus for online harassment research. In: *Proceedings of the 2017 ACM on web science conference*. New York, NY, USA: Association for Computing Machinery, 229–233.

Goodfellow I, Bengio Y, Courville A. 2016. Deep learning. Cambridge: MIT Press.

Gröndahl T, Pajola L, Juuti M, Conti M, Asokan N. 2018. All You Need is" Love" evading hate speech detection. In: *Proceedings of the 11th ACM workshop on artificial intelligence and security*. 2–12.

Halevy A, Norvig P, Pereira F. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems* **24(2)**:8–12.

Heinzerling B, 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)*. 2989–2993.

Hovy D, Spruit SL. 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)*. Berlin, Germany: Association for Computational Linguistics, 591–598 DOI 10.18653/v1/P16-2096.

Huang X, Xing L, Dernoncourt F, Paul MJ. 2020. Multilingual twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In: *Proceedings of the 12th language resources and evaluation conference*. Marseille, France: European Language Resources Association, 1440–1448.

Indurthi V, Syed B, Shrivastava M, Chakravartula N, Gupta M, 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*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 70–74 DOI 10.18653/v1/S19-2009.

Jigsaw . 2018. Toxic comment classification challenge. *Available at https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge* .

Jin X, Wei Z, Du J, Xue X, 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, 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*. 427–431.

Karan M, Šnajder J. 2018. Cross-domain detection of abusive language online. In: *Proceedings of the 2nd workshop on abusive language online (ALW2)*. Brussels, Belgium: Association for Computational Linguistics, 132–137 DOI 10.18653/v1/W18-5117.

Kennedy B, Atari M, Davani AM, 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, Dehghani M. 2018. The Gab Hate Corpus: A collection of 27k posts annotated for hate speech. Technical report, PsyArXiv. DOI 10.31234/osf.io/hqjxn.

Kennedy B, Jin X, Mostafazadeh Davani A, Dehghani M, Ren X. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In: *Proceedings of the 58th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, 5435–5442 DOI 10.18653/v1/2020.acl-main.483.

Kim JY, Ortiz C, Nam S, Santiago S, Datta V. 2020. Intersectional bias in hate speech and abusive language datasets. ArXiv preprint. arXiv:2005.05921.

Kiritchenko S, Mohammad SM. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In: *Proceedings of *SEM*. 43–53.

Kolhatkar V, Wu H, Cavasso L, Francis E, Shukla K, Taboada M. 2019. The SFU opinion and comments corpus: a corpus for the analysis of online news comments. *Corpus Pragmatics* **4(2)**:155–190 DOI 10.1007/s41701-019-00065-w.

Koufakou A, Pamungkas EW, Basile V, 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*. Association for Computational Linguistics, 34–43 DOI 10.18653/v1/2020.alw-1.5.

Kumar R, Ojha AK, Malmasi S, Zampieri M. 2018a. Benchmarking aggression identification in social media. In: *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 1–11.

Kumar R, Reganti AN, Bhatia A, 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, Jung K. 2018. Comparative studies of detecting abusive language on twitter. In: *Proceedings of the 2nd workshop on abusive language online (ALW2)*. Brussels, Belgium: Association for Computational Linguistics, 101–106 DOI 10.18653/v1/W18-5113.

Levy O, Goldberg Y. 2014. Dependency-based word embeddings. In: *Proceedings of the 52nd annual meeting of the association for computational linguistics (Volume 2: Short Papers)*. 302–308.

Liu P, Li W, Zou L. 2019. NULI at SemEval-2019 Task 6: transfer learning for offensive language detection using bidirectional transformers. In: *Proceedings of the 13th international workshop on semantic evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 87–91 DOI 10.18653/v1/S19-2011.

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. 2019. RoBERTa: a robustly optimized BERT pretraining approach. ArXiv preprint. arXiv:1907.11692.

Magu R, Luo J. 2018. Determining code words in euphemistic hate speech using word embedding networks. In: *Proceedings of the 2nd workshop on abusive language online (ALW2)*. Brussels, Belgium: Association for Computational Linguistics, 93–100 DOI 10.18653/v1/W18-5112.

