Information Processing and Management 58 (2021) 102524

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets?

Paula Fortuna [a,*], Juan Soler-Company [a], Leo Wanner [b,a]

[a] Natural Language Processing Group, Department of Communication and Information Technologies, Pompeu Fabra University, Spain
[b] Catalan Institute for Research and Advanced Studies (ICREA), Spain

ARTICLE INFO

MSC:
00-01
99-00

Keywords:
Hate speech
Offensive language
Classification
Generalization

ABSTRACT

A considerable body of research deals with the automatic identification of hate speech and related phenomena. However, cross-dataset model generalization remains a challenge. In this context, we address two still open central questions: (i) to what extent does the generalization depend on the model and the composition and annotation of the training data in terms of different categories?, and (ii) do specific features of the datasets or models influence the generalization potential? To answer (i), we experiment with BERT, ALBERT, fastText, and SVM models trained on nine common public English datasets, whose class (or category) labels are standardized (and thus made comparable), in intra- and cross-dataset setups. The experiments show that indeed the generalization varies from model to model and that some of the categories (e.g., 'toxic', 'abusive', or 'offensive') serve better as cross-dataset training categories than others (e.g., 'hate speech'). To answer (ii), we use a Random Forest model for assessing the relevance of different model and dataset features during the prediction of the performance of 450 BERT, 450 ALBERT, 450 fastText, and 348 SVM binary abusive language classifiers (1698 in total). We find that in order to generalize well, a model already needs to perform well in an intra-dataset scenario. Furthermore, we find that some other parameters are equally decisive for the success of the generalization, including, e.g., the training and target categories and the percentage of the out-of-domain vocabulary.

1. Introduction

The transformation of the internet into a universal communication forum used by nearly everyone has led to a considerable increase of online hate speech, in particular in social media. While in the US, courts have repeatedly ruled that hate speech is covered by the Freedom of Speech Amendment of the US Constitution (Herz & Molnár, 2012), the EU Code of Conduct requires the social media industry to identify and remove hate speech in less than 24 h after its appearance.[1] To combat hate speech, some social media companies (cf. e.g., Facebook and Instagram) automatically remove messages if they appear to be similar enough to the messages in their hate speech reference database. However, a mere removal is often not sufficient in legal terms: some types of hate speech, including, for instance, aggression, threats, racism, etc. are subject to criminal persecution by law enforcement agencies. So far, only human moderation, as relied upon, e.g., by Twitter, ensures that these types of hate speech are identified. But human

* Correspondence to: C/ Roc Boronat, 138, 08018 Barcelona, Spain.
E-mail addresses: paula.fortuna@upf.edu (P. Fortuna), juan.soler@upf.edu (J. Soler-Company), leo.wanner@upf.edu (L. Wanner).

[1] https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en.

https://doi.org/10.1016/j.ipm.2021.102524
Received 9 August 2020; Received in revised form 11 November 2020; Accepted 20 January 2021
Available online 9 February 2021

moderation is expensive and, in addition, often implies a significant negative psychological impact on the moderators (Roberts, 2019). This calls for automation of the procedure.

Over the last decade, a considerable body of work has been carried out in the area of automatic identification and classification of hate speech and related phenomena in terms of fine-grained categories such as, e.g., 'racism', 'sexism', 'hate speech', etc.; cf., e.g., Davidson, Warmsley, Macy, and Weber (2017), Fortuna and Nunes (2018), Mossie and Wang (2020), Salawu, He, and Lumsden (2020), Schmidt and Wiegand (2017), Swamy, Jamatia, and Gambäck (2019) and Waseem and Hovy (2016). However, as argued, e.g., in Badjatiya, Gupta, Gupta, and Varma (2017), most of the research focused so far on a single dataset, with data collected from a single social media platform. Furthermore, in contrast to other research areas, so far no datasets have been established as standard benchmarks. Researchers often compile datasets of limited size and diversity. Moreover, guidelines for annotation of abusive language content that focus on one specific dataset are prone to shortcuts, dataset-specific rules and over-simplifications (Vidgen & Derczynski, 2021). This is critical since the use of different, sometimes theoretically ambiguous or misleading, terms for equivalent categories impedes the reuse of resources (Kumar, Ojha, Malmasi, & Zampieri, 2018; Vidgen et al., 2019). It also raises the question on how state-of-the-art models trained on one or several commonly used datasets perform on cross-dataset classification. Should they perform well, they can be used, e.g., for assisting manual moderation of the large-scale heterogeneous data in social media.

Recently, several studies tackled the question of cross-dataset classification; cf., e.g., Arango, Pérez, and Poblete (2019, 2020), Chandrasekharan, Samory, Srinivasan, and Gilbert (2017), Karan and Šnajder (2018), Salminen et al. (2020), Swamy et al. (2019) and Waseem, Thorne, and Bingel (2018). However, these studies consider only a limited number of datasets (and thus a limited number of categories), as, e.g., Waseem et al. (2018), or they merge all categories into one positive[2] category (e.g., 'abusive'), which is then contrasted to a negative category (e.g., 'not abusive'), as, e.g., Karan and Šnajder (2018) and Swamy et al. (2019). Both the limitation in scope and the fusion of the original categories of the different datasets into one generic category impede a conclusive answer on the generalization potential of abusive language classification models across datasets. To address this problem, we analyze the cross-dataset performance of two state-of-the-art models, BERT (Devlin, Chang, Lee, & Toutanova, 2019) and ALBERT (Lan et al., 2020), and two baselines, fastText (Bojanowski, Grave, Joulin, & Mikolov, 2017), and SVM (Pedregosa et al., 2020), trained on nine of the most common abusive languagedatasets in English.[3] To be able to compare the performance of such models across datasets, we normalize the original dataset categories using the conversion schema proposed by Fortuna, Soler, and Wanner (2020). To better understand the generalization potential of the classifiers, we assess the performance of 450 BERT, 450 ALBERT, 450 fastText, and 348 SVM binary abusive language classifiers (1698 in total) based on different features, using a Random Forest model.

Overall, our contribution is threefold:

- using a standardized category schema, we are able to identify dataset categories which favor the generalization potential of a model;
- we provide clear evidence that, in contrast to the argumentation in some recent works, e.g., Gröndahl, Pajola, Juuti, Conti, and Asokan (2018), the model itself is equally of high relevance for generalization;
- we identify the dataset features and models that are responsible for a higher degree of generalization.

In the next section, we review the related work on the cross-dataset abusive language model generalization. Section 3 outlines the set up of our experiments. Section 4 introduces the results of the experiments and discusses them and, finally, Section 5 draws some conclusions and outlines some lines of our future work.

2. Related work

In what follows, we present in Section 2.1 the most common datasets collected from a single platform, which we use in our experiments. Since our main goal is to explore the cross-dataset generalization potential of abusive language classification models, we limit our review of the related work in Section 2.2 to the relevant works on cross-dataset classification.[4]

2.1. Main datasets in hate speech research

Table 1 lists the main datasets used in the field and the categories these datasets are annotated with; the dataset identifiers (Ids) in the table are used henceforth to refer to each specific corpus.

[2] With "positive category" we refer to the class that we want to detect, as is traditionally done in machine learning, e.g., Sebastiani (2002).

[3] In the first phase of our experiments, we assessed the use of a variety of other deep learning models. However, based on previous works, cf., e.g., Pedregosa et al. (2020), who report poor cross-dataset performance when using CNN+GRU, LSTM, and ULMFiT for abusive language detection on several hate speech datasets, and on our own experience, we opted to focus on transformer models and on FastText, which shows results that are on par with above mentioned learning models in terms of accuracy and an order of magnitude faster in performance. We further run a number of experiments with SVMs to assess the performance of models that have been very prominent in the field until very recently.

[4] A research area in which recently data from different social media platforms have also been merged in order to identify messages in an out-of-domain context is content moderation (Chandrasekharan et al., 2017). However, the discussion of the works in this area is beyond our scope.

Table 1
Main hate speech datasets ('Id': identifier we use as reference; 'Categories': annotation categories; 'Size': number of instances, 'Source': data source; 'Performance': best score achieved on the dataset and reference (F1 for all datasets, except *Stormfront*; accuracy for *Stormfront*); 'Reference': dataset reference.

Id	Categories	Size	Source	Performance	Reference
W&H	racism, sexism	16,914	Twitter	0.88 (Swamy et al., 2019)	Waseem and Hovy (2016)
Waseem	racism, sexism, both	16,914	Twitter	0.74 (Karan & Šnajder, 2018)	Waseem (2016)
Davidson	hate speech, offensive	24,802	Twitter	0.96 (Swamy et al., 2019)	Davidson et al. (2017)
Ami	misogynous	5000	Twitter	0.66 (Pamungkas & Patti, 2019)	Fersini, Nozza and Rosso (2018)
Stormfront	hate speech	9656	Stormfront	0.73 (de Gibert, Perez, García-Pablos, & Cuadros, 2018)	de Gibert et al. (2018)
TRAC	covert aggression, overt aggression, aggression	15,000	Facebook	0.77 (Swamy et al., 2019)	Kumar et al. (2018)
Hateval	hate speech, aggression	10,000	Twitter	0.65 (Basile et al., 2019)	Basile et al. (2019)
Offenseval	offensive	14,000	Twitter	0.83 (Zampieri et al., 2019)	Zampieri et al. (2019)
Kaggle	threat, identity hate, severe toxic, insult, obscene, toxic	159,571	Wikipedia	0.78 (van Aken, Risch, Krestel, & Löser, 2018)	Jigsaw (2019)
Founta	normal, spam, abusive, hateful	80,000	Twitter	0.93 (Swamy et al., 2019)	Founta et al. (2018)
Kol	toxicity	730	Newspaper	0.82 (Karan & Šnajder, 2018)	Kolhatkar et al. (2020)
Gao	hateful	1069	Fox news	0.49 (Karan & Šnajder, 2018)	Gao and Huang (2017)
Wul1	aggressive	69,526	Wikipedia	0.70 (Karan & Šnajder, 2018)	Wulczyn, Thain, and Dixon (2017)
Wul2	attack	69,526	Wikipedia	0.71 (Karan & Šnajder, 2018)	Wulczyn et al. (2017)
Wul3	toxic	95,692	Wikipedia	0.75 (Karan & Šnajder, 2018)	Wulczyn et al. (2017)
Salminen	hierarchy with 30 classes	137,098	YouTube	0.79 (Salminen et al., 2018)	Salminen et al. (2018)
Almerekhi	toxicity	10,100	Reddit	0.83 (Almerekhi, Kwak, Jansen, & Salminen, 2019)	Almerekhi et al. (2019)
Golbeck	harassment	35,000	Twitter	0.39 (Pamungkas & Patti, 2019)	Golbeck et al. (2017)

2.2. Cross-dataset classification studies

As already mentioned above, a number of studies address the question of the generalization potential of models in "cross-dataset" abusive language classification tasks.[5] Waseem et al. (2018) experiment on three Twitter datasets, *W&H*, *Waseem*, and *Davidson*.[6] *Waseem* is an extension of *W&H*; both are merged and contrasted against the *Davidson* dataset. The authors show that the performance of cross-dataset classification is low; to improve it, training data from the other dataset is needed in that either the different datasets are merged, or the models trained on one dataset are fine-tuned using transfer learning on the data of the other dataset. Gröndahl et al. (2018) also report poor cross-dataset performance, but on more datasets and with different experimental setups. The authors use linear regression, character-based multilayer perceptron, CNN+GRU, LSTM, and ULMFiT for abusive language detection on *W&H*, *Davidson*, *Wul2* and *Zhang* datasets, and show that good performance is achieved only when tested on the same dataset. Karan and Šnajder (2018) use a broader range of nine different datasets: *W&H*, *Waseem*, *TRAC*, *Kol*, *Gao*, *Kaggle*, *Wul1*, *Wul2*, and *Wul3*.[7] Prior to the experiments, the labels of all datasets are binarized into 'positive' (abusive language) and 'negative' (not abusive language). This implies that the distinction between the original categories gets lost, which impedes a detailed analysis of the characteristics of each of them and a fine-grained abusive language classification. Support Vector Machines (SVM) with unigram-count models are first trained on each of the (re-labeled) datasets and tested on the other eight datasets. Transfer learning is then used (as in another previous work Waseem et al., 2018), namely, FEDA ("Frustratingly Easy Domain Adaptation"), to obtain a certain generalization. The authors conclude that for a good performance on the target dataset classification, it is crucial to have as training data at least some data from the target dataset. Note, however, that this conclusion is not consistent with the work of Swamy et al. (2019) and also with our analysis; see Section 4.

All three studies from above also assess the influence of the characteristics of the datasets on cross-dataset classification. Thus, Waseem et al. (2018) state that the *Davidson* dataset is easier to classify than the *W&H* dataset since the vocabulary in the *Davidson* dataset contains a high percentage of African American Vernacular English and is thus more homogeneous. Karan and Šnajder (2018) further hypothesize that differences in cross-dataset performance from dataset to dataset are due to the differences between the categories of the datasets and the dataset sizes. Gröndahl et al. (2018) also argue that the type of data and labeling criteria are of higher relevance than the model. However, Swamy et al. (2019) show that with state-of-the-art models such as BERT it is possible to obtain a language model that achieves some generalization, which highly depends on the training data. As Karan and Šnajder (2018), and Swamy et al. (2019) merge the categories of the considered datasets into two generic categories, 'positive' (abusive) and 'negative' (non-abusive or "benign") – although not all of the used datasets capture the same type of abusive language. BERT models (Devlin et al., 2019) are applied to four Twitter datasets (*W&H*, *Davidson*, *Offenseval*, and *Founta*). The authors state that a model will generalize better if it is used on data that is more similar to the data used for training. Thus, a model trained on the *Founta* dataset performs well when tested on the similar *Offenseval* dataset and vice versa. In a separate experiment, Swamy et al.

[5] In the literature, often the term "out-(of-)domain" is used. We prefer the more precise term "cross-dataset" since in some cases, different datasets from the same source and of the same domain are considered.

[6] Cf. Table 1 for the dataset identifiers.

[7] Note, however, that *Wul1*, *Wul2*, and *Wul3* share data.

(2019) build models with all the categories present in the *Offenseval* dataset and test them also on all the categories of the other three datasets. This facilitates the identification of some overlap between the considered datasets.

Swamy et al. (2019) also observe a performance drop when going from a large training dataset to a small test set and vice versa; this is in line with a related conclusion by Karan and Šnajder (2018) that datasets with a larger percentage of positive samples tend to generalize better than datasets with fewer positive samples, in particular, when tested against dissimilar datasets. For instance, models trained on the *Davidson* dataset, which contains in it majority offensive instances, perform well when tested on the *Founta* dataset, which contains in its majority non-offensive instances.

In another study (Pamungkas & Patti, 2019), the authors confirm that a model trained on datasets with a broader coverage of phenomena is able to also detect other kinds of abusive language than those it has been trained on. The authors use the *W&H*, *Hateval*, *Offenseval* and *Golbeck* datasets, with linear SVM with bag-of-words (BOW) and LSTM as models. However, it should be noted that the generalization quality in this experiment is – with a maximum F1 score of 0.55 – rather moderate.

Arango et al. (2019) bring up an additional characteristic that should be taken into account in the context of the cross-dataset hate speech classification, namely the number of authors of the material captured in a dataset. They show that the generalization potential of the *Waseem* hate speech dataset, whose messages marked as hateful ('sexist' or 'racist') stem from few accounts, increases when the dataset is enriched by hate speech examples from other accounts taken from the *Davidson* dataset (Davidson et al., 2017). But even in this case, the achieved F1 score is merely 0.54 for the hate speech category. That is, more diverse datasets are useful, but they do not solve the problem of poor cross-dataset classification.

Another recent study, which aims to overcome the limited generalization potential of models across domains and thus datasets, was presented by Salminen et al. (2020). The authors argue that models trained on datasets compiled from one online platform can be assumed to be restricted *a priori* to that platform and that a cross-platform dataset ensures more generalizability of a model. They use four datasets collected from YouTube (Salminen et al., 2018), Reddit (Almerekhi et al., 2019), Wikipedia (Jigsaw, 2019), and Twitter (Davidson et al., 2017). Similarly to Karan and Šnajder (2018) and Swamy et al. (2019), they merge the samples into two generic categories, positive ('hateful') and negative ('non-hateful'). Several classification algorithms (Logistic Regression, Naïve Bayes, Support Vector Machines, XGBoost, and Neural Networks) and feature representations (Bag-of-Words, TF–IDF, Word2Vec, BERT, and their combination) are then applied to the generic categories. With XGBoost and all features, the best performance is reported to be 0.92 F-score.

As far as the use of models is concerned, previous studies on model generalization draw upon a range of different supervised classification models. Some use SVMs (e.g., Karan & Šnajder, 2018; Pamungkas & Patti, 2019), mostly as baseline; others use deep learning (e.g., Gröndahl et al., 2018). More recently, authors have been using transformer-based models such as BERT, which render better performance; cf., e.g., Salminen et al. (2020) and Swamy et al. (2019).

Transfer learning is also being considered; see, e.g., Karan and Šnajder (2018) and Waseem et al. (2018). For instance, in one recent study (Mozafari, Farahbakhsh, & Crespi, 2019), the authors introduce a novel transfer learning approach based on BERT. More specifically, they investigate the ability of BERT for capturing hateful context within social media content by using new fine-tuning methods based on transfer learning. $\text{BERT}_{\text{BASE}}$ with an Inserted CNN layer proved to be the best model, leading to F1 of 0.88 on the *W&H* and of 0.92 on *Davidson* datasets.

Even if multilingual hate speech classification is not in the focus of our presented work, it is worth mentioning that recently considerable advances have been made in this area; cf., e.g., Ousidhoum, Lin, Zhang, Song, and Yeung (2019) for a multilingual and multi-aspect approach. In Pamungkas and Patti (2019), a promising cross-lingual classification approach to both in-domain (with F1 of 0.69) and out-domain (with F1 of 0.59) abusive language is presented.

2.3. Open questions

Recent studies have shown that model generalization in the context of hate speech and abusive language classification can be achieved by merging different datasets (Salminen et al., 2020). However, we are interested in exploring to what extent a model can generalize across datasets when trained on the category or categories of one given dataset. In this context, at least two important questions are still open: **1. Are the models or the datasets decisive for cross-dataset generalization?** Gröndahl et al. (2018) defend that the nature and composition of the datasets are more important than the models. However, Swamy et al. (2019) provide evidence that by using state-of-the-art models such as BERT at least a certain generalization can be achieved. **2. Which model and dataset characteristics are important for generalization, after all?** The authors refer, e.g., to dataset size, similarity between categories (Karan & Šnajder, 2018; Swamy et al., 2019), advantage of classes with broader coverage of phenomena (Pamungkas & Patti, 2019), and author diversity (Arango et al., 2019). However, none of these studies carries out an exhaustive analysis that would assess the effect of dataset characteristics and their correlation. In our experiments, we address this issue. As Salminen et al. (2020) and Swamy et al. (2019), we use BERT as one of the state-of-the-art models. In order not to rely on one model only, we also use ALBERT, fastText, and SVM. In contrast to the works of Pamungkas and Patti (2019), Salminen et al. (2020), and Swamy et al. (2019), we do not merge all categories into two generic categories, but, rather, standardize the category labels across datasets, which ensures that the original differentiation remains and, at the same time, enables the comparison of classifiers trained on different datasets. We assess each considered category with respect to its generalization capacity across datasets via binary classification for all original and standardized labels (<label> vs. NIL). To address the question of the influence of model and dataset characteristics on their generalization potential, we apply a Random Forest classifier.

Table 2
Used dataset properties and summarized origin of the data.

Dataset name	Mutually exclusive classes	Original classes	Origin of the data
W&H	No	racism, sexism	Search Twitter for common slurs and terms used to refer to religious, sexual, gender, and ethnic minorities.
Davidson	Yes	hate speech, offensive	Search Twitter with the Hatebase lexicon.
Ami	Yes	misogynous	Search Twitter for representative slurs, monitoring of potential victims' and perpetrators accounts.
Stormfront	Yes	hate speech	Subset of 22 Stormfront sub-forums covering diverse topics and nationalities was random-sampled.
TRAC	Yes	covert aggression, overt aggression	Search Twitter for keywords and constructions that are often included in offensive messages, such as 'she is', 'antifa', 'conservatives'.
HatEval	No	hate speech, aggression	Collection of tweets directed against immigrants and women.
Offenseval	Yes	offensive	Search Twitter for keywords and constructions that are often included in offensive messages, such as 'she is' or 'to:BreitBartNews' in the Twitter API
Kaggle	No	threat, identity hate, severe toxic, insult, obscene, toxic	Not provided.
Founta	Yes	normal, spam, abusive, hateful	Search Twitter for a mixture of a random sample with tweets that have strong negative polarity and contain at least one offensive word.

3. Experimental setup

3.1. Datasets

We use nine publicly available datasets from the list in Table 1 that cover different offensive, abusive language related categories: *W&H*, *Davidson*, *Ami*, *Stormfront*, *TRAC*, *Kaggle*, *Hateval*, *Offenseval*, and *Founta*.

From most of the datasets, we consider only the training sets of the datasets since their test sets were not always available and in cases they were, the splits between the training and test sets varied. We split these training sets randomly into 70% for training and 30% for testing our models. The exceptions are *Hateval* and *Offenseval*, of which we use both the training and the test sets in their original 70%–30% split. This is because we had them at our disposal from our previous work (Fortuna, Soler-Company and Nunes, 2019).

As already mentioned above, to obtain an objective picture of the generalization potential of the models across different datasets, we use a procedure for category standardization presented in our previous work (Fortuna et al., 2020), extending it to the three additional datasets that are considered in the present study (*Stormfront*, *Offenseval* and *Founta*). This procedure relies on analyzing dataset categories, properties, definitions, and data collections as summarized in Table 2.

Let us illustrate, in what follows, the essence of this procedure. We keep the original labels when possible. For instance, in the case of the *W&H* dataset, the 'sexism' and 'racism' categories are considered both as separate categories, but also as subcategories of 'hate speech'. Hence, a new category ('hate speech') is added to this dataset. Furthermore, the 'sexism' category in this dataset is assumed to be equivalent to the 'misogynous' category of the *Ami* dataset since in the literature no clear distinction between these two is provided. The resulting standardized cross-dataset label is called 'misogyny-sexism'. For the *Davidson* dataset, we use a new category 'toxicity' that subsumes the union of its 'hate speech' and 'offensive' categories. 'Toxicity' is an umbrella term that aims to capture general offense and different types of 'aggression' (Kolhatkar et al., 2020). *Ami's* 'misogynous' category is assumed to be equivalent to the 'sexism' category in *W&H*'s dataset, as already mentioned. The *TRAC* dataset contains the categories 'overt aggression' and 'covert aggression', which we merge into a new category 'aggression'. We could also convert it to a general category such as 'toxicity', however, we opt not to do it as *TRAC* aims to identify subtler aggression, which is a dimension not mentioned in the *Kaggle* dataset. Regarding the *Hateval* dataset, its aggression category covers a specific type of aggression as it is a subset of 'hate speech'. In this case, we do not merge the two categories into 'toxicity', as this dataset aims to classify only 'hate speech', and considers 'aggression' only when it happens in the context of 'hate speech'. Moreover, we consider it 'aggressive hate speech' and also not equivalent to the 'aggression' category in *TRAC* dataset. For the *Kaggle*'s dataset, we kept the original labels, as the dataset was already annotated in a multiclass manner. For *Stormfront*, we kept the original category since the authors use a 'hate speech' definition that focuses on the target characteristics of this type of communication (e.g., gender and age). This is similar to previous definitions found in the literature, and we want to test if the different 'hate speech' annotated datasets generalize within themselves. In the *Offenseval* dataset, only 'offense' is annotated as a general category that is meant to cover all types of 'offensive' speech. Therefore, we converted it to 'toxicity', such that it becomes comparable with the equally general terms found in *Davidson*'s and *Kaggle*'s datasets (again, using the criteria defined in Fortuna et al., 2020). Finally, the *Founta* dataset is annotated with 'hateful', 'abusive', 'spam', and 'normal' labels. We considered 'spam' to fall into the category 'normal', as we are interested in abusive speech only. In our standardized category scheme, we mark both as 'none'. In contrast, we keep the original 'abusive' labels. Although the authors mention that 'abusive' significantly correlates with 'aggression' and 'offensive' categories, we left it as it is since 'abusive' is the most popular among the three, the most central in *Founta*, and it is the label that the authors preferred for their dataset. Furthermore, it seems that this category is not equivalent to 'aggression' from *TRAC*, as it does not include the covert and overt dimensions. In both cases, we discarded their conversion to the 'offensive' category of *Davidson*, since we believe that with the conversion we would lose information. Finally, the category 'hateful' is converted to 'hate speech' to be in line with the definition in the literature. The resulting conversion is presented on Table 3.

Table 3
Standardized categories used in our study (for convenience, we refer in the text to 'misogyny-sexism' as 'sexism').

Dataset	Original category	Standardized category
W&H	sexism	misogyny-sexism
	racism	racism
	sexism or racism	hate speech
Davidson	hate speech	hate speech
	offensive	offensive
	hate speech or offensive	toxicity
Ami	misogynous	misogyny-sexism
TRAC	covert aggression	covert aggression
	overt aggression	overt aggression
	overt or covert aggression	aggression
Hateval	hate speech	hate speech
	aggression	aggressive hate speech
Kaggle	threat	threat
	identity hate	hate speech
	severe toxic	severe toxic
	insult	insult
	obscene	obscene
	toxic	toxicity
Stormfront	hate speech	hate speech
Offenseval	offensive	toxicity
Founta	hateful	hate speech
	abusive	abusive
	spam	none
	hateful or abusive	toxicity

3.2. Experiments

3.2.1. Intra-dataset model evaluation

Our first experiment is similar to the one from Swamy et al. (2019), but with more datasets. We first create binary intra-dataset classification models for BERT, ALBERT, fastText and SVM and our nine datasets and respective standardized categories. BERT is the model with state-of-art performance in abusive language related tasks. We also experiment with ALBERT because it has been reported to outperform BERT (Lan et al., 2020) in some cases, and with fastText and SVM as a baselines. The macro averaged F1 score is used for evaluation of all the models.

For BERT, ALBERT and FastText, we use off-the-shelf models.[8] As already mentioned above, we split the training sets of all datasets, except *Offenseval* and *Hateval*, for which we kept the original training and test sets, randomly into 70% for training and 30% for testing. This training-test set division per dataset (cf. Table 4) is kept all over the experiments, hence also for all the standardized categories of a given dataset. For the SVM experiments, we use 10-fold cross-validation instead of a 70%/30% split.[9]

The training of both BERT and ALBERT is carried out on a TPU in COLAB with Tensorflow 1.15.[10] In the case of BERT and ALBERT, we refer to the number of layers or Transformer blocks as L, to the hidden size as H, and to the number of self-attention heads as A.

BERT. We use $\text{BERT}_{\text{LARGE}}$ ($L = 24$, $H = 1024$, $A = 16$) with 340M parameters in total, which outperformed $\text{BERT}_{\text{BASE}}$ across all tasks, especially those with very little training data (Devlin et al., 2019), as is the case with some of our datasets. We use a batch size of 32 and fine-tune for 3 epochs over the data of all datasets. The dropout probability is set to 0.1 for all layers; the Adam optimizer is used with a learning rate of 2e−5.

ALBERT. We use $\text{ALBERT}_{\text{XXLARGE}}$ ($L = 12$, $H = 4096$, $A = 64$) model with about 70% of $\text{BERT}_{\text{LARGE}}$'s parameters for an available trained ALBERT model (Lan et al., 2020), which we fine-tune to all nine datasets as described in Lan et al. (2020). We use a batch size of 32, the dropout probability is set to 0.1 for all layers and the Adam optimizer is used with a learning rate of 1e−5.

Fasttext. FastText is similar to the CBOW model of Mikolov, Chen, Corrado, and Dean (2013). We ran the model in its version 0.9.2 with 300 dimension-FastText pretrained vectors (Mikolov, Grave, Bojanowski, Puhrsch, & Joulin, 2018), Skipgram Hierarchical softmax loss function, learning rate of 1.0, considering 1 as minimal number of word occurrences, bi-grams, and 25 epochs.

[8] https://fasttext.cc/docs/en/english-vectors.html https://github.com/google-research/bert https://github.com/google-research/albert.
[9] For this purpose, we merge the training and test sets of *Offenseval* and *Hateval*.
[10] https://github.com/paulafortuna/IP-M_abusive_models_generalize.

Table 4
Dataset and respective standardized category (st. category), total number of instances for training (train total N), total number of instances for test (test total N), total number of positive instances for training (train total pos), and percentage of positive instances in the training set (train perc positive).

Dataset	st. category	Train total N	Test total N	Train total pos	Train perc positive
W&H	sexism	11 835	5073	2407	0.20
W&H	racism	11 835	5073	1377	0.12
W&H	hate speech	11 835	5073	3784	0.32
Davidson	hate speech	17 348	7435	975	0.06
Davidson	offense	17 348	7435	13 517	0.78
Davidson	toxicity	17 348	7435	14 492	0.84
Ami	sexism	2800	1200	1249	0.45
TRAC	covert aggression	12 000	3000	4240	0.35
TRAC	overt aggression	12 000	3000	2708	0.23
TRAC	ov cov aggression	12 000	3000	6948	0.58
Hateval	hate speech	9000	1000	3783	0.42
Hateval	aggressive hs	9000	1000	1559	0.17
Kaggle	toxicity	111 699	47 872	10 856	0.10
Kaggle	hate speech	111 699	47 872	977	0.01
Kaggle	severe toxicity	111 699	47 872	1107	0.01
Kaggle	insult	111 699	47 872	5593	0.05
Kaggle	obscene	111 699	47 872	6008	0.05
Kaggle	threat	111 699	47 872	336	0.00
Stormfront	hate speech	3501	1500	705	0.20
Offenseval	toxicity	13 240	319	4400	0.33
Founta	hate speech	64 366	27 587	2885	0.05
Founta	abusive	64 366	27 587	14 463	0.23
Founta	toxicity	64 366	27 587	17 348	0.27

BOW + SVM. In the BOW+SVM experiments, we use the Scikit Learn models (Pedregosa et al., 2011).[11] For the Bag-Of-Words (BOW) extraction, we remove stopwords and consider only words with a frequency ≥ 1%. For SVM classification, we use most of its default parameters, except for the kernel, which was set to the linear kernel. Due to the time complexity of the parameter extraction and training procedures, we use SVM with bagging. The time complexity, paired with the size of the dataset, forces us to exclude the *Kaggle* dataset from this classification task.

Figs. 1 and 2 display the results of BERT/ALBERT/fastText/SVM for intra-dataset classification in terms of the macro averaged F1 score, grouped by standardized categories (Fig. 1) and datasets (Fig. 2).

3.2.2. Inter-dataset model evaluation

The obtained intra-dataset models are tested in a second experiment in a cross-dataset scenario. More precisely, each model, trained on a certain dataset, is tested against all the remaining datasets and corresponding standardized categories. We use the same training and test set divisions for intra-dataset and cross-dataset experiments. As already before, the macro averaged F1 measure is used for evaluation. The results of the experiment on inter-(or cross-)dataset classification are displayed in Table 5.[12]

Due to space constraints, we show only the results with F1 score ≥ 0.60.[13] We assume that there is a better cross-dataset generalization if at least one of the four algorithms (BERT, ALBERT, fastText, or SVM) achieves with its model an F1 score of ≥ 0.70.

3.2.3. Model performance classification

To systematically study which model and dataset features lead to a better generalization in abusive language-related models, we run an experiment on the relation between the performance figures obtained when applying BERT, ALBERT, fastText, and SVM and 16 prominent features of the models and datasets considered in the literature as good generalization predictors; cf. Table 6. For this purpose, we group the 1698 binary BERT/ALBERT/fastText/SVM models (450 of each for BERT/ALBERT/fastText and 348 for SVM) into models that generalize better (those with an F1 score ≥ 0.70; 136 in total) and models that generalize worse (those with an F1 score < 0.70; 1562 in total). The goal is to train a classifier on the above 16 features to predict whether a model belongs to the better generalizing models or worse generalizing models. As classifier, we use a Random Forest with 50 estimators (Pedregosa et al., 2011) with 5 Fold cross-validation. We have chosen Random Forest since it is a general-purpose classifier with weak statistical assumptions. To rank the different features used for classification, we use the permutation feature importance algorithm,[14] which directly measures feature importance by observing how random re-shuffling of each predictor influences model performance, without changing the distribution of the variable. Before using this model, we normalize data with the Z-score method (Kreyszig, 1960).

[11] For BOW, we use the *CountVectorizer* class, for SVM the SVC class and for bagging the *BaggingClassifier* class.

[12] The shades in the table cells reflect the F1 score: from white (F1 ≤ 0.69) to green (F1 = 1.0).

[13] Note that in what follows, we use dataset name abbreviations as introduced in Table 5. The complete results will be provided in the project repository upon acceptance.

[14] https://explained.ai/rf-importance/index.html.

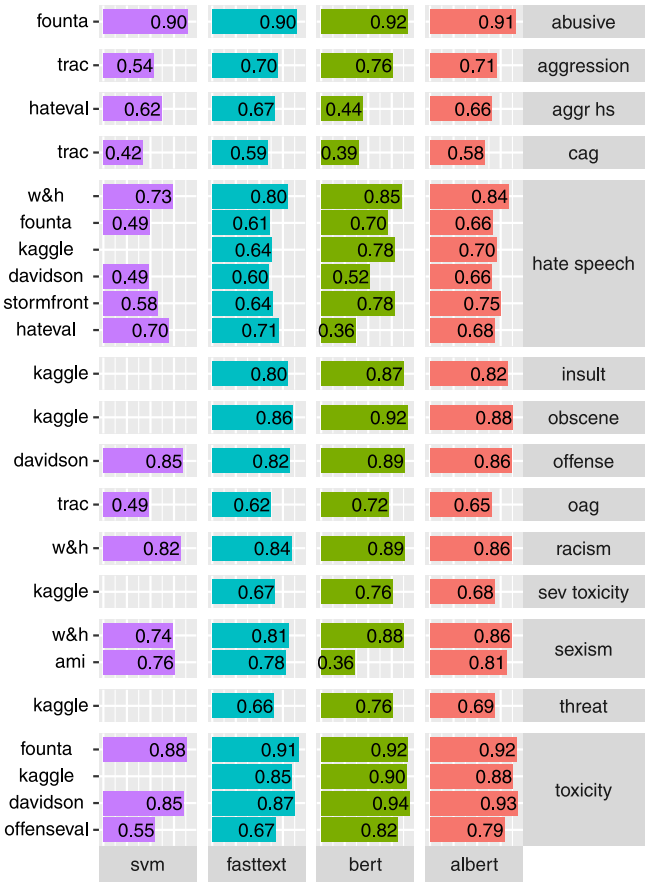

Fig. 1. Macro F1 scores, by standardized categories ('cag': covert aggression; 'oag': overt aggression, 'sev toxicity': severe toxicity, 'aggr hs': aggressive hate speech).

4. Discussion

In this section, we discuss in detail the outcome of the experiments.

4.1. Outcome of intra-dataset classification

As can be observed in Figs. 1 and 2, in general (and as expected from previous works), SVM was the model that performed worst, with some exceptions. Thus, in general, it performed poorer than fastText, except for the 'abusive' category of the *Founta* dataset, where both scored equally and the 'offense' category of the *Davidson* dataset, where SVM was slightly better. It also performed worse than ALBERT for all cases, except for 'hate speech' in the *Hateval* dataset, where it was slightly better.

FastText performed generally worse than BERT and ALBERT, which is in line with d'Sa, Illina, and Fohr (2020) and Uglow, Zlocha, and Zmyslony (2019), who also report a poorer performance of fastText compared to BERT.

Even though BERT and ALBERT achieved an overall better performance than the baseline models, BERT's performance was not good (lower than 0.52) in some categories: 'hate speech' in *Davidson*, 'sexism' in *Ami*, both categories in *Hateval*, 'covert aggression' in *TRAC* and 'hate speech' in *Stormfront*. In these cases, both SVM and fastText, or at least one of them, obtained a better performance than BERT. This may be explained by the fact that BERT is unstable on smaller datasets (Devlin et al., 2019). BERT is more unstable than ALBERT and fastText, both in terms of the same category (cf., e.g., 'hate speech', 'sexism') Fig. 1, and same dataset (*TRAC*, *Hateval*, *Davidson*) Fig. 2. However, BERT also achieves the best performance on the largest number of categories.

As illustrated in Fig. 1, from the categories present in more than one dataset, 'toxicity' proved to be the easiest category to classify, followed by 'sexism' and then 'hate speech', the latter one with more unstable results. From the dataset-specific categories, 'abusive' was the easiest to classify, followed by 'obscenity', 'offensive', 'racism', and 'aggression'. The remaining categories showed worse performance.

According to Fig. 2, among the datasets with more than one category, *W&H* shows good and stable results for all categories, while *TRAC* and *Hateval* show worse performances. For the other datasets, the performance varies from category to category. *Davidson*

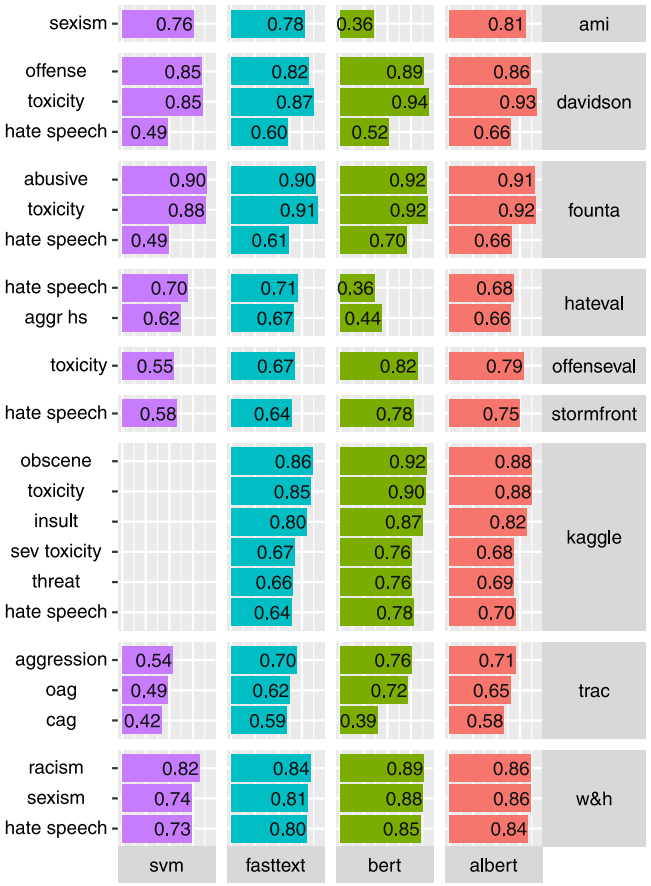

Fig. 2. Macro F1 scores, by datasets.

and *Founta* both show good performances, except for 'hate speech'. *Kaggle* shows good performance for 'obscene', 'toxicity' and 'insult', but worsens for 'hate speech', 'threat' and 'severe toxicity'. Among the datasets with only one category, *Ami*, *Offenseval*, and *Stormfront* had the best scores.

When we compare our results in Figs. 1 and 2 with the figures reported in the literature (cf. Table 4), we see that the ALBERT and BERT models achieve similar performance as reported in other transformer-based works (such as Swamy et al., 2019) for the classification of, e.g., 'hate speech' in *W&H*, 'toxicity' in *Davidson*, 'aggression' in *TRAC* and 'toxicity' in *Founta*. These models outperform works that do not use transformers; cf., e.g., Pamungkas and Patti (2019) for classification of 'sexism' in *Ami*; de Gibert et al. (2018) for classification of 'hate speech' in *Stormfront*; Basile et al. (2019) for classification of 'hate speech' in *Hateval*, and van Aken et al. (2018) for classification of 'toxicity' in *Kaggle*.

4.2. Outcome of inter-dataset classification

The results of the inter-dataset classification experiment provide some interesting insights with respect to both models and datasets. Both are discussed in the following subsections.

4.2.1. Discussion of the models

Table 5 shows that BERT and ALBERT models generalize better more often than fastText and SVM. Thus, the results for fastText are worse than for BERT and/or ALBERT, except for the generalization of *fount.tox* to *kaggl.obsc* (F1 of 0.82) and *kaggl.insu* (0.78), and *kaggl.insu* to *fount.abus* (0.70). For this last case, it was the only model capable to generalize. SVM performed generally worse than fastText, BERT and ALBERT, being better only when trained on *offen.tox* and tested on *kaggl.obsc* or *kaggl.insu* (0.81 and 0.74 respectively), or trained on *ami.sex* and tested on *david.offe* (0.74). In contrast to fastText, SVM does not show a distinct generalization potential, i.e., it is capable to generalize only if BERT, ALBERT or fastText also are.

BERT generalizes best for almost all the models; ALBERT is close to BERT many times and is even better for the models trained with *hateval.hs* and *ami.sex*.

Table 5
Model generalization evaluation of BERT (be), ALBERT (al), fastText (ft) and SVM (svm) in terms of macro F1 score. The second row from the top in bold indicates the dataset and standardized category used for training. The remaining rows (second column) of the test data (*Offenseval* ('offen'), *Davidson* ('david'), *Founta* ('fount'), *Kaggle* ('kaggl'), *Hateval* ('hatev') and categories toxicity ('tox'), obscene ('obsc'), insult ('insu'), aggression ('aggr'), offensive ('offe'), overt aggression ('oag'), covert aggression ('cag'), abusive ('abus'), hate speech ('hs'), sexism ('sex'), racism ('race'), severe toxicity ('stox'), aggressive hate speech ('aghs')).

		1					2					3				
		train ->	fount abus				train ->	david offe				train ->	david tox			
		model ->	svm	ft	be	al	model ->	svm	ft	be	al	model ->	svm	ft	be	al
Testing	1	kaggl obsc	**0.83**	**0.82**	**0.87**	**0.71**	fount tox	**0.84**	0.64	**0.89**	**0.84**	fount tox	**0.83**	0.62	**0.91**	**0.86**
	2	kaggl tox	**0.74**	**0.73**	**0.85**	**0.72**	fount abus	**0.86**	0.62	**0.89**	**0.84**	fount abus	**0.82**	0.59	**0.89**	**0.85**
	3	offen tox	**0.75**	**0.70**	**0.83**	**0.77**	kaggl obsc	**0.79**	0.43	**0.82**	**0.74**	kaggl tox	0.67	0.45	**0.85**	**0.77**
	4	kaggl insu	**0.77**	**0.77**	**0.81**	0.67	kaggl tox	**0.73**	0.47	**0.79**	**0.74**	kaggl obsc	0.66	0.39	**0.81**	**0.73**
	5	david offe	0.60	0.58	**0.71**	0.69	offen tox	**0.76**	0.58	**0.79**	**0.79**	offer tox	**0.71**	0.53	**0.81**	**0.80**
	6	david tox	0.56	0.54	0.69	0.67	kaggl insu	**0.73**	0.42	**0.75**	0.69	kaggl insu	0.62	0.39	**0.77**	**0.70**
	7	ami sex	0.66	0.67	0.67	0.65	ami sex	0.68	0.50	0.67	0.65	storm hs	0.54	0.38	0.66	0.62
	8	kaggl stox	0.64	0.67	0.61	0.54	kaggl stox	0.61	0.37	0.58	0.55	ami sex	0.66	0.44	0.64	0.66
	9	hatev aghs	0.62	0.59	0.59	0.58	hatev aghs	0.60	0.46	0.58	0.56	hatev hs	0.59	0.53	0.42	0.61
	10						w&h sex	0.51	0.52	0.55	0.60					
		train ->	fount tox				train ->	offen tox				train ->	trac oag			
		model ->	svm	ft	be	al	model ->	svm	ft	be	al	model ->	svm	ft	be	al
Testing	11	kaggl tox	**0.74**	**0.77**	**0.87**	**0.74**	fount tox	**0.81**	**0.82**	**0.91**	**0.85**	fount tox	0.42	0.62	**0.82**	**0.75**
	12	offen tox	**0.74**	**0.75**	**0.84**	**0.76**	fount abus	**0.84**	**0.81**	**0.88**	**0.83**	fount abus	0.44	0.61	**0.80**	**0.75**
	13	kaggl obsc	**0.81**	**0.82**	**0.79**	0.69	kaggl tox	**0.71**	0.69	**0.83**	**0.75**	kaggl tox	0.48	0.62	**0.79**	**0.72**
	14	kaggl insu	**0.75**	**0.78**	**0.77**	0.67	david tox	0.34	0.63	**0.76**	**0.73**	offer tox	0.43	0.53	**0.72**	**0.71**
	15	david tox	0.56	0.64	**0.77**	**0.74**	kaggl obsc	**0.81**	0.65	**0.73**	0.67	kaggl insu	0.49	0.63	**0.70**	0.69
	16	david offe	0.60	0.66	**0.73**	**0.71**	david offe	0.37	0.64	**0.72**	**0.72**	kaggl obsc	0.49	0.62	**0.70**	0.68
	17	storm hs	0.47	0.56	**0.71**	0.34	kaggl insu	**0.74**	0.64	**0.72**	0.66	ami sex	0.36	0.50	0.69	0.67
	18	w&h hs	0.47	0.51	0.67	0.54	trac oag	0.44	0.59	**0.70**	0.65	storm hs	0.44	0.51	0.68	0.59
	19	trac oag	0.46	0.52	0.67	0.44	storm hs	0.49	0.58	0.69	0.65	davic tox	0.15	0.26	0.67	0.63
	20	ami sex	0.66	0.68	0.66	0.44	ami sex	0.54	0.65	0.64	0.65	davic offe	0.19	0.29	0.66	0.64
	21	w&h race	0.50	0.53	0.63	0.28	hatev hs	0.50	0.58	0.60	0.58	hatev hs	0.37	0.49	0.63	0.58
	22	kaggl stox	0.62	0.63	0.56	0.52	kaggl stox	0.64	0.52	0.53	0.51	w&h race	0.47	0.65	0.62	0.54
	23	hatev aghs	0.61	0.60	0.56	0.56						w&h hs	0.41	0.54	0.61	0.55
		train ->	kaggl tox				train ->	trac aggr				train ->	kaggl insu			
		model ->	svm	ft	be	al	model ->	svm	ft	be	al	model ->	svm	ft	be	al
Testing	24	fount tox	-	**0.88**	**0.92**	**0.88**	fount tox	0.32	0.58	**0.84**	**0.70**	ami sex	-	0.66	**0.74**	0.69
	25	fount abus	-	**0.88**	**0.90**	**0.87**	fount abus	0.29	0.54	**0.80**	0.67	fount tox	-	0.69	**0.70**	0.64
	26	offen tox	-	**0.73**	**0.85**	**0.84**	david tox	0.50	0.52	**0.74**	0.61	fount abus	-	0.70	0.69	0.64
	27	david tox	-	0.59	**0.78**	**0.75**	offen tox	0.30	0.46	**0.74**	0.57	davic offe	-	0.49	0.64	0.62
	28	david offe	-	0.59	**0.73**	**0.71**	david offe	0.48	0.52	0.69	0.60	davic tox	-	0.47	0.63	0.62
	29	ami sex	-	0.67	0.66	0.66	kaggl tox	0.24	0.39	0.66	0.49	hatev aghs	-	0.60	0.62	0.59
	30	trac oag	-	0.56	0.65	0.61	w&h hs	0.30	0.53	0.61	0.55	offer tox	-	0.60	0.61	0.60
	31	storm hs	-	0.53	0.63	0.58	hatev hs	0.35	0.52	0.60	0.56					
	32						storm hs	0.30	0.50	0.60	0.55					
		1					2					3				
		train ->	kaggl obsc				train ->	w&h sex				train ->	storm hs			
		model ->	svm	ft	be	al	model ->	svm	ft	be	al	model ->	svm	ft	be	al
Testing	1	fount abus	-	**0.88**	**0.90**	**0.87**	ami sex	0.35	0.66	**0.73**	0.68	w&h race	0.54	0.54	0.69	0.61
	2	fount tox	-	**0.85**	**0.89**	**0.86**	david offe	0.19	0.53	0.69	0.61	kaggl insu	0.50	0.51	0.65	0.60
	3	offen tox	-	**0.71**	**0.77**	**0.82**	david tox	0.15	0.48	0.65	0.57	kaggl hs	0.51	0.53	0.64	0.57
	4	david tox	-	0.55	**0.73**	**0.75**	kaggl obsc	0.49	0.56	0.64	0.56	kaggl obsc	0.50	0.51	0.63	0.59
	5	david offe	-	0.58	**0.73**	**0.74**	kaggl insu	0.49	0.57	0.63	0.56	kaggl tox	0.49	0.50	0.61	0.58
	6	ami sex	-	0.67	**0.71**	0.69	kaggl stox	0.50	0.59	0.60	0.53					
		train ->	ami sex				train ->	w&h hs				train ->	hatev hs			
		model ->	svm	ft	be	al	model ->	svm	ft	be	al	model ->	svm	ft	be	al
Testing	7	kaggl insu	0.62	0.65	0.49	0.66	ami sex	0.35	0.66	**0.70**	0.61	kaggl insu	0.57	0.59	0.49	0.64
	8	kaggl obsc	0.63	0.65	0.49	0.67	kaggl insu	0.49	0.56	0.66	0.57	kaggl obsc	0.57	0.60	0.49	0.65
	9	kaggl tox	0.59	0.63	0.48	0.65	kaggl obsc	0.49	0.56	0.66	0.57	kaggl tox	0.54	0.61	0.48	0.66
	10	w&h sex	0.56	0.56	0.44	0.62	david offe	0.19	0.56	0.63	0.54	storm hs	0.50	0.55	0.44	0.64
	11	hatev aghs	0.61	0.59	0.44	0.57	david tox	0.15	0.51	0.63	0.51	fount abus	0.52	0.56	0.44	0.64
	12	fount abus	0.54	0.60	0.44	0.62	kaggl tox	0.48	0.55	0.62	0.55	fount tox	0.51	0.57	0.42	0.64
	13	offen tox	0.47	0.62	0.43	0.58	kaggl hs	0.51	0.52	0.62	0.54	ami sex	**0.73**	**0.99**	0.36	**0.84**
	14	fount tox	0.52	0.59	0.42	0.61	kaggl stox	0.51	0.53	0.61	0.54	davic offe	0.68	0.62	0.19	**0.70**
	15	david offe	**0.74**	0.65	0.19	**0.71**	storm hs	0.47	0.57	0.61	0.57	davic tox	0.63	0.58	0.15	0.67
	16	david tox	0.69	0.61	0.15	0.66	hatev hs	0.38	0.52	0.60	0.59					
		train ->	fount hs				train ->	kaggl hs				train ->	david hs			
		model ->	svm	ft	be	al	model ->	svm	ft	be	al	model ->	svm	ft	be	al
Testing	17	w&h race	0.47	0.54	**0.77**	0.51	fount hs	-	0.51	0.64	0.60	w&h race	0.47	0.48	0.47	0.62
	18	storm hs	0.44	0.47	**0.71**	0.48										
	19	w&h hs	0.40	0.44	0.63	0.53										
	20	trac oag	0.43	0.45	0.62	0.47										
	21	kaggl hs	0.50	0.55	0.60	0.58										
		train ->	hatev aghs				train ->	w&h race								
		model ->	svm	ft	be	al	model ->	svm	ft	be	al					
⊢	22	kaggl stox	0.53	0.61	0.50	0.60	storm hs	0.47	0.46	0.66	0.55					

With respect to the first open question on the relevance of models formulated in Section 2.3, we can thus state that models are indeed important for generalization and not all models are equally good for all datasets/categories. BERT achieves better generalization in more cases, but ALBERT and fastText generalize in some cases where BERT does not. On the other hand, an SVM model trained on BOW features generalizes worse than the remaining models. In general, we can observe (and as has already been shown by Swamy et al., 2019 using BERT) that transformer-based models are able generalize better, while other models are less suitable for this task; cf., e.g., Gröndahl et al. (2018), Karan and Šnajder (2018) and Waseem et al. (2018). Our experiments

Table 6
Dataset and model features used for predicting model performance.

Feature	Description
Training dataset	One of *W&H*, *Davidson*, *Ami*, *TRAC*, *Hateval*, *Kaggle*, *Stormfront*, *Offenseval*, or *Founta*.
Training dataset size	Number of instances used for training.
Training dataset percentage of positive instances	Number of instances in the positive class divided by the number of instances used for training.
Original F1	Performance of the model when trained and tested with data from the same class, as provided in Figs. 1 and 2.
Training category	All the distinct standardized categories in Table 3.
Training social network	Social network of the data: Twitter, Facebook, Wikipedia or Stormfront.
Test dataset	One of the same dataset list as in the training dataset.
Test dataset size	Number of instances in the test dataset.
Test category	One of the same category list as in the training category.
Test social network	One as in the same social network list as in the training dataset.
BERT, ALBERT, fastText or SVM	Model used to classify.
Is same category	Boolean indicating whether training and test sets belong to the same category.
Is same social network	Boolean indicating whether training and test set belong to the same social network.
Train and test set proportion	Training dataset size divided by the test dataset size.
Vocabulary-train	After removing stop words (by using NLTK) and keeping only distinct words, we compute the percentage of words that is present in the positive class of the test set.
Vocabulary-test	With the same procedure as for 'Vocabulary test present training', we compute the percentage of distinct words from the positive class that are present in the positive class of the training set.

also show that the generalization capability of a model equally depends on the chosen dataset, and, even more importantly, on the targeted categories; cf. below.

4.2.2. Discussion of the datasets

BERT models that are trained on the category 'toxicity' of a dataset (*Offenseval*, *Davidson*, *Founta*, and *Kaggle*) generalize well over the same category of the other test sets; cf., when trained on *offen.tox*: (0.91;0.83;0.76)[15]; on *david.tox*: (0.91;0.81;0.85), on *fount.tox*: (0.84;0.87;0.77); and on *kaggl.tox*: (0.92;0.85;0.78). This shows that 'toxicity' is homogeneous across different datasets.

'Offensive' and 'abusive' are also consistently predicted well and when a model is trained on one of them, it predicts well 'toxicity'. E.g., when trained on *david.offen* BERT predicts well *fount.tox* (0.89) and *kaggl.tox* (0.79) and ALBERT predicts well *offen.tox* (0.79). BERT also predicts well *david.offen* when trained on these datasets (0.73;0.72;0.73); and when trained on *fount.abus*, it predicts well *offen.tox* (0.83), *kaggl.tox* (0.85), and *david.tox* (with a borderline result of 0.69). *fount.abus* is also predicted well when BERT is trained on these datasets (0.88;0.90;0.89).

Categories such as 'abusive', 'offensive', 'aggression' and 'toxicity', whose definitions tend to overlap, generalize well between each other, which indicates that these labels are conceptually similar or that they represent the same phenomenon. Thus, when trained on *david.offen*, BERT predicts well *fount.abus* (0.89), *kaggl.tox* (0.79), and *offen.tox* (0.79). The prediction of *david.offen* is also accurate when BERT is trained on each of these datasets (0.71;0.73;0.72). The same holds for training on *fount.abus* and predicting *david.offen* (0.71) and the reverse (0.89), for training on *fount.abus* and testing on *kaggl.tox* (0.85) and the reverse (0.90), and for training with *fount.abus* and testing on *offen.tox* (0.83) and the reverse (0.88).

Datasets that include the categories 'abusive', 'offensive', 'aggression' or 'toxicity' also include 'obscene'. 'Obscene' from *Kaggle* (*kaggl.obsc*) as training set obtains good results in different cases: 'toxicity' by BERT (cf. fount.tox: 0.89) and ALBERT (cf. *offen.tox*: 0.82, and *david.tox*: 0.75); 'offensive' by ALBERT (*david.offen*: 0.74); and 'abusive' by BERT (*fount.abus*: 0.90). Another category that is related to 'abusive', 'offensive', 'aggression' and 'toxicity' is 'insult'. When using any of the former for training, models generalize reasonably well over *kaggl.insu* (cf., for BERT *fount.tox*: 0.77, *david.offen*: 0.75, and *fount.abus*: 0.81, *david.tox*: 0.77, for SVM *offen.tox*: 0.74), proving that insults are also commonly subsumed by these categories. In view of the co-occurrence of 'obscene' and 'insult' with 'toxicity' reported in Fortuna et al. (2020), these generalizations are not surprising. Additionally, *Ami* 'sexism' seems to contain many insults: BERT generalizes well over *Ami.sex* when trained on Kaggle insult (cf. *kaggl.insu*:0.74).

BERT trained on *TRAC* 'overt aggression' or on 'aggression' is capable of predicting 'abusive', 'offensive', and 'toxicity'-related categories. Thus, training on *trac.oag*, predicts well *founta.abus* (0.80), *fount.tox* (0.82), *kaggl.tox* (0.79) and *offen.tox* (0.72). This is similar to when it is trained on *trac.aggr* (0.80; 0.84; 0.66 and 0.74, respectively), which was to be expected since both *TRAC* 'aggressive' and *TRAC* 'overt aggressive' datasets share data. BERT also generalizes better over *kaggl.insu* when trained on *trac.oag* (0.70). On the other side, when these predicted categories are used for training, the models generalize worse over *TRAC* 'overt aggressive', and they do not generalize over *TRAC* 'covert aggressive' or *TRAC* 'aggressive'. This can be due to the fact that *TRAC* contains covert and overt instances of harmful behavior in general. As a result, when models trained on *TRAC* are applied to other datasets, they can still detect and flag the positive instances of more explicit harm. However, models trained on other datasets struggle to deal with data that mostly contain covert aggression.

Table 5 also shows that models trained on the 'hate speech' category of the different datasets generalize much worse. Thus, BERT trained on *Founta* 'hate speech' generalizes over *Stormfront* with a worse performance, and poorly over *W&H* (trained on *fount.hs*, BERT's performance on *storm.hs* is 0.71 and on *W&H.hs* 0.63).

[15] If not mentioned otherwise, we cite the BERT figures.

Poor performance is also observed for BERT trained on *Kaggle* 'hate speech' over *Founta* (trained on *kaggl.hs*, BERT's performance on *fount.hs* is 0.64. For the remaining datasets with hate speech categories (*W&H*, *Davidson*, *Hateval*, and *Stormfront*) the achieved generalization performance was even worse.

However, it is to be noted that in the case of more specific hate speech categories, a better generalization is observed; cf., e.g., the generalization over *W&H* 'racism' when trained on *Founta* 'hate speech' and over *Ami* 'sexism' when trained on *W&H* 'sexism' (trained on *fount.hs*, BERT's performance on *W&H.race* is 0.77, and when trained on *W&H.sex*, its performance on *Ami.sex* is 0.73), which opposes Arango et al.'s (2019) conclusion that *W&H* is a dataset with a low generalization potential due to its composition by messages of a low number of authors. Just the contrary: we found that certain categories of this dataset generalize when classifying sexism from *Ami*'s dataset.

Furthermore, SVM trained on *Ami* 'sexism' generalizes over *Davidson*'s 'offensive' test set (0.74), which indicates that *Davidson* may contain sexist offensive content. On the other side, as expected, *Hateval* 'hate speech' generalizes extremely well over *Ami* 'sexism' because both datasets share data (Fersini, Nozza et al., 2018). *Hateval* targets hate speech against immigrants and women, and Ami targets only hate speech against women (i.e misogyny). So this second dataset will miss the immigrant hate messages from *Hateval*. It also generalizes over 'offensive' in *Davidson* (trained on *hatev.hs*, ALBERT's performance on *david.offe* is 0.70 and trained on *hatev.hs*, fastText's performance on *ami.sex* is 0.99). This suggests that these three categories may be related and some sexist hate speech may be present in the 'offensive' samples of *Davidson*. This is surprising since the *Davidson* dataset is annotated with respect to both 'offensive' and 'hate speech' and both categories are mutually exclusive in this case. This means that there is probably sexist content annotated in the *Davidson* dataset as 'offensive', but not as 'hate speech'.

4.2.3. Comparison with previous studies

In this subsection, we compare the outcome of our cross-dataset experiments with those reported in previous works mentioned in Section 2.

First, it is difficult to compare our figures with those presented in Waseem et al. (2018), as we apply binary classification to all the standardized dataset categories, while Waseem et al. use multiclass classification. Still, some relevant observations can be made. Thus, the authors conclude that poor generalization values are achieved when training BOW and Average of Subword Embeddings on the *Davidson* dataset and testing on the *W&H* dataset (F1 score of 0.58) and the reverse (F1 score of 0.57). From the values in Table 5, we can equally conclude that these two datasets do not generalize very well among each other and that the generalization between both datasets is always below 0.70. Consider, e.g., the BERT model trained on *W&H*'s 'sexism' and tested on *Davidson*'s 'offensive' (0.69) and 'toxicity' (0.65), and the BERT model trained on *W&H*'s 'hate speech' and tested on *Davidson*'s 'offensive' (0.63) and 'toxicity' (0.63); or the ALBERT model trained on *Davidson*'s 'offensive' and tested on *W&H*'s 'sexism' (0.60).

On the other hand, it is easier to compare our results with the results of the other generalization studies, which tag all abusive language-related messages as 'positive' and the remaining messages as 'negative'. For instance, in Gröndahl et al. (2018) the same two datasets as by Waseem et al. (2018) are used in their binarized version. The reported macro F1 scores are below 0.49 for all of the setups. This performance is lower than what we reported above for our experiments. Karan and Šnajder also binarize the labels of the *W&H*, *TRAC*, and *Kaggle* datasets (Karan & Šnajder, 2018). They report a generalization performance across the different categories of F1 scores < 0.48. We achieved better scores when training BERT with *TRAC*'s 'aggression' and testing on *Kaggle*'s 'toxicity' (F1 score of 0.66) and *W&H*'s 'hate speech' (F1 score of 0.61); and when training BERT on *W&H*'s 'hate speech' and testing on *Kaggle*'s 'toxicity' (F1 score of 0.62).

Swamy et al. (2019) binarize the *W&H*, *Davidson*, *Offenseval*, and *Founta* datasets. Since they also use BERT as we do, in the majority of the cases, their and our results are comparable. In some cases, our model achieved a slightly higher performance. This is the case when BERT is trained on *Offenseval*'s 'toxicity' and tested on *Founta*'s 'toxicity' (0.91) and *Davidson*'s 'toxicity' (0.76); when it is trained on *Davidson*'s 'toxicity' and tested on *Founta*'s 'toxicity' (0.91) and when the training and test sets are reversed (0.77). Our figures are better when BERT is trained on *Davidson*'s 'toxicity' and tested on *Offenseval*'s 'toxicity' (0.81); when it is trained on *Founta*'s 'toxicity' and tested on *Offenseval*'s 'toxicity' (0.84) and *W&H* 'hate speech' (0.67); and when it is trained on *W&H*'s 'hate speech' and tested on *Davidson*'s 'toxicity' (0.63). The overall (slightly) better performance in our experiments may be due to the fact that we use $\text{BERT}_{\text{LARGE}}$, while Swamy et al. use $\text{BERT}_{\text{BASE}}$.

It is to be noted that Swamy et al. carry out an additional separate experiment in which they build models with all the categories present in the *Offenseval* dataset and test them also on all the categories of the other three datasets. Given that they report their results in terms of accuracy and not F1, we do not compare them with the outcome of our experiments.

In another study (Pamungkas & Patti, 2019), the authors binarize the categories of *W&H*, *Hateval*, and *Offenseval* datasets. For the experiments, they use LSTM and SVM, which both render a poorer performance than the one we achieved in our experiments. Compare, for instance, the case when BERT is trained on *Offenseval*'s 'toxicity' and tested on *Hateval*'s 'hate speech', and when it is trained on *W&H*'s 'hate speech' and tested on *Hateval*'s 'hate speech' (both with F1 = 0.60).

The above comparison of the outcome of our experiments with previous studies shows that the deep models we tested perform, in general, better. As a look at (Swamy et al., 2019) furthermore shows, different variants of BERT (in this case, $\text{BERT}_{\text{BASE}}$ vs. $\text{BERT}_{\text{LARGE}}$) also perform differently.

Table 7
Random forest model feature importance.

Features	Imp.
Original F1	0.22
Train category — toxicity	0.11
Vocabulary test	0.10
fastText	0.10
Train and test set proportion	0.07
Vocabulary train	0.06
test concept — toxicity	0.06
Training dataset size	0.06
BERT	0.05
Test concept — hate speech	0.05
Train concept — overt aggression	0.04
Test concept — offense	0.03
Training dataset percentage of positive instances	0.03
SVM	0.03
Test dataset size	0.02
ALBERT	0.02
Train concept — hate speech	0.02
Test dataset — davidson, founta	0.02
Test sn — twitter, facebook	0.01
Is same social network	0.01
Train dataset — trac, founta	0.01
Train concept — insult, obscene	0.01
Test dataset — trac, ami	0.01
Test concept — overt aggression, abusive, sexism, obscene	0.01
Remaining features	0.00

4.3. Outcome of model performance classification

The model performance classification aims to answer the second open research question raised in Section 2.3. Table 7 displays the feature importance of the 16 features obtained in the Random Forest classification experiment (F1 score of 0.64).

Four features are most informative. The importance of "original F1" (0.22) shows that for cross-dataset generalization it is relevant to start with a model that performs well in an intra-dataset scenario — something which has been ignored in previous studies. "Vocabulary-test" (0.10) proves also to be relevant. It is also worth pointing out that the generalization relies more on "vocabulary-test" (0.10) and less on "vocabulary-training" (0.06). It is advantageous to have in the training set a higher share of vocabulary of the test data in order to avoid 'out-of-vocabulary' words. It seems to be of no advantage to have in the test set a high percentage of vocabulary appearing in the training set.

FastText (0.10) proved to be a good predictor of a poor generalization potential. BERT (0.05), SVM (0.03), and ALBERT (0.02) had lower relevance, which suggests that they add little to the already considered fastText variable.

Regarding categories, the feature "Train category — toxicity" is also of relevance (0.11). This is in line with Karan and Šnajder (2018), who already pointed out that different dataset categories could affect generalization. In this case, 'toxicity' as a training set category led to good performance. One could expect that the feature "Test category — toxicity" is also of relevance; however, it seems not to offer any further information to "Train category — toxicity", since generalization profits from the use of 'toxicity' in both the training and test sets. The other category-related features contribute less, no matter whether they are used for training or testing. This also applies to all datasets and social network features.

With the works of Karan and Šnajder (2018) and Swamy et al. (2019) in mind, we expected the dataset size-related features to be of high importance. However, "Train and test set proportion" obtained only 0.07, "Training dataset size" 0.06, "Training dataset percentage of positive instance" 0.03, and "Test dataset size" 0.02. This might be due to the fact that dataset size-related variables have actually low variability in this and previous studies, but, rather, depend on the considered abusive language datasets. This would imply that both our experiments and previous research in the field are not the most suitable for assessing the effect of the dataset size-related variables. In this regard, our study provides a further insight that the presence of categories with different performance in the same dataset makes it even more difficult to find possible correlations between dataset size-related variables and performance; cf. Table 8.

4.4. Implications of the results of the experiments

The results of our experiments have some clear implications for the definition of the categories, new dataset development, and model construction in the context of abusive language detection. In what follows, we present the main insights that our experiments provide concerning a better model generalization in this field.

Table 8
Correlation between intra and cross model performance and dataset size features.

	Intra-dataset	Cross-dataset
	F1 macro	F1 macro
Training data total size	0.14	0.01
Training data total positive	0.31	0.30
Training data percentage positive	0.07	0.16
Testing data total size	0.15	0.24

Use carefully coarse-grained categories. Categories like 'toxicity', 'offensive', and 'abusive' correlate well between each other and lead to a good cross-dataset generalization. With this in mind, we could have concluded that broader coverage terms work well, and this would be aligned with the findings of Pamungkas and Patti (2019). However, there is evidence that this happened with generic umbrella terms not because they supposedly cover more general concepts. If this would be the case, other generic concept categories such as e.g., 'hate speech', would generalize better than more specific concept categories like, e.g., 'sexism'. On the other side, we observe that while 'hate speech' is difficult to generalize, models trained for an umbrella term (e.g., 'toxicity') that subsumes 'hate speech' lead to a rather good performance. We hypothesize that training with such general categories leads to generalization due to the subclasses' imbalance. In a previous study (Fortuna et al., 2020), which inspected the *Kaggle* dataset, we verified that the majority of the instances of this dataset belong to obscenity and insult and only a small percentage is labeled as 'hate speech'. The study also provided evidence that the Perspective API performance has high variability, which depends on the targeted 'toxicity' subcategory. For instance, concepts such as 'obscene' are better detected than 'hate speech'. Our present study shows (cf. Section 4.2.2) that general umbrella terms lead to a good generalization with 'obscenity' and 'insult' detection models, but not with 'hate speech'. We conclude that coarse-grained general categories may serve well. However, it is necessary to clearly define and quantify the particular phenomenon that a general category in a dataset is supposed to cover. The more specific categories, which subcategorize a generic category, should also be annotated such that an error analysis on the model performance can be conducted and it can be assessed whether models equally detect all the subcategories.

Prioritize fine-grained categories. Our results suggest that in the case of 'hate speech', more fine-grained categories would be more appropriate. When models were trained and tested on fine-grained categories such as 'sexism' or 'racism', better levels of generalization have been achieved. Thus, we can observe that the use of categories such as 'hate speech' does not help that much in terms of generalization; and that they are likely to contain message samples that largely vary across the datasets with respect to content and style and thus do not serve well as training categories. This also further buttresses the argumentation for a more fine-grained classification, e.g., in Fortuna, Rocha da Silva, Soler-Company, Wanner and Nunes (2019) and Salminen et al. (2018). A more fine-grained classification implies that during the dataset compilation and category definition phase, specific phenomena that define each category should be identified (cf. Table 2 for some details on the procedure for data collection). Thus, if during the dataset compilation, only messages targeting sexism and racism are collected, a model trained on this dataset will not generalize well with another hate speech dataset that targets, for instance, homophobia.

Avoid redundant labels. Categories such as 'toxicity', 'offensive' and 'abusive' correlate well between each other and lead to a good cross-dataset generalization when used as training categories. We assume that this is because they contain similar data across the datasets, which implies that there is some label redundancy and points (again) to the need to establish a coherent cross-dataset annotation schema of the type we introduced in Fortuna et al. (2020). With this in mind, it might be appropriate to introduce a new generic category term 'Abuse and Harms', to replace 'toxicity', 'offensive' and 'abusive'.[16] Also, 'sexism' in *W&H* and *Ami* achieved similar performance, which indicates that using the label of 'sexism' to refer to both avoids the need for an extra label.

Use standardized categories. Our experiments have shown that in order to be able to assess the generalization potential of the categories across datasets it is of utmost importance not to merge all positive dataset categories (as done in the majority of the previous works Karan & Šnajder, 2018; Pamungkas & Patti, 2019; Salminen et al., 2020; Swamy et al., 2019), but, rather, standardize the category labels across datasets, preserving the original labels (as much as possible) when building models.

Control dataset size variables. As already mentioned above, with the works by Karan and Šnajder (2018) and Swamy et al. (2019) in mind, we expected the dataset size and proportion of positive and negative classes to be relevant for generalization. However, our experiments did not confirm this hypothesis. When using different categories for the same dataset (e.g., 'hate speech', 'offensive' and 'toxicity' for the *Davidson* dataset), we can see that the performance of a model depends more on the training category than on the dataset size and there is no clear correlation between size and model performance. This is even more surprising as these figures seem to contradict the general idea in machine learning that large (in particular, training) data sets will lead to better model performance and generalization (Halevy, Norvig, & Pereira, 2009). However, this might be due to the fact that dataset size-related variables depend on the considered datasets. A thorough study of the effect of size-related variables should ensure that size variables are independent of datasets. In order to guarantee this, different samples from one dataset should be taken and compared.

[16] This term would also capture the recent insight of the Community reflected by the change of the title of the most popular workshop in the area from 'Abusive Language Workshop' to 'Workshop on Online Abuse and Harms' https://www.aclweb.org/portal/content/fourth-workshop-online-abuse-and-harms.

Better balance explicit vs. implicit abuse. Our results indicate that models still have problems with the identification of covert aggression (i.e., "implicit abuse"). This problem is at least partially due to the explicit search for specific key words during the early phases of data collection. Thus, it is common to search for offensive words from Hatebase (e.g., Davidson et al., 2017); for instance, in the *Founta* dataset, messages are partially selected because they contain offensive words and negative sentiment (cf. Table 2). Such a selection procedure introduces bias and reinforces the collection of more explicit abuse.

Provide information on users. We have shown that it is possible to reach a reasonable generalization level when using data from few authors (for instance, 'sexism' and 'racism' categories from the *W&H* dataset stem from few authors and were apt for training models that achieve a certain generalization when tested against certain categories from other datasets). In view of the results in Arango et al. (2019), we believe that with a higher number of authors in a dataset, a higher degree of generalization could have been achieved. Unfortunately, for most of the datasets that we used in our study, no information on the message authors is available, which undermines the study of this variable.

Evaluate classification models. We can also observe that the choice of the proper model is equally of primary relevance. Only a model that performs well in an intra-dataset classification setup has also a chance to perform well in an inter-dataset setup (transformer-based models are an example of such models). The evaluation of classification models with respect to their generalization potential and robustness is also of relevance in the context of cross-dataset studies. In a young research area such as abusive language detection, where so much subjectivity still prevails, testing models in a cross-dataset scenario brings valuable insights on the quality of dataset categories.

5. Conclusions and future work

We addressed two open questions related to the cross-dataset model generalization in the context of abusive online language: **1. Are the models or the datasets decisive for cross-dataset generalization?** and **2. Which model and dataset characteristics are important for generalization after all?** Regarding the 1st question, we have shown that both models and the nature of the categories within the datasets are relevant and should be taken into account when creating models that generalize. Regarding the 2nd question, we found that the intra-dataset model performance is the most relevant generalization predictor and have identified the types of categories that are more suitable as training categories for models with a generalization potential. Compared to the previous works on model generalization in the field of abusive language, our work is the first that attempts to predict generalization based on dataset features and model properties, by means of applying a Random Forest classifier. We use only public datasets and make our code available, such that it is replicable and reusable by the community and thus contributes to a hate speech-free internet.

Apart from answering the above two open questions, our work has shown how a cross-dataset generalization study can be used to detect similarity between datasets and dataset categories and help to come up with a uniform dataset categorization. As already (Vidgen et al., 2019), our study revealed the need for accurate and non-overlapping definitions of categories.

Several issues still need to be tackled. The use of merged datasets, with the application of a category conversion schema before the merge, which would allow for a more fine-grained classification, is another promising line of research as recent works have demonstrated good performance on merged datasets (Salminen et al., 2020). Another open question is whether the number of annotators for each dataset is a relevant feature for generalization.

Furthermore, we did not tackle so far multilingual hate speech classification and cross-data set generalization, which becomes increasingly relevant in the field (cf., e.g., Pamungkas, Basile, & Patti, 2020; Pamungkas & Patti, 2019). In order to obtain a first intuition, we carried out some preliminary experiments on sexism classification with 5 datasets, two for English (Fersini, Nozza et al., 2018; Waseem & Hovy, 2016), 1 for Spanish (Fersini, Rosso and Anzovino, 2018), 1 for Italian (Fersini, Nozza et al., 2018), and 1 for Portuguese (Fortuna, Rocha da Silva et al., 2019), using multilingual BERT. The experiments showed a poor generalization between multilingual datasets (cf. Table 9). Only the generalization from English (Fersini, Nozza et al., 2018) to Portuguese (Fortuna, Rocha da Silva et al., 2019) (macro F1 = 0.67) indicated a borderline better generalization. The best result is achieved when generalizing across both English datasets (macro F1 = 0.70). This indicates that a multilingual generalization approach is *per se* likely to have a poorer performance than an intra-lingual approach — although many questions remain open. Further investigation is needed as a multilingual approach may work better with other concepts, models, and possibly combined with dataset merging.

CRediT authorship contribution statement

Paula Fortuna: Conceptualization, Methodology, Software, Validation, Investigation, Resources, Data curation, Writing - first submitted version, Writing - revised version & editing, Funding acquisition. **Juan Soler-Company:** Conceptualization, Supervision, Project administration, Writing - first submitted version, Writing - revised version & editing. **Leo Wanner:** Conceptualization, Supervision, Writing - first submitted version, Writing - revised version & editing, Funding acquisition.

Acknowledgments

The first author is supported by the research grant SFRH/BD/143623/2019, provided by the Portuguese national funding agency for science, research and technology, Fundação para a Ciência e a Tecnologia (FCT), within the scope of Operational Program *Human Capital* (POCH), supported by the European Social Fund and by national funds from MCTES. The work of the second and third authors has been supported by the European Commission in the context of the H2020 Research Program under the contract numbers 700024 and 786731.

Table 9
Intra and cross dataset model generalization evaluation of Multilingual BERT on 5 sexism related datasets, *Ami_sex_En* (Fersini, Nozza et al., 2018), *Ami_sex_It* (Fersini, Nozza et al., 2018), *Ami_sex_Sp* (Fersini, Rosso et al., 2018), *Fort_sex_Pt* (Fortuna, Rocha da Silva et al., 2019) and *W&H_sex_En* (Fersini, Nozza et al., 2018; Waseem & Hovy, 2016).

		Train				
		Ami En	Ami It	Ami Sp	Fort Pt	W&H En
Test	Ami En	0.81	0.34	0.63	0.57	0.70
	Ami It	0.44	0.87	0.52	0.39	0.54
	Ami Sp	0.59	0.44	0.76	0.47	0.52
	Fort Pt	0.67	0.25	0.63	0.80	0.62
	W&H En	0.67	0.20	0.57	0.52	0.86

References

Almerekhi, H., Kwak, H., Jansen, B. J., & Salminen, J. (2019). Detecting toxicity triggers in online discussions. In *HT '19, Proceedings of the 30th ACM conference on hypertext and social media* (pp. 291–292). New York, NY, USA: Association for Computing Machinery, ISBN: 9781450368858.

Arango, A., Pérez, J., & Poblete, B. (2019). Hate speech detection is not as easy as you may think: A closer look at model validation. In *SIGIR'19, Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval* (pp. 45–54). New York, NY, USA: Association for Computing Machinery, ISBN: 9781450361729.

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*, Article 101584.

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* (pp. 759–760). International World Wide Web Conferences Steering Committee.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., et al. (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* (pp. 54–63). Minneapolis, Minnesota, USA: Association for Computational Linguistics.

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.

Chandrasekharan, E., Samory, M., Srinivasan, A., & Gilbert, E. (2017). The bag of communities: Identifying abusive behavior online with preexisting internet data. In *CHI '17, Proceedings of the 2017 CHI conference on human factors in computing systems* (pp. 3175–3187). New York, NY, USA: Association for Computing Machinery, ISBN: 9781450346559.

Davidson, T., Warmsley, D., Macy, M. W., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the eleventh international conference on web and social media, ICWSM 2017, MontréAl, QuéBec, Canada, May 15-18, 2017* (pp. 512–515). AAAI Press.

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)* (pp. 11–20). Brussels, Belgium: Association for Computational Linguistics.

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)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.

d'Sa, A. G., Illina, I., & Fohr, D. (2020). BERT and fasttext embeddings for automatic detection of toxic speech. In *SIIE 2020-information systems and economic intelligence*.

Fersini, E., Nozza, D., & Rosso, P. (2018). Overview of the evalita 2018 task on automatic misogyny identification (AMI). In T. Caselli, N. Novielli, V. Patti, & P. Rosso (Eds.), *CEUR workshop proceedings: vol. 2263, Proceedings of the sixth evaluation campaign of natural language processing and speech tools for Italian. final workshop (EVALITA 2018) co-located with the fifth italian conference on computational linguistics (CLiC-It 2018), Turin, Italy, December 12-13, 2018*. CEUR-WS.org.

Fersini, E., Rosso, P., & Anzovino, M. (2018). Overview of the task on automatic misogyny identification at ibereval 2018. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, & J. C. de Albornoz (Eds.), *CEUR Workshop Proceedings: vol. 2150, Proceedings of the third workshop on evaluation of human language technologies for iberian languages (IberEval 2018) co-located with 34th conference of the spanish society for natural language processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018* (pp. 214–228). CEUR-WS.org, URL http://ceur-ws.org/Vol-2150/overview-AMI.pdf.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computer Surveys, 51*(4), 85:1–85:30.

Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L., & Nunes, S. (2019). A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the third workshop on abusive language online* (pp. 94–104). Florence, Italy: Association for Computational Linguistics.

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* (pp. 6788–6796). Marseille, France: European Language Resources Association.

Fortuna, P., Soler-Company, J., & Nunes, S. (2019). Stop propaghate at semeval-2019 tasks 5 and 6: are abusive language classification results reproducible?. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 745–752). Minneapolis, Minnesota, USA: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/S19-2131, URL https://www.aclweb.org/anthology/S19-2131.

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., et al. (2018). Large scale crowdsourcing and characterization of Twitter abusive behavior. In *Proceedings of the twelfth international conference on web and social media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018* (pp. 491–500). 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* (pp. 260–266). Varna, Bulgaria: INCOMA Ltd..

Golbeck, J., Ashktorab, Z., Banjo, R. O., Berlinger, A., Bhagwan, S., Buntain, C., et al. (2017). A large labeled corpus for online harassment research. In P. Fox, D. L. McGuinness, L. Poirier, P. Boldi, K. Kinder-Kurlanda (Eds.), *Proceedings of the 2017 ACM on web science conference, WebSci 2017, Troy, NY, USA, June 25 - 28, 2017* (pp. 229–233). ACM, http://dx.doi.org/10.1145/3091478.3091509.

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018). All you need is "Love": Evading hate speech detection. In *AISec '18, Proceedings of the 11th ACM workshop on artificial intelligence and security* (pp. 2–12). New York, NY, USA: Association for Computing Machinery, ISBN: 9781450360043.

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

Herz, M., & Molnár, P. (2012). *The content and context of hate speech: rethinking regulation and responses*. Cambridge University Press.

Jigsaw (2019). Toxic comment classification challenge: identify and classify toxic online comments. Available in https://www.kaggle.com/c/jigsaw\discretionary{-}{}{}toxic-comment-classification-challenge, accessed last time in November 2019.

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

Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., & Taboada, M. (2020). The SFU opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, *4*(2), 155–190.

Kreyszig, E. (1960). *Advances engineering mathematics*. Wiley Eastern.

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

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: a lite BERT for self-supervised learning of language representations. In *8th international conference on learning representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In Y. Bengio, & Y. LeCun (Eds.), *1st international conference on learning representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, workshop track proceedings*.

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).

Mossie, Z., & Wang, J.-H. (2020). Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, *57*(3), Article 102087.

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2019). A BERT-based transfer learning approach for hate speech detection in online social media. In H. Cherifi, S. Gaito, J. F. Mendes, E. Moro, & L. M. Rocha (Eds.), *Studies in Computational Intelligence*: *vol. 881*, *Complex networks and their applications VIII - Volume 1 proceedings of the eighth international conference on complex networks and their applications COMPLEX NETWORKS 2019, Lisbon, Portugal, December 10-12, 2019* (pp. 928–940). Springer, http://dx.doi.org/10.1007/978-3-030-36687-2_77.

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., & Yeung, D.-Y. (2019). Multilingual and multi-aspect hate speech analysis. 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)* (pp. 4675–4684). Hong Kong, China: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D19-1474, URL https://www.aclweb.org/anthology/D19-1474.

Pamungkas, E. W., Basile, V., & Patti, V. (2020). Misogyny detection in Twitter: a multilingual and cross-domain study. *Information Processing & Management*, *57*(6), Article 102360.

Pamungkas, E. W., & 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* (pp. 363–370). Florence, Italy: Association for Computational Linguistics.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., & and, O. G. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2020). Scikit-learn support vector classification. Available at https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html, accessed last time October.

Roberts, S. T. (2019). *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.

Salawu, S., He, Y., & Lumsden, J. (2020). Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions Affective Computing*, *11*(1), 3–24.

Salminen, J., Almerekhi, H., Milenkovic, M., Jung, S., An, J., Kwak, H., et al. (2018). Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Proceedings of the twelfth international conference on web and social media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018* (pp. 330–339). AAAI Press.

Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S.-g., Almerekhi, H., & Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, *10*(1), 1.

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* (pp. 1–10). Valencia, Spain: Association for Computational Linguistics.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47.

Swamy, S. D., Jamatia, A., & Gambäck, B. (2019). Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)* (pp. 940–950). Hong Kong, China: Association for Computational Linguistics.

Uglow, H., Zlocha, M., & Zmyslony, S. (2019). An exploration of state-of-the-art methods for offensive language detection. CoRR abs/1903.07445.

van Aken, B., Risch, J., Krestel, R., & Löser, A. (2018). Challenges for toxic comment classification: an in-depth error analysis. In *Proceedings of the 2nd workshop on abusive language online (ALW2)* (pp. 33–42). Brussels, Belgium: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/W18-5105, URL https://www.aclweb.org/anthology/W18-5105.

Vidgen, B., & Derczynski, L. (2021). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, *15*(12), 1–32.

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* (pp. 80–93). Florence, Italy: Association for Computational Linguistics.

Waseem, Z. (2016). Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science* (pp. 138–142). Austin, Texas: Association for Computational Linguistics.

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* (pp. 88–93). San Diego, California: Association for Computational Linguistics.

Waseem, Z., Thorne, J., & Bingel, J. (2018). Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In *Online harassment* (pp. 29–55). Springer.

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *WWW '17*, *Proceedings of the 26th international conference on world wide web* (pp. 1391–1399). Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee.

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