Personal and Ubiquitous Computing
https://doi.org/10.1007/s00779-021-01609-1

ORIGINAL ARTICLE

Towards multidomain and multilingual abusive language detection: a survey

Endang Wahyu Pamungkas[1] · **Valerio Basile**[1] · **Viviana Patti**[1]

Received: 14 February 2021 / Accepted: 12 July 2021

Abstract
Abusive language is an important issue in online communication across different platforms and languages. Having a robust model to detect abusive instances automatically is a prominent challenge. Several studies have been proposed to deal with this vital issue by modeling this task in the cross-domain and cross-lingual setting. This paper outlines and describes the current state of this research direction, providing an overview of previous studies, including the available datasets and approaches employed in both cross-domain and cross-lingual settings. This study also outlines several challenges and open problems of this area, providing insights and a useful roadmap for future work.

Keywords Abusive language detection · Hate speech detection · Literature review · Multidomain · Multilingual

1 Introduction

Abusive language is becoming a relevant issue in social media platforms such as Facebook and Twitter. The rise of the phenomenon is also due to the anonymity given to users and to the lack of effective regulation provided by these platforms. On the one hand, social media provide a facility for improving connectedness between people with their relations. On the other hand, this facility is often exploited to propagate toxic content such as hate speech or other forms of abusive language. Given the current rate of user-generated content produced in every minute, manually monitoring abusive behavior in social media is impractical. Facebook and Twitter also made efforts to eliminate abusive content from their platforms[1] by providing clear policies on hateful conducts[2], implementing user report mechanisms, and employing content moderators to filter the abusive posting. However, these efforts are not a scalable and long-term solution to this problem.

Several studies from the Natural Language Processing (NLP) field have been done to tackle the problem of abusive language in social media. Most studies proposed a supervised approach to detect abusive content automatically using various models ranging from traditional machine learning approaches to recent neural-based approaches. Moreover, the majority of current studies only focused on a single language, i.e., English, and a single abusive language phenomenon, e.g., hate speech, sexism, racism, and so on, rather than multiple phenomena and how they are interconnected. However, abusive language in social media is not limited to specific languages, and it features multiple abusive phenomena. As a matter of fact, the most popular social media, such as Twitter and Facebook, are multilingual, as users are encouraged to express themselves spontaneously in their mother tongue, and online social conversations are characterized with multiple different topics. Therefore, in a variety of languages and contexts there is a considerable urgency to prevent online hate speech from spreading virally, becoming a significant factor in grave crimes against minorities or vulnerable categories. Specifically, robust approaches are needed for abusive language detection in a multidomain and multilingual environment, which will also enable the implementation of effective tools that could be employed to support both monitoring and content moderation activities such as

[1] https://time.com/5739688/facebook-hate-speech-languages/

[2] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

✉ Endang Wahyu Pamungkas
pamungka@di.unito.it

Valerio Basile
valerio.basile@unito.it

Viviana Patti
viviana.patti@unito.it

[1] Dipartimento di Informatica, University of Turin, Turin, Italy

Published online: 11 August 2021

automatic moderation and flagging of potentially hateful users and posts, also for guaranteeing a better compliance to governments demands to counteract the phenomenon [37]. A few works initiated cross-domain and cross-lingual studies on abusive language detection to tackle these aspects of the problem [48, 64, 94, 134]. However, some difficulties and issues still remain to obtain a robust model to detect abusive language across different domains and languages.

In this paper, we summarize the recent development of studies in the detection of abusive language in social media across domains and languages. Through this survey, we present a systematic overview of research conducted in this research area, providing a comprehensive view of the state of the art and datasets that are centered on this area. Our main objective is to draw a conclusion on the state of the art and to provide several possible opportunities for future work based on the existing open problems.

After the introduction, in Section 2, we describe several previous surveys on abusive language detection and related topics. We discuss the existing studies in multidomain abusive language detection, including a review of available datasets that could be exploited for this task in Section 3. Section 4 presents a comprehensive review of multilingual abusive language detection studies which covers state-of-the-art approaches and available datasets. An analysis of challenges and opportunities for this particular task in future work is discussed in Section 5. Finally, Section 6 presents conclusive remarks of this survey.

2 Related work

Few recent works focused on analyzing the current challenges in abusive language detection task based on existing works. Jurgens et al. [63] presented a position paper that outlines current challenges to fight online abuse and proposes several strategies to address them. They argued that most existing studies only focus on a narrow definition of abuse, and expanding the problem scope is needed in order to deal with more subtle but serious forms of abusive behavior, such as microaggressions. Secondly, they opined that we need to develop proactive technologies to counter abusive in the future, rather than only focus on the automatic detection perspective. Finally, they postulated that the community should take a role in contextualizing its effort inside the broader framework of justice, including explicit capabilities, restorative justice, and procedural justice, to support and promote a healthier community. Another work by Vidgen et al. [137] presented challenges and frontiers in abusive content detection. They outlined several challenges of the abusive content detection task from three different perspectives. From a research point of view, there are three challenges: the difficulties in categorizing abusive content, recognizing abusive content, and accounting for context. The dataset creation and distribution as well as ethical issues are the main challenges from the community perspective. They also outlined challenges based on research frontiers, which cover several issues in multimedia content that are not yet much explored, implementation of fairness and explainability, and cross-domain applications. MacAveney et al. [71] outlined and explored the current challenges of hate speech detection tasks in text. To understand the problem, they proposed a straightforward multi-view SVM approach to provide better interpretability than more complex neural models. Based on the experiment, they found two remaining issues in hate speech detection in the text, namely (i) the change of perspectives towards topic or issue over time; (ii) hate speech detection is a closed-loop system that only focuses on the current characteristics of the phenomenon, while the spreader of hate speech always looks at ways to outsmart the system.

The scientific study of abusive language, especially in the NLP field, has been growing incredibly fast in the last five years. The work of Schimdt and Wiegand [125] was the first study to provide a short, comprehensive, and systematic overview of hate speech detection tasks. This work presented what has been done so far in the hate speech detection task, focusing on the feature extraction approach. However, they also have several dedicated sections to describe bullying, classification approaches, available datasets with their annotation procedure, and the overall challenges of hate speech detection tasks. The work of Fortuna and Nunes [43] complements the aforementioned work by providing a more in-depth critical review of this area. Firstly, they presented more detailed discussion on the definition of hate speech based on several previous proposals from other studies. They also reviewed the feature extraction approach by classifying it into generic text mining features and specific hate speech detection features. A complete description of available datasets, including their collection and annotation approaches, was also provided. Finally, they outlined challenges and opportunities as outcome of their study, to provide better insight into future research development. Mishra et al. [76] also aimed to provide a comprehensive view on online abuse detection tasks. This study outlined the existing datasets and reviews the approaches to deal with this issue, including analyzing their strengths and weaknesses. In their conclusions, they highlighted the remaining challenges in the field and provide insights for future development: (i) the study of abusive language detection is still only focusing on specific languages and also specific abusive phenomena; (ii) most current approaches are vulnerable to the obfuscation of words; (iii) the difficulty to deal with the implicit abuse; (iv) the ever-changing nature of abusive phenomena makes the detection of new emerging phenomena difficult.