Mandl T, Modha S, Majumder P, Patel D, Dave M, Mandlia C, Patel A. 2019. Overview of the HASOC track at FIRE 2019: hate speech and offensive content identification in indo-european languages. In: *Proceedings of the 11th forum for information retrieval evaluation, FIRE '19*. Kolkata, India: Association for Computing Machinery, 14–17.

Meyer JS, Gambäck B. 2019. A platform agnostic dual-strand hate speech detector. In: *Proceedings of the third workshop on abusive language online*. Florence, Italy: Association for Computational Linguistics, 146–156 DOI 10.18653/v1/W19-3516.

Mikolov T, Grave E, Bojanowski P, Puhrsch C, Joulin A. 2018. Advances in pre-training distributed word representations. In: *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA),.

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. 2013. Distributed representations of words and phrases and their compositionality. In: *Advances in Neural Information Processing Systems*. 3111–3119.

Mishra P, Yannakoudakis H, Shutova E. 2018. Neural character-based composition models for abuse detection. In: *Proceedings of the 2nd workshop on abusive language online (ALW2)*. Brussels, Belgium: Association for Computational Linguistics, 1–10 DOI 10.18653/v1/W18-5101.

Mishra P, Yannakoudakis H, Shutova E. 2019. Tackling online abuse: a survey of automated abuse detection methods. ArXiv preprint. arXiv:1908.06024.

Mishra S, Mishra S. 2019. 3Idiots at HASOC 2019: fine-tuning transformer neural networks for hate speech identication in indo-european languages. *FIRE* 6.

Mozafari M, Farahbakhsh R, Crespi N. 2020a. A BERT-based transfer learning approach for hate speech detection in online social media. In: Cherifi H, Gaito S, Mendes JF, Moro E, Rocha LM, eds. *Complex networks and their applications VIII. Studies in computational intelligence*, Cham: Springer International Publishing, 928–940.

Mozafari M, Farahbakhsh R, Crespi N. 2020b. Hate speech detection and racial bias mitigation in social media based on BERT model. *PLOS ONE* **15(8)**:e0237861 DOI 10.1371/journal.pone.0237861.

Nejadgholi I, Kiritchenko S. 2020. On cross-dataset generalization in automatic detection of online abuse. In: *Proceedings of the fourth workshop on online abuse and harms*. Association for Computational Linguistics, 173–183 DOI 10.18653/v1/2020.alw-1.20.

Nobata C, Tetreault JR, Thomas A, Mehdad Y, Chang Y. 2016. Abusive language detection in online user content. In: Bourdeau J, Hendler J, Nkambou R, Horrocks I, Zhao BY, eds. *Proceedings of the 25th international conference on world wide web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*. ACM, 145–153 DOI 10.1145/2872427.2883062.

Nozza D, Volpetti C, Fersini E. 2019. Unintended bias in misogyny detection. In: *IEEE/WIC/ACM international conference on web intelligence, WI '19*. New York, NY, USA: Association for Computing Machinery, 149–155.

Ousidhoum N, Song Y, Yeung D-Y. 2020. Comparative evaluation of label-agnostic selection bias in multilingual hate speech datasets. In: *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*. Association for Computational Linguistics, 2532–2542 DOI 10.18653/v1/2020.emnlp-main.199.

Pamungkas EW, Basile V, Patti V. 2020. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing & Management* **57(6)**:102360 DOI 10.1016/j.ipm.2020.102360.

Pamungkas EW, 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*. Florence, Italy: Association for Computational Linguistics, 363–370 DOI 10.18653/v1/P19-2051.

Park JH. 2018. Finding good representations of emotions for text classification. ArXiv preprint. arXiv:1808.07235.

Park JH, Shin J, Fung P. 2018. Reducing gender bias in abusive language detection. In: *Proceedings of the 2018 conference on empirical methods in natural language processing*. Brussels, Belgium: Association for Computational Linguistics, 2799–2804 DOI 10.18653/v1/D18-1302.

Pennington J, Socher R, Manning C. 2014. GloVe: global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 1532–1543 DOI 10.3115/v1/D14-1162.

Pinter Y, Guthrie R, Eisenstein J. 2017. Mimicking word embeddings using subword RNNs. In: *Proceedings of the 2017 conference on empirical methods in natural language processing*. Copenhagen, Denmark: Association for Computational Linguistics, 102–112 DOI 10.18653/v1/D17-1010.

Poletto F, Basile V, Sanguinetti M, Bosco C, Patti V. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation* DOI 10.1007/s10579-020-09502-8.

Qian J, ElSherief M, Belding E, Wang WY. 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)*. New Orleans, Louisiana: Association for Computational Linguistics, 118–123 DOI 10.18653/v1/N18-2019.

Razavi AH, Inkpen D, Uritsky S, Matwin S. 2010. Offensive language detection using multi-level classification. In: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Vardi MY, Weikum G, Farzindar A, Keelj V, eds. *Advances*

in artificial intelligence. Vol. 6085. Berlin: Springer Berlin Heidelberg, 16–27. Series Title: Lecture Notes in Computer Science.

Razo D, Kübler S. 2020. Investigating sampling bias in abusive language detection. In: *Proceedings of the fourth workshop on online abuse and harms*. Association for Computational Linguistics, 70–78 DOI 10.18653/v1/2020.alw-1.9.

Sanguinetti M, Cassidy L, Bosco C, Çetinoğlu Ö, Cignarella AT, Lynn T, Rehbein I, Ruppenhofer J, Seddah D, Zeldes A. 2020. Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations. ArXiv preprint. arXiv:2011.02063.

Sap M, Card D, Gabriel S, Choi Y, Smith NA. 2019. The risk of racial bias in hate speech detection. In: *Proceedings of the 57th annual meeting of the association for computational linguistics*. Florence: Association for Computational Linguistics, 1668–1678 DOI 10.18653/v1/P19-1163.

Sap M, Gabriel S, Qin L, Jurafsky D, Smith NA, 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*. Association for Computational Linguistics, 5477–5490 DOI 10.18653/v1/2020.acl-main.486.

Schmidt A, 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*. Valencia, Spain: Association for Computational Linguistics, 1–10 DOI 10.18653/v1/W17-1101.

Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J. 2019. Fairness and abstraction in sociotechnical systems. In: *Proceedings of the conference on fairness, accountability, and transparency, FAT* '19*. New York, NY, USA: Association for Computing Machinery, 59–68.

Serrà J, Leontiadis I, Spathis D, Stringhini G, Blackburn J, 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*. Vancouver, BC, Canada: Association for Computational Linguistics, 36–40 DOI 10.18653/v1/W17-3005.

Sharma S, Agrawal S, 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)*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 106–112.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* **15(1)**:1929–1958.

Swamy SD, Jamatia A, Gambäck B. 2019. Studying generalisability across abusive language detection datasets. In: *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*. Hong Kong, China: Association for Computational Linguistics, 940–950 DOI 10.18653/v1/K19-1088.

Tatman R. 2017. Gender and dialect bias in YouTubes automatic captions. In: *Proceedings of the first ACL workshop on ethics in natural language processing*. 53–59.

Taylor J, Peignon M, Chen Y.-S. 2017. Surfacing contextual hate speech words within social media. ArXiv preprint. arXiv:1711.10093.

Uban A.-S, Dinu LP. 2019. On transfer learning for detecting abusive language online. In: Rojas I, Joya G, Catala A, eds. *Advances in computational intelligence. Lecture notes in computer science*, Cham: Springer International Publishing, 688–700.

Vaidya A, Mai F, Ning Y. 2019. Empirical analysis of multi-task learning for reducing model bias in toxic comment detection. ArXiv preprint. arXiv:1909.09758.

Van Aken B, Risch J, Krestel R, Löser A. 2018. Challenges for toxic comment classification: an in-depth error analysis. Brussels, Belgium: Association for Computational Linguistics, 33–42 DOI 10.18653/v1/W18-5105.

Vidgen B, Derczynski L. 2020. Directions in abusive language training data: garbage in, garbage out. ArXiv preprint. arXiv:2004.01670.

Vidgen B, Hale S, Guest E, Margetts H, Broniatowski D, Waseem Z, Botelho A, Hall M, Tromble R. 2020. Detecting east asian prejudice on social media. In: *Proceedings of the fourth workshop on online abuse and harms*. Association for Computational Linguistics, 162–172 DOI 10.18653/v1/2020.alw-1.19.

Vidgen B, Harris A, Nguyen D, Tromble R, Hale S, Margetts H. 2019. Challenges and frontiers in abusive content detection. In: *Proceedings of the third workshop on abusive language online*. Florence, Italy: Association for Computational Linguistics, 80–93 DOI 10.18653/v1/W19-3509.

Warner W, Hirschberg J. 2012. Detecting hate speech on the world wide web. In: *Proceedings of the second workshop on language in social media*. Montréal, Canada: Association for Computational Linguistics, 19–26.

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*. Austin, Texas: Association for Computational Linguistics, 138–142 DOI 10.18653/v1/W16-5618.

Waseem Z, Davidson T, Warmsley D, Weber I. 2017. Understanding abuse: a typology of abusive language detection subtasks. In: *Proceedings of the first workshop on abusive language online*. Vancouver, BC, Canada: Association for Computational Linguistics, 78–84 DOI 10.18653/v1/W17-3012.

Waseem Z, Hovy D. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: *Proceedings of the NAACL student research workshop*. San Diego, California: Association for Computational Linguistics, 88–93 DOI 10.18653/v1/N16-2013.

Waseem Z, Thorne J, Bingel J. 2018. Bridging the gaps: multi task learning for domain transfer of hate speech detection. In: Golbeck J, ed. *Online Harassment. HumanComputer Interaction Series*, Cham: Springer International Publishing, 29–55.

Wiedemann G, Yimam SM, 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*. Barcelona (online): International Committee for Computational Linguistics, 1638–1644.

Wiegand M, Ruppenhofer J, 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, 602–608 DOI 10.18653/v1/N19-1060.

Wieting J, Bansal M, Gimpel K, Livescu K. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the Association for Computational Linguistics* **3**:345–358 DOI 10.1162/tacl_a_00143.

Wulczyn E, Thain N, Dixon L. 2017. Ex machina: personal attacks seen at scale. In: Barrett R, Cummings R, Agichtein E, Gabrilovich E, eds. *Proceedings of the 26th international conference on world wide web, WWW 2017, Perth, Australia, April 3-7, 2017*. ACM, 1391–1399 DOI 10.1145/3038912.3052591.

Xia M, Field A, Tsvetkov Y. 2020. Demoting racial bias in hate speech detection. In: *Proceedings of the eighth international workshop on natural language processing for social media*. Association for Computational Linguistics, 7–14 DOI 10.18653/v1/2020.socialnlp-1.2.

Zampieri M, Malmasi S, Nakov P, Rosenthal S, Farra N, Kumar R. 2019a. Predicting the type and target of offensive posts in social media. In: *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 1415–1420 DOI 10.18653/v1/N19-1144.

Zampieri M, Malmasi S, Nakov P, Rosenthal S, Farra N, Kumar R. 2019b. SemEval-2019 Task 6: identifying and categorizing offensive language in social media (OffensEval). In: *Proceedings of the 13th international workshop on semantic evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 75–86 DOI 10.18653/v1/S19-2010.

Zampieri M, Nakov P, Rosenthal S, Atanasova P, Karadzhov G, Mubarak H, Derczynski L, Pitenis Z, Çöltekin Ç. 2020. SemEval-2020 Task 12: multilingual offensive language identification in social media (OffensEval 2020). In: *Proceedings of the fourteenth workshop on semantic evaluation*. Barcelona (online): International Committee for Computational Linguistics, 1425–1447.

Zhang Z, Luo L. 2018. Hate speech detection: A solved problem? The challenging case of long tail on twitter. ArXiv preprint. arXiv:1803.03662.

Zhang Z, Robinson D, Tepper J. 2018. Detecting hate speech on twitter using a convolution-GRU based deep neural network. In: Gangemi A, Navigli R, Vidal M.-E, Hitzler P, Troncy R, Hollink L, Tordai A, Alam M, eds. *The Semantic Web. Lecture notes in computer science*, Cham: Springer International Publishing, 745–760.

Zhao R, Zhou A, Mao K. 2016. Automatic detection of cyberbullying on social networks based on bullying features. In: *Proceedings of the 17th international conference on distributed computing and networking*. 1–6.