If we focus on the topic of *resources* for the detection of abusive phenomena in particular, there are two very recent survey studies, providing a critical review of the available resources, datasets and benchmark corpora for abusive language detection. Vidgen and Derczynski [136] presented a critical analysis of available abusive language datasets by discussing the goals underlying their development, the introduced taxonomies, and the annotation procedure. They also elaborated on the different ways to share datasets, including the introduction of the website https://hatespeechdata.com/, which is meant as a constantly updated catalogue of datasets annotated for hate speech, online abuse, and offensive language. Finally, they presented best practices for creating abusive language datasets based on the findings of the study. Similarly, Poletto et al. [103] provided a systematic review of resources and benchmarks for hate speech detection tasks. They described different strategies to develop datasets for hate speech detection based on five comparison perspectives, including type, topical focus, data source, annotation procedure, and language. They also provided an overview of all available resources for hate speech detection tasks based on their type, which covers corpora, resources released for shared tasks, and also lexica. Finally, they introduced a reflection on the impact of keywords used to collect the data when creating the hate speech corpora. Overall, these recent surveys on language resources capture and underline a great availability of benchmark datasets for the evaluation of abusive language and hate speech detection systems in several languages and with several topical focuses. The take away message is that such availability lays the foundation to address the urgent challenge of investigating architectures which are stable and well-performing across different languages and abusive domains. However, none of these works cover the multilingual and multidomain perspective and the related challenges specifically and extensively, while this is the main issue we address in the current work, with the main aim to develop a roadmap for scholars active in the field and a compass for future work.

3 Multidomain abusive language detection

Abusive language behavior is multifaceted and available datasets are characterized by different topical focuses. *Abusive language* is generally used as an umbrella term [143], covering several sub-categories, such as cyberbullying [55, 131], hate speech [33, 144], toxic comments [150], offensive language [152] and online aggression [67]. Several datasets have been proposed having different topical focuses, e.g., misogyny, racism, sexism, and so on, and sourced from different platforms, e.g., Facebook and Twitter. Most studies in this area also tend to focus on one topical focus, which makes difficult to quantify whether a model or feature set which perform well in one dataset is transferrable to other datasets [125, 145].

However, the abusive language phenomena are not constrained to one particular topical focus and platform. Therefore, having a robust model to detect abusive language across different topical focuses and platforms is important. Some existing studies proposed cross-domain abusive language detection [48, 64, 94, 134]. A model is trained on one specific dataset with a specific domain and tested in another dataset with a different domain. In this study, the domain term is used to describe both topical focuses and platforms. It has been stated that ensuring that a model can detect abusive language across different domains is one of the main challenges and an important frontier [137]. The cross-domain setting is also explored by Wiegand et al. [146] to prevent bias contained in the training data, as they experimentally found several biases in currently popular abusive language datasets, including topic bias and author bias. In this section, we discuss recent studies on cross-domain abusive language detection. We review available datasets that could be exploited for this task, focusing on English. Furthermore, we also describe several approaches that have been proposed in this research direction.

3.1 What datasets are available for multidomain abusive language detection?

In this section, we collect information about the available datasets from existing studies on abusive language detection across different domains. Several previous works in abusive language detection defined a domain as a topical focus [94, 134], such as hate speech, cyberbullying, and offensiveness. In contrast, some others describe it as platforms [48, 64] such as Twitter, Facebook, and Youtube. We select English datasets by focusing on topical focus and platform variety. The collection of abusive language datasets in languages other than English is also available in Section 4. We mainly extract this information from the two most recent survey studies on abusive language resources. First, Vidgen and Derczynski [136] provided the analysis of available training data for abusive language detection tasks and proposed best practices in creating training data of abusive language based on existing studies. Meanwhile, Poletto et al. [103] presented a more comprehensive study on resources and benchmarks available for hate speech detection tasks based on several aspects. We also add datasets from several shared tasks that were not covered by these works and a few datasets from very recent studies that were not available when these articles were published. Table 1 summarizes our findings on the available datasets for this research purpose. We discuss a more in-depth comparison between datasets and other aspects we need to consider when using these

Table 1 Summarization of available abusive language dataset across different topical focuses and sources (English only)

Topical focus	Sources	Entries	Available	Ref
Hate speech	Twitter	24,802	Yes	[33]
	Twitter	27,330	Yes	[36]
	Twitter	62 millions	No	[46]
	Stormfront	10,568	Yes	[47]
	Youtube	24,840	No	[54]
	Twitter and Reddit	150 millions	No	[86]
	Gab and Reddit	56,100	Yes	[105]
	Twitter	3.5 millions	No	[106]
	Twitter	18,667	No	[107]
	Twitter	4,000	No	[138]
	Twitter	16,907	Yes	[144]
	Twitter	13,000	Yes	[11]
	Facebook	1,288	Yes	[25]
	Twitter	4,972	Yes	[114]
	Twitter	5,647	Yes	[89]
	Twitter	149,823	Yes	[50]
	Twitter & Facebook	7,005	Yes	[73]
Toxicity	News Site	1,043	No	[65]
	Wikipedia	115,737	Yes	[150]
	Twitter	6,774	Yes	[108]
Cyberbullying	Youtube	2,235	No	[127]
	Gaming Platforms	34,329	No	[16]
	Formspring4	13,160	Yes	[116]
	Twitter & Formspring3	13,000	No	[155]
	Instagram	25,000	No	[56]
	Tweet	9,484	No	[21]
Offensiveness	Reddit	168 millions	No	[84]
	Reddit	11 millions	No	[124]
	Twitter	14,100	Yes	[153]
	Twitter	9 millions	Yes	[154]
Abusiveness	News Site	3.1 millions	No	[85]
	Twitter	80,000	Yes	[45]
	News Sites	2,000	No	[22]
Flames	News Site	5,077	Yes	[133]
Harassment	Twitter	25,000	Yes	[113]
	Twitter	35,000	Yes	[49]
Misogyny	Twitter	3,977	Yes	[41]
	Twitter	5,000	Yes	[39]
	Twitter	6,000	Yes	[40]
Sexism	Twitter	712	Yes	[61]
Aggressiveness	Facebook	15,000	Yes	[67]
	Twitter and Facebook	5,000	Yes	[68]

datasets for multidomain abusive language study based on existing works in the following.

Topical focus The motivation for several multidomain abusive language detection studies is to have a robust model that generalizes the problem across different topical focuses. Topical focus usually includes the addressed abusive phenomena, as well as the specific targets of the abusive behavior. However, some topics overlap with each other, i.e., misogyny and sexism or xenophobia and

racism, due to a certain degree of subjectivity in defining these phenomena. The topical focus information presented in Table 1 is based on the information provided in the publications which accompany the proposed resources. However, some of these papers did not include a clear definition of the addressed phenomena. We observe that hate speech is the most covered topic by previous studies. However, on some hate speech datasets, we also discover other abusive phenomena such as offensiveness [33], racism [144], and sexism [144]. In this manner, a cross-domain abusive language detection setting means training a model on one or more topical focuses and testing it on completely different topical focuses.

Sources Another objective of abusive language detection in the multidomain setting is to have a robust model to detect abusive content across different platforms. This task is also challenging since the available datasets are retrieved from various platforms, and every platform has different characteristics and uniqueness. Based on the information presented in Table 1, Twitter is the most studied platform for capturing the abusive phenomena. This is possibly due to the convenience of scraping tweet samples using the available Twitter API and the less strict policy on making the data publicly available. Facebook is another popular social media which becomes a data source by several studies. Other studies exploited news sites, online forums, and Youtube comments for gathering their data. Most studies used several defined keywords to query the data from the platforms mentioned above. Some of them used offensive words [32, 39, 41, 89], which are usually a strong signal of abusive content, while other studies decided to use more neutral keywords to maintain a real-world approach to the problem [11], or even both offensive and neutral keywords [152]. Some other works also exploited specific keywords related to some events that trigger abusive phenomena [144].

Availability In Table 1, we provide information about the availability of the datasets. We manually check the published papers and mark a dataset as available when the authors explicitly mention the link to the dataset repository or state that the dataset is available for research purposes upon request. We can see that 26 out of 39 datasets were made available by their authors.[3] Most available datasets were obtained from Twitter, likely due to their policy or other regulation restricting data sharing from other sources such as Reddit, Youtube, and news sites. However, we also notice that some Twitter datasets are shared by only providing the tweet identifier [45, 144] and allow users to download them by using the publicly available Twitter API. In this case, the number of entries could decrease due to the data decay (tweets were already deleted or are simply not available anymore).

Annotation scheme This information is not provided in Table 1, but we perform a manual inspection regarding the annotation scheme of every dataset. Most datasets have binary labels, including abusive and not abusive class. Some other datasets have a multiclass annotation, capturing different abusive phenomena. For example, Davidson et al. [33] labeled not only the hateful tweets but also their offensiveness. Similarly, Waseem and Hovy [144] proposed to label racism and sexism separately. Some studies also proposed a finer-grained annotation scheme to capture more in-depth abusive phenomena. For example, Fersini et al. [39, 41] provided three layers of annotation to capture the misogyny phenomenon (misogyny or not), misogyny category and behavior (stereotype, dominance, derailing, sexual_harassment, and discredit), and the target of misogyny (active or passive). In a multidomain or cross-domain classification task, one of the most important steps is to unify the label annotation of every dataset. Most existing works modeled this task as a binary classification task [64, 94]. Therefore, they cast the multiclass annotation to binary annotation by combining different abusive phenomena into one class. In the case of finer-grained annotation, they only took the first layer of annotation, where the data is mainly annotated as either abusive or not abusive.

Data distribution Data distribution also needs to be considered in the multidomain and cross-domain classification task, especially the percentage of abusive samples in the dataset. The different label distribution between training and testing sets would make the performance evaluation and comparison between systems difficult [92, 94]. Specifically, when systems are trained on skewed distributions of labels, with few examples in the abusive class, they will struggle to detect the abusive class on the test set, resulting in a higher rate of false negatives. Pamungkas et al. [93] observed that balancing the distribution in the training set improves the f1-score of the positive class significantly. Based on our investigation, the class distribution of abusive language datasets varies considerably, mostly depending on how the data is sampled and on the source of the data. However, we observe that most abusive language datasets have a lower percentage of abusive content than neutral content, with some datasets only containing less than 20% of abusive instances [45, 47, 150]. Some studies experimentally found that systems often struggle to detect the under-represented class, resulting in low f1-scores on the positive class (abusive label), which is an issue for real-world abusive language detection systems [58, 93]. Maintaining a uniform label

[3] The links to the available datasets can be seen in Appendix Table 5.

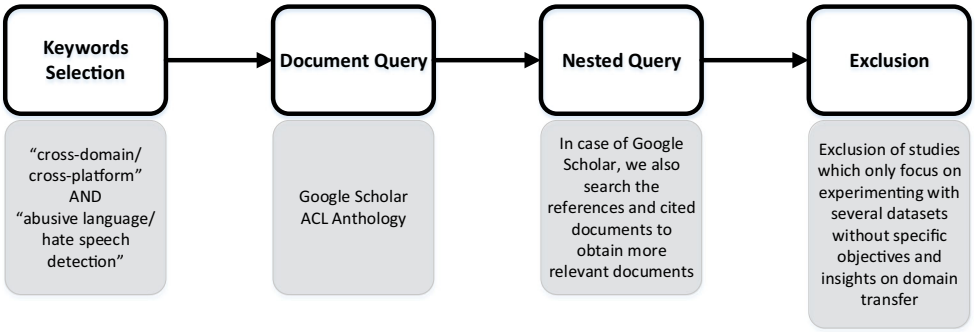

Fig. 1 Documents collection methodology

distribution between training and test set was an approach often followed to provide a comparable evaluation in cross-domain classification [58, 134]. This approach, however, does not necessarily provide an accurate estimate of the robustness of the model in a realistic scenario, where the amount of abusive language could drastically change.

3.2 What has been done so far in multidomain abusive language detection study?

This section presents studies that have been done in abusive language detection, which focus on building robust models across different domains. We collect any publication found on Google Scholar by using four main keywords, namely "cross-domain abusive language detection", "cross-domain hate speech detection", "cross-platform abusive speech detection", and "cross-platform hate speech detection". These keywords are chosen after several observations using different keyword combinations. We limit our query to the first five pages for each keyword and sort results based on relevance, without a time filter. Furthermore, we also check each document's cited documents and references on the first five pages to get more relevant publications. To avoid missing on the very recent works, we also exploit the same keywords on the proceeding of the last three years' main NLP conferences on ACL Anthology platforms[4]. Finally, we exclude some works which only experiment with different datasets, without any objectives and insights about domain-agnostic models. Figure 1 summarizes the methodology for the document collection in this survey study.

We carefully read each work to obtain several key pieces of information to be discussed in this study. Table 2 summarizes the full list of works in this direction. Most studies only focused on English, and we only observe two studies that work on Italian [27] and Arabic [24]. Most of the chosen studies conducted a cross-domain experiment, where the domain can be refer to topical focuses or platforms. We also noticed that this research focus is still relatively new, with the earliest works were initiated in 2018 [64, 145, 147]. All studies adopted a supervised approach by training a model on a training set and predicting instances on the test set. Following, we provide a deeper discussion to compare each work based on the models (traditional machine learning based, neural based, or transformer based), features (a very wide variant of features), and approaches adopted to deal with domain-shift specifically.

3.2.1 Models

A wide variety of models was adopted to deal with this task. Some studies exploited traditional machine learning approaches such as linear support vector machine classifiers (LSVC) [64, 92, 94], logistic regression (LR) [121], and support vector machine (SVM) [24, 147]. Their argument for adopting the traditional approach was to provide better explainability of the knowledge transfer between domains. Some other studies adopted several neural-based models, including convolutional neural networks (CNN) [75, 141], long short-term memory (LSTM) [8, 75, 92, 94, 145], bidirectional LSTM (Bi-LSTM) [115], and gated recurrent unit (GRU) [27]. The most recent works focus more on investigating transferability or generalizability of state-of-the-art transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT) [19, 48, 66, 79, 83, 90, 92, 134] and its variant like RoBERTa [48] in the cross-domain abusive language detection task.

In the early phases of cross-domain abusive language detection, specific models which adopt joint-learning [115] and multi-task [145] architectures achieved the best performance. These architectures were proven to be effective for transferring knowledge between domains. However, in the latest studies, transformer-based models succeed in achieving state-of-the-art results. The most recent study by Glavas et al. [48] shows that ROBERTa outperformed other models such as BERT in the cross-domain setting of the hate speech detection task. This result confirms a recent finding on other natural language processing tasks [18], i.e., that a pre-training language

[4] https://www.aclweb.org/anthology/

Table 2 Summary of approaches adopted by existing studies for cross-domain abusive language detection tasks

Models	Approach	Year	Ref
Traditional model	Proposed to employ a SVM model with a novel abusive lexicon and exploited it in the cross-domain abusive language detection task, providing domain-independent knowledge.	2018	[147]
Traditional model	Employed a linear SVM coupled with a domain adaptation approach called FEDA, which works by duplicating features several times across domains to allow the model to learn domain-dependent weights for each feature.	2018	[64]
Neural based	Experimented with a multitask learning approach, which allows the model to learn the task from two or more tasks sequentially by sharing the learning parameters and combining the loss functions of the respective tasks.	2018	[145]
Neural based	This work experimented by combining datasets from different platforms to train the GRU-based model and exploit different sets of features.	2019	[27]
Neural based	Exploited a specific hateful lexicon called HurtLex to provide domain-independent features for two supervised models including a linear SVM and a LSTM in a cross-domain abusive language task.	2019	[94]
Neural based	Proposed a joint-learning architecture based on ELMo Embeddings, which allows the model to learn the task from two datasets sequentially, obtaining more robust performance.	2019	[115]
Transformer based	This work aims to study the transferability of the current state-of-the-art BERT model, so no specific approach is proposed to tackle domain transfer.	2019	[134]
Neural model	This study proposed several LSTM-based models that only focuses on using text information (char n-grams and word embedding) representation for building platform-agnostic hate speech detector, but they did not conduct any cross or multidomain experiment to evaluate their model.	2019	[75]
Transformer based	Experimented with a BERT-based classifier and topic modeling approach, which show that removing domain-specific instances improve the model's out-domain performance	2020	[83]
Neural based	Proposed several representations including target, content, and linguistic behavior and used cross attention gate flow to refine these representations, providing better domain-transfer knowledge.	2020	[141]
Transformer based	Infused specific hateful lexicon called HurtLex into BERT model to transfer knowledge across domains.	2020	[66]
Multiple models	Besides experimented with a wide coverage of models including traditional (linear SVM), (LSTM), and (BERT), they also exploited HurtLex as domain-independent features for knowledge transfer between domains.	2020	[92]
Neural based	Experimented with augmenting all training data from different domains, resulting in the performance improvement of the models based on BERT and RoBERTa representation.	2020	[48]
Transformer based	It is proposed to retrain BERT with a big abusive language corpus obtained from Reddit called HateBERT, which shows a promising result in the cross dataset experiment.	2020	[19]
Traditional models	They tested the generalizability of wide-coverage traditional models (logistic regression, naïve Bayes, support vector machine, XGBoost, feed-forward neural network) coupled with also a wide range of feature representation in detecting hate speech across different platforms.	2020	[121]
Transformer based	Experimented by combining several datasets to train the model based on BERT and proven to be effective in detecting uncivil language across multiple domains, it outperformed several fine-tuning strategies.	2020	[90]
Transformer based	Proposed to use existing regularization method to re-weight input samples which succeeded to decrease the racial bias of the dataset, resulting in the improvement of the BERT-based model's performance in cross-domain classification settings	2020	[79]
Neural based	This study reproduced the state-of-the-art models to evaluate the dataset bias issue in abusive language task based on the cross-dataset classification study.	2020	[8]

Table 2 (continued)

Models	Approach	Year	Ref
Traditional model	This study proposed a novel multiplatform abusive language dataset. The proposed model for the experiment is the standard SVM without a specific approach to deal with domain-shift issue.	2020	[24]
Neural based	This work proposed a deep exploration to deal with cross-domain and cross-target hate speech detection. They also proposed a multitask architecture, which allows the model to learns hate speech detection task and target classification task sequentially.	2020	[23]

model trained on huge corpora provides a more general representation for knowledge transfer.

3.2.2 Feature representation

A wide range of features was also exploited in this particular task, ranging from straightforward n-gram representations to the most recent contextual language representations. Several text representation were used for the traditional machine learning model, including n-grams [24, 64, 75, 92, 94], TF-IDF [121], and word2vec [121]. Some studies also proposed to use linguistic features such as emoji information [27] and lexical [27, 92, 94, 147] features by using a specific lexicon. Most of the neural models in this task used word embedding as the text representation model. Several pre-trained models were exploited, such as FastText [27, 92, 94], GloVe [75, 134] and ELMo [115]. Finally, the transformer-based models use pre-trained models based on a very big corpus such as BERT [19, 48, 66, 79, 83, 90, 92, 134] and RoBERTa [48]. However, we also observe a study that proposes to re-train the BERT representation on a specific corpus related to abusive language [19]. Finally, the work by Nejadgholi and Kiritchenko [83] proposed to use unsupervised topic modeling approach to generate the features for obtaining better topic generalization on cross-dataset abusive language detection experiment.

Our study discover that several state-of-the-art pre-trained models provide the best feature representation and better generalization to deal with domain-shift in the cross-domain abusive language detection task. Interestingly, some studies proposed using external resources to facilitate the knowledge transfer between domains by delivering domain-independent features. These additional features were infused into either traditional models [147] or neural-based models [94] and succeeded in improving the prediction performance. Wiegand et al. [147] show the effectiveness of additional features from their novel abusive words lexicon in a cross-domain abusive language detection setting. The additional features were represented as a score based on the confidence learned by an SVM classifier. Similarly, Pamungkas et al. [92, 94] exploited the HurtLex lexicon, which contains a list of abusive words in 17 categories. The features were represented as a 17-column binary vector, to indicate the presence of each word category in the document. The vector was then concatenated to the representation of the message computed by LSTM network.

3.2.3 Domain transfer

The main challenge of cross-domain classification is the domain shift between training and testing data. Several methods have been proposed by studies in more mature areas, such as sentiment analysis [35, 95, 151]. These techniques are usually called domain-adaptation or domain-transfer, a specific approach to allow the model to learn domain-independent features, intersecting between two or more different domains. In the abusive language detection task, several features could represent an important signal for knowledge transfer between domains, such as the use of abusive words [147], emotional information [109, 119], and some other linguistic features [27, 66, 92, 94]

Table 2 shows that studies have different approaches to cope with the domain-shift problem. Some works proposed to combine the training sets from several different domains dataset [27, 48, 90, 92]. This straightforward approach allows the trained model to obtain wider domain coverage for detecting abusive language. Most aforementioned studies found that this simple approach was proven to be effective in this task. However, there is still a possibility that the trained model would struggle when applied to data from the totally unseen domain. Several other studies experimented with the use of lexicon as a domain-independent feature to bridge the domain-transfer. Wiegand et al. [147] used their novel lexicon automatically induced from HateBase, a platform that provides several keywords related to hate speech. Meanwhile, Pamungkas et al. [92, 94] and Corazza et al. [27] exploited HurtLex, a manually built lexicon by DeMauro [34], which contains offensive words structured in 17 different categories. Additional features from these lexica were also proven helpful to facilitate the transfer of knowledge between domains.

We also found some works that tried to modify the input sample for training the model in order to minimize the domain-shift issue between source and target domains.

For example, Nejadgholi and Kiritchenko [83] used the topic modeling approach and proposed to remove the domain-specific instances from the training set, resulting in the improvement of the model's performance. Another effort by Karan and Snajder [64] adopted a domain adaptation approach called FEDA (Frustratingly Easy Domain Adaptation), which works by duplicating features across domains to allow the model to learn domain-dependent weights for each feature. Finally, Mozafari et al. [79] proposed to deal with the racial bias on the abusive language dataset by re-weighting the input samples using the existing regularization approach. Their approach was shown to be effective in decreasing the dataset bias issue, which was found as one of the main problems in cross-domain classification.

We also notice that some studies focus more on providing better representation to improve the model's domain generalization. Wang et al. [141] proposed a multi-aspect embedding, which combines several representations, including target, content, and linguistic behavior, to provide domain-transfer knowledge. Then, Caselli et al. [19] proposed to retrain state-of-the-art BERT with a huge abusive language corpus to obtain a more specific representation for abusive language detection tasks.

Furthermore, we discover two studies proposed new architectures to tackle cross-domain abusive language detection task specifically. Rizoiu et al. [115] proposed a joint-learning model based on Bi-LSTM, which allows the model to learn from two datasets sequentially, obtaining better generalization. In addition, Waseem et al. [145] proposed a multitask learning architecture based on LSTM to learn the problem from two or more tasks sequentially, providing a medium for knowledge transfer between domains. The rest of the works more focused on investigating the transferability of some models, including BERT in the cross-domain abusive language detection [121, 134]. They found that using BERT only without a specific approach for bridging domain-shift already achieves a solid result.

4 Multilingual abusive language detection

Another prominent challenge in abusive language detection is the multilinguality issue. Even if in the last years abusive language datasets were developed for other languages, including Italian [15, 41], Spanish [41], and German [148], English remains by far the most represented language. Recently, deep learning approaches have been applied, achieving state-of-the-art results for some languages [9, 78]. However, most of the proposed models are tested in monolingual settings, mostly in English. Since the most popular social media such as Twitter and Facebook are highly multilingual, fostering their users to interact in their primary language, there is a considerable urgency to develop a robust approach for abusive language detection in a multilingual environment, also for guaranteeing a better compliance to governments demands for counteracting the phenomenon — see, e.g., the recently issued EU commission *Code of Conduct on countering illegal hate speech online* [37].

Similarly to other natural language processing tasks [62], detecting abusive language in less-resourced languages is a prominent and timely challenge. For example, the escalation of hate speech against Muslims in Rohingya Myanmar was also affected by the failure to stop spreading hate comments on Facebook due to the difficulty of processing Burmese text automatically [5]. The current availability of datasets in many languages [103], makes the time ripe for addressing the multilingual challenge. Cross-lingual transfer learning is the common approach to transfer knowledge from one language (usually with more available resources) to another language (usually with less resources) [69, 126]. In this approach, models are trained and optimized on a dataset from one language (called *source* language), and then tested on another language (called *target* language). Zero-shot learning is an extreme case of transfer learning, where a model trained on one language (such as in this work) or one domain is employed to predict samples from a totally unseen language or domain [51]. The less extreme form of transfer learning is few-shot learning, where a percentage of samples from unseen data (target language) is added to the training set, allowing the model to learn a better generalization between two languages or domains [126].

In this section, we discuss the development of studies in building robust models to detect abusive language across multiple languages. Specifically, we focus on the abusive language detection task in a cross-lingual setting. We review the available abusive language datasets in languages other than English, which could be exploited for this task. Importantly, we also deeply discuss several existing approaches that have been proposed in this task, mainly focusing on the method to transfer knowledge between languages.

4.1 What datasets are available for multilingual abusive language detection study?

In this section, we present information regarding the available datasets for abusive language tasks across different languages. Since we already presented the English datasets in the cross-domain part, in this section we only review the available datasets in languages other than English, which we

[5] https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/

will call *lower-resourced* languages for the rest of this article. We obtain this information based on the two most recent reviews [103, 136] which focused on the available resources in abusive language tasks. In addition, we also add more uncovered resources from the most recent shared tasks in the abusive language field, such as Misogyny@EVALITA2020 [40], HaSpeeDe@EVALITA2020 [122], and OffensEval@SemEval2020 [154]. We also search for the recently available resources from the last edition of Language Resources and Evaluation Conference (LREC) 2020[6] and Workshop on Online Abuse and Harms (WOAH) 2020[7], where we discover some datasets that are still not covered in these surveys. Table 3 summarizes the information of these lower-resourced languages datasets for abusive language detection task. We provide an in-depth discussion focusing on the comparison of these resources in the following.

Language In Table 3, we use the ISO 639-1 language code to represent the language names. We provide the list of languages with their corresponding code in Appendix Table 6. Based on Table 3, the abusive language datasets were already available in 18 different languages. Despite being not as many as in English, we notice that some languages have more resources than others, such as Arabic (AR), Hindi (HI), and Italian (IT). However, some other languages only have one resource available such as Czech (CS), Croatian (HR), Poland (PL), Swedish (SW), Turkish (TR), and Vietnamese (VI). The availability of these lower-resourced datasets indicates that this research direction is still growing. However, we notice that these resources are more centered on Indo-European languages. We still could not find datasets in the Niger-Congo language family which are mostly used in some African regions. The datasets in Afro-Asiatic, Austronesian, and other language families are also far less than Indo-European languages. Moreover, we observe Hindi-English (HI-EN) code-mixed datasets, all focusing on detecting hate speech. The first dataset of hate speech in Hindi-English code-mixed was proposed by Bohra et al. [14]. Mandl et al. presented a new collection created for a shared task, Hate Speech and Offensive Content Identification (HASOC), at FIRE 2019. Recently, Rani et al. [111] proposed the first Hindi-English hate speech dataset containing tweets written in both Roman and the native Devanagari script. Additionally, a Swahili-English code-mixed hate speech dataset was recently published [87]. They gathered their dataset from Twitter, mainly related to the 2017 general election in Kenya. It is worth mentioning the work by Oriola and Kotze [88] proposing a code-mixed Twitter dataset containing 14,896 tweets written in a mix of four different languages, namely English, Afrikaans, IsiZulu, and Sesotho.

Topical focus Similarly to the English datasets, these lower-resourced languages datasets also feature different topical focuses, where hate speech is the most used phenomenon to describe the resource. Other datasets cover several abusive phenomena such as offensiveness, abusiveness, misogyny, aggressiveness, and cyberbullying. The topical focus is also an important aspect to be considered in the cross-lingual abusive language detection task. A study found that topic bias was one of the main issues in cross-lingual abusive language detection [8]. If we do not want to deal with topic-shift between languages, we notice some datasets which only focus on one topic and cover more than one language, such as hate speech and misogyny. We also aware that there are a lot of datasets that have hate speech topics. However, different approaches in collecting the data could potentially introduce another bias issue when exploited in cross-lingual settings. As observed by Arango et al. [8], several biases such as user bias, racial bias, and sampling bias could be an issue in cross-lingual abusive language detection task. Otherwise, we can freely choose the available datasets if we want to tackle both domain-shift and language-shift.

Data source Most resources were retrieved from social media platforms such as Twitter, Facebook, and Instagram. Twitter is the most convenient platform which provides API and a more friendly policy to retrieve and distribute the samples gathered from its platforms. We can see from Table 3 that almost 60% of abusive language datasets were obtained from Twitter. Some other datasets were obtained from comments on news sites, online forums such as Reddit and Youtube comments. In a multilingual or cross-lingual setting, we also need to pay attention to the source of the data. Every source has its own specific characteristics, such as stylistic aspects and formality levels. Twitter data have some specific features, such as hashtags and user mentions. Language in social media platforms is usually used more informal language than other sources such as news site comments.

Availability Based on the manual check, most of the abusive language datasets in lower-resourced language were made publicly available. We only discover 4 out of 60 resources were not shared publicly by their authors. However, some authors decided to provide only the tweet identifier due to some Twitter policies and allowed us to retrieve the tweets by using the Twitter public API. The restricted datasets are mostly obtained from other sources than Twitter, which provides a more strict policy for sharing the data.

[6] https://lrec2020.lrec-conf.org/en/

[7] https://www.workshopononlineabuse.com/

Table 3 Summary of available abusive language datasets across different languages

Lang.	Topical focus	Sources	Entries	Available	Ref
AM	Hate speech	Facebook	4,882	No	[77]
AR	Hate speech	Twitter	6,000	Yes	[4]
	Hate speech	Multiple sources	6,039	Yes	[53]
	Offensiveness	Twitter	1,100	Yes	[80]
	Offensiveness	Youtube	15,050	Yes	[3]
	Hate speech	Twitter	5,846	Yes	[81]
	Hate speech	Twitter	3,353	Yes	[89]
	Offensiveness	Twitter	10,000	Yes	[154]
BN	Hate speech	Facebook	5,126	Yes	[60]
	Aggressiveness	Youtube	5,000	Yes	[68]
	Misogyny	Youtube	5,000	Yes	[68]
CS	Flamming	News sites	5,077	Yes	[133]
DA	Offensiveness	Twitter, Facebook, Reddit	3,600	Yes	[128]
DE	Hate speech	Twitter	541	Yes	[117]
	Flamming	News sites	5,077	Yes	[133]
	Hate Speech	Twitter	4,669	Yes	[72]
	Offensiveness	Facebook	5,836	Yes	[17]
	Offensiveness	Twitter	8,541	Yes	[148]
EL	Abusiveness	News sites	1.5 millions	Yes	[97]
	Offensiveness	Twitter	10,287	Yes	[101]
ES	Misogyny	Twitter	4,138	Yes	[41]
	Hate speech	Twitter	6,600	Yes	[11]
	Aggressiveness	Twitter	11,000	Yes	[7]
	Hate speech	Twitter	6,000	Yes	[99]
FR	Hate speech	Other	15,024	Yes	[25]
	Flamming	News sites	5,077	Yes	[133]
	Hate speech	Twitter	4,014	Yes	[89]
HI	Offensiveness	Twitter	3,679	No	[74]
	Aggressiveness	Facebook	15,000	Yes	[67]
	Aggressiveness	Youtube	5,000	Yes	[68]
	Misogyny	Youtube	5,000	Yes	[68]
HI-EN	Hate speech	Twitter	4,575	Yes	[14]
	Hate speech	Twitter	5,983	Yes	[72]
	Hate speech	Facebook, Twitter	3,367	Yes	[111]
HR	Abusiveness	News site	17 millions	Yes	[70]
ID	Hate speech	Twitter	1,100	Yes	[5]
	Abusiveness	Twitter	2,016	Yes	[57]
	Hate speech	Twitter	13,169	Yes	[58]
IT	Homophobic	Twitter	1,859	No	[2]
	Hate speech	Other	15,024	Yes	[25]
	Hate speech	Instagram	6,710	No	[27]
	Hate speech	Facebook	6,502	No	[139]
	Hate speech	Twitter	4,000	No	[102]
	Flamming	News sites	5,077	Yes	[133]
	Hate speech	Twitter	6,009	Yes	[123]
	Misogyny	Twitter	5,000	Yes	[39]
	Misogyny	Twitter	6,000	Yes	[40]
	Hate speech	Twitter, Facebook	4,000	Yes	[15]

Table 3 (continued)

Lang.	Topical Focus	Sources	Entries	Available	Ref
	Hate speech	Twitter, news site	8,602	Yes	[122]
	Cyberbulying	WhatsApp	14,600	Yes	[131]
PL	Cyberbullying	Twitter	11,041	Yes	[104]
PT	Offensiveness	Twitter	7,672	Yes	[82]
	Offensiveness	News site	1,250	Yes	[98]
	Hate speech	Twitter	3,059	Yes	[44]
SL	Abusiveness	News site	13,000	Yes	[42]
	Abusiveness	News site	7.6 millions	Yes	[70]
SV	Hate speech	Web fora	3,056	No	[38]
SW-EN	Hate speech	Twitter	25,000	No	[87]
TR	Hate speech	Twitter	36,232	Yes	[26]
	Offensiveness	Twitter	35,000	Yes	[154]
VI	Hate speech	Facebook	25,431	Yes	[140]

Annotation scheme Similar to the cross-domain setting, in the cross-lingual experiment, we also need to uniform the labels of every dataset. Most previous studies decided to binarize the label into two classes, namely abusive and not abusive. Based on our investigation, some datasets have more than two labels to capture a finer-grained phenomenon instead of merely limiting it to binary labels. Previous studies proposed to combine some labels when some of them can be safely merged into one class [64, 94]. For example as adopted by Karan and Snajder [64], they combined *overtly aggressive* and *covertly aggressive* labels as abusive class and *not aggressive* as not abusive class of the TRAC-1 datasets by Kumar et al. [67]. Otherwise, we can remove the data with a specific label when it is too problematic to merge some classes into one class. For example, the proposed dataset by Ousidhoum et al. [89] introduces some classes, including *hate speech*, *abusive*, *offensive*, *disrespectful*, *fearful*, and *normal*. In this case, we can combine *hate speech*, *abusive*, and *offensive* into one *abusive* class, but it is quite problematic to include the *disrespectful* and *fearful* label in the class, as proposed by Aluru et al. [6].

Data distribution In the cross-lingual setting of abusive language detection task, we also need to consider the data distribution of training (in source languages) and testing (in target languages) data. Based on our manual inspection, most of the resources have more positive (abusive) samples than negative (not abusive) ones. As mentioned in the cross-domain part, maintaining the same class distribution of training and testing data is important to have a more reliable evaluation and avoid bias in the models [92, 94]. Therefore, if the test set only contains 20% of abusive instances, a similar distribution can be imposed on the training set in the source language by adding or removing instances.

4.2 What has been done so far in multilingual abusive language detection study?

This section presents the existing studies focusing on building robust models to detect abusive language across different languages automatically. Overall, we use the same approach, as shown in Fig. 1, to collect related studies from several publication repositories. The only difference is the keywords used to query the relevant publications. For this purpose, we employ four keywords, namely "cross-lingual abusive language detection", "cross-lingual hate speech detection", "multilingual abusive language detection", and "multilingual hate speech detection". We use these keywords in two scientific publication repositories, namely Google Scholar and ACL Anthology. In the case of Google Scholar, we limit the query only to the first five pages of each keyword, without any limitation on publication time. We also check the cited documents and references for each document shown in the query result. Finally, we also remove some studies which did not provide any objective and insight to build a robust model to detect abusive instances across languages. For example, we notice some experiments with different models to cope with datasets in different languages.

Table 4 summarizes the existing works found on abusive language detection across different languages. We notice that the study in this direction is still relatively new, with the first study found in 2019. The works are more centered on languages from the Indo-European family, such as English, French, Spanish, Italian, German, and Hindi, in line with

Table 4 Summary of approaches adopted in existing studies on cross-lingual abusive language detection tasks

Models	Approach	Year	Ref
Traditional model	Proposed to use the *bleaching* approach [52] with a model based on SVM to conduct cross-lingual experiments between Italian and English	2018	[10]
Traditional model	Experimented with a gradient-boosting model and proposed to concatenate two sentence embeddings obtained from LASER Embedding and Multilingual BERT as a language-agnostic representation.	2019	[120]
Traditional models	Experimented with the use of machine translation tools to translate the training data to the target language and exploited a wide range of traditional models including SVM, naïve Bayes, and random forest.	2019	[59]
Neural based	Proposed a joint-learning architecture based on LSTM coupled with features from HurtLex to transfer knowledge between domains and languages.	2019	[94]
Transformer based	Proposed multichannel architecture based on BERT model, which learns the task sequentially in three languages: source languages, English, and Chinese.	2019	[130]
Neural based	Proposed multitask architecture based on Sluice Networks coupled with Babylon cross-lingual embedding.	2019	[89]
Transformer based	Proposed to continue training multilingual BERT and XLM-RoBERTa via masked language modeling (intermediate MLM-ing) as a language and domain adaptation approach.	2020	[48]
Transformer based	Proposed to use XLM-RoBERTa by inter-language and inter-task language transfer learning for conducting cross-lingual classification of offensive languages.	2020	[110]
Transformer based	Employed multilingual BERT and proposed two data augmentation techniques for the cross-lingual transfer by adding training set with filtered samples from the semi-supervised dataset and samples from languages other than target languages.	2020	[1]
Transformer based	Proposed a hybrid emoji-based masked language model (MLM) on the top of XLM architecture to leverage the common information conveyed by emojis as a language-agnostic feature.	2020	[28]
Multiple models	Proposed to infuse features from multilingual hate lexicon called HurtLex into traditional (SVM) and neural models (LSTM) for transferring knowledge across different languages.	2020	[92]
Transformer based	Exploited cross-lingual representation based on XLM-RoBERTa for building multilingual models and tested on five different languages.	2020	[30]
Transformer based	Proposed a novel architecture consisting of a frozen Transformer Language Model (TLM) and Attention-Maximum-Average Pooling (AXEL) to deal with zero-shot and few-shot cross-lingual learning.	2020	[132]
Transformer based	Proposed a multichannel BERT architecture that learns the task from both source and target languages.	2020	[20]
Transformer based	Proposed to convert Hindi-English code-switched data into the high resource languages (English) for exploiting both monolingual and cross-lingual settings by using the state-of-the-art cross-lingual language model XLM-RoBERTa.	2020	[31]
Multiple models	Conducted an exploratory work using logistic regression, several deep learning models (CNN-GRU, and BERT based) and multilingual language representations (LASER, MUSE, and multilingual BERT) to deal with multilingual hate speech detection in nine languages.	2020	[6]
Neural based	Conducted an extensive experiment to build a language-agnostic model based on recurrent neural networks (RNN) by exploiting several language-agnostic features.	2020	[29]
Transformer based	Proposed a single multilingual hate speech model based on the multilingual BERT model, which is trained on datasets in five different languages.	2020	[100]

Table 4 (continued)

Models	Approach	Year	Ref
Multiple models	Experimented with several models including traditional models (logistic regression), neural models (CNN-LSTM), and transformer models (BERT) to build a multilingual system trained on code-switched datasets in English and Hindi by adopting a transfer learning approach.	2021	[135]
Multiple models	Experimented with several models including a joint-learning architecture which allow the model to learn from source and target languages sequentially.	2021	[93]

the available resources. Most of them tried to transfer the knowledge from a resource-rich language (English) to other languages with the lower resource available. All studies proposed a supervised approach, where most of them utilized a multilingual language representation as a basis for knowledge transfer between languages. Following, we discuss the gathered studies in this direction, focusing on several aspects, including the model adopted, features used, and approaches proposed to deal with language-shift.

4.2.1 Models

Based on Table 4, most studies implemented transformer-based architecture to deal with abusive language detection in a cross-lingual setting. However, we also observe some works that exploited a traditional machine learning approach, such as logistic regression [6, 10, 135], linear support vector machines [92, 94], and support vector machines [59]. They used multilingual language representation or simple translation tools (to translate the data training to the target languages) for the knowledge sharing between languages. Some studies also exploited several neural-based models such as LSTM [29, 92, 94, 135], Bi-LSTMs [29], and GRU [6, 28]. The more recent works adopted several transformer-based architectures due to the availability of multilingual transformer models such as Multilingual BERT [1, 6, 48, 92, 100, 132, 135], RoBERTa [30, 31], XLM [28, 132], and XLM-RoBERTa [30, 31, 48, 110]. Interestingly, we also notice some works that proposed a multichannel architecture based on the multilingual BERT model [20, 130], which allows the model to learn the task in several languages sequentially. Finally, we also discover a study proposed to adapt a multitask approach to deal with this task [89].

Based on our investigation, transformer-based models with multilingual language representations effectively deal with language-shift in the zero-shot cross-lingual abusive language detection task. A recent study shows that XLM-RoBERTa provided a more robust performance than other multilingual language models, including multilingual BERT and RoBERTa [30, 48, 110]. However, the most recent study shows that the use of a straightforward English BERT pre-trained model with the help of translation tools already achieved a competitive result. The more complex approaches that adopt joint-learning [94], multi-channel [130], or multi-task [89] architectures obtained more competitive results compared to previously mentioned models.

4.2.2 Feature representation

For the traditional models, some works used the LASER Embedding model, which provides a language-agnostic representation across 93 languages. A study by Basile et al. [10] proposed to use TF-IDF representation of bleached characters n-grams. Other studies simply translated the training data to the target language and used the word n-grams feature representation [6, 59, 92, 94]. Meanwhile, most neural-based models were coupled by multilingual word embedding models, including Facebook MUSE (Multilingual FastText) [6, 89, 92, 94] and Babylon Embeddings [89]. Finally, the transformer-based architectures exploited the multilingual pre-trained model trained on the very big corpus such as Multilingual BERT [1, 6, 48, 92, 100, 132, 135], RoBERTa [30, 31], ULMFit [31], and the recent XLM-RoBERTa [30, 31, 48, 110]. It is worth noting that we also discover that some features were introduced to complement the language representation, providing language-agnostic information for knowledge transfer such as a hate-specific lexicon (HurtLex) [29, 92, 94] and emotion features based on emoji presence [28].

Overall, almost all cross-lingual abusive language detection studies exploited multilingual language models as the main feature representation. In particular, the most recent studies found that a multilingual representation based on XLM-RoBERTa obtained the most robust result and outperformed other multilingual language models [30, 48, 110]. Several studies also presented the interesting finding that infusing language-agnostic features extracted from hate-specific lexicons HurtLex, in particular) [29, 94] and

emoji-based features [28] could improve abusive language detection systems in a multilingual setting. In the case of HurtLex, the feature was represented as a one-hot vector which indicates the word presence in 17 HurtLex categories [94]. Meanwhile, Corazza etS al. [28] exploited common information conveyed by emoji for building a pre-trained Masked Language Model (MLM).

4.2.3 Language transfer approaches

Cross-lingual transfer learning is the common approach to transfer knowledge from one language to another language [69, 126]. In this approach, models are trained and optimized on a dataset from one language (called *source* language), and then tested on another language (called *target* language). In this task, a specific model or approaches is needed to facilitate the knowledge transfer between language. In this subsection, we discuss several approaches proposed by existing works to bridge the language-shift in cross-lingual abusive language detection task.

Several works proposed the most straightforward approach by **utilizing machine translation tools to align data training and testing language**. Most of them used Google Translate, which provides reliable translation results. Pamungkas et al. [92, 94] exploited Linear Support Vector Classifier with TF-IDF feature representation of translated data by Google Translated. Some other works also tried to align the language of test data to the source language before feeding them to state-of-the-art English BERT pre-trained models [6, 92]. The translation tools were also used to obtain parallel corpora in some studies which propose a joint learning or multichannel architecture. These architectures require these corpora to allow the model to learn the task in two or more languages sequentially [20, 92, 94, 130].

Some existing studies proposed to experiment by **infusing language-agnostic features as language-independent information for transferring knowledge between languages**. Pamungkas et al. [92, 94] and Corazza et al. [29] used features extracted from HurtLex [12], a multilingual lexicon that specifically contains abusive words. Another work by Corazza et al. [28] exploited a language-agnostic feature provided by emoji in the Twitter data. They argued that emoji could give some signals related to emotion information.

A **novel architecture was also proposed** by several works to obtain a better learning representation across different languages. Glavas et al. [48] proposed to continue the training process of Multilingual BERT and XLM-RoBERTa models via masked language modeling. Pamungkas et al. [92, 94] presented a joint-learning architecture model to learn the task in source and target languages sequentially. Then, Casula et al. [20] and Sohn et al. [130] introduced a similar architecture by introducing a multichannel model based on multilingual pre-trained models. Then, Stappen et al. [132] introduced novel architecture consisting of a frozen Transformer Language Model (TLM) and Attention-Maximum-Average Pooling (AXEL) to deal with the zero-shot cross-lingual classification. Finally, Ousidhoum et al. [89] proposed a multitask architecture based on Sluice Network [118] coupled with Babylon cross-lingual word embedding [129], which allows the model to share the same parameters from other related tasks.

The cross-lingual task heavily relied on the machine translation tools for a long time before the emergence of multilingual language representation in recent years. Some prior studies **conducted an exploratory experiment to test the robustness of these multilingual language representation models** in abusive language detection tasks, without any specific knowledge transfer approaches between languages. Pamungkas et al. [92, 94] and Aluru et al. [6] used a straightforward logistic regression model coupled with Multilingual LASER Embedding. In addition, they also experimented with the Multilingual FastText embedding. Then, several other works [6, 92, 94, 100] also tested the robustness of the Multilingual BERT model to tackle cross-lingual abusive language detection. Meanwhile, Ranasinghe et al. [110] and Dadu and Pant [30, 31] experimented with the recent state-of-the-art multilingual language representation XLM-RoBERTa to deal with this task. Finally, we observe work that proposed two data augmentation techniques for cross-lingual transfer by adding a training set with filtered other data samples and using an ensemble model based on the Multilingual BERT pre-trained model [1].

5 Challenge and opportunities

The analysis of the relevant literature done so far gives us a picture of cross-domain and cross-lingual abusive language detection as challenging tasks. Several challenges emerged, summarized as follows:

– *Bias issue on the existing datasets.* Several studies mentioned that dataset bias is one of the main issue which contributes to the difficulties of abusive language detection in both cross-domain and cross-lingual settings. Several kinds of bias were found, including topic bias [146], author bias [146], and

racial bias [32]. Among these biases, topic bias is the most influential issue, as noticed also by some works in cross-domain [94] and cross-lingual [8] abusive language detection task.

- *The insufficient ability of current multilingual language models to transfer knowledge between languages in the specific hate speech detection task.* Especially, in the use of some swear words which are very culture-dependant and vary from a language to another. Similar issues were also observed by Pamungkas et al. [91, 94], where swear words have an important role in a cross-lingual setting of abusive language detection in which some of them are not directly translatable by using machine translation tools.
- *Language- and topic-shift.* Language-shift is not the only issue to deal with in a cross-language setting, but also the topic-shift between one dataset and others [48]. This due to the differences in task formulation and the nature of the abusive language datasets. This issue is related to the first challenge mentioned in this list, which is also in line with the findings of a recent study in cross-lingual hate speech detection [8].
- *Unstable performance of models in different target languages* [48]. Existing works show that the performance in more resource-rich languages is higher than in lower-resource languages. This may be related to multilingual language representation models being trained on different amounts of data in different languages [149].
- *Difficulties in producing a dataset that encompasses multiple facets and targets of abusive language online* [48, 134]. The effort to merge several datasets with different topical focuses still does not obtain a significant result [92]. Actually, this issue is not only an issue for this research area, but rather for every task in which manually annotating a new dataset is very labor intensive and a highly subjective task.
- *Intrinsic complexity in defining abusive phenomena and variety of definitions.* Different concepts and terms were introduced across studies for similar abusive phenomena [137]. This issue contributes to the difficulties in providing a better experimental setting for the cross-domain abusive language detection task.

Based on these challenges, we also point out several opportunities for future studies in this research direction, which are summarized below.

- *Applying debiasing approaches on the available dataset.* Several studies have explored this direction by adapting debiasing techniques from other research topics to reduce the bias issue in abusive language datasets [96, 112]. They proved that reducing or removing bias on either the language model or the datasets could improve the model performance in detecting abusive language automatically.
- Having an *abusive dataset covering multiple facets of abusive behavior* could be a first important step important to develop robust systems which are stable and well-performing across different abusive domains [134]. However, this is not a trivial task, since obtaining broad samples of abusive instances of real online discourse is very difficult.
- *Developing a pre-trained word embedding model, specifically for abusive language detection tasks*, also in a multilingual setting [134]. Often text in abusive utterances has specific characteristics compared to traditional text, which involve either explicit mention of abusive words, obfuscated words and implicit abuse, i.e., indicating negative stereotypes. Several studies have been proposed to deal with this solution, which we think need to be followed up [13]. This solution may could help to cope with cross-domain and cross-lingual task difficulties.
- Several previous studies showed the effectiveness of the *infused features from a domain or language-independent resource* both in cross-domain and cross-lingual settings. The further development of exploiting other resources could also help the model to transfer knowledge between domains and languages. For example, some studies highlighted the importance of *emotion information* in abusive language detection task [109, 119]. Therefore, exploring the use of emotion information as domain- or language-independent features for knowledge transfer would be valuable. Another study by Pamungkas et al. [93] also proved the usefulness of external features extracted from HurtLex, a multilingual lexicon that contains offensive words structured in 17 different categories. HurtLex contains a wide range of hateful words, organized in general categories sometimes related to cultural stereotypes, ranging from ethnic slurs to insulting words that target physical disabilities and derogatory senses in different languages. Specifically, they found that HurtLex can help the knowledge transfer of abusive language detection across languages, which often make use of rhetorical figures (e.g., metaphors, synecdoche, metonymy) and idiomatic expressions, and they are highly sensitive to geographical, temporal, and cultural variations, especially when the derogatory meaning is linked to stereotype and prejudice.
- *Tackling the lesser performance of multilingual language representations in low-resource languages also needs to be considered.* Even a study by Wu and Dredze [149] found that 30% of languages in the multilingual BERT model with lower pretraining resources obtain

worse performance than without using a pre-trained language model. Several possible solutions could be considered. One of the main answers is by developing monolingual embedding with sufficient training data. Since, Wu and Dredze [149] also found that monolingual language always obtains better performance than multilingual BERT when sufficient data available to develop the pre-trained model. Another possible solution is to extend the current multilingual model to improve its language coverage as proposed by Wang et al. [142].

- Focusing on the *model and architecture engineering* to facilitate the language and domain transfer is another possible, solid solution for such endeavors. Existing studies show that some techniques such as joint-learning, multitask learning, and MLM-ing effectively alleviate the models' performance. For future work, it could be interesting to implement other transfer learning approaches, by exploiting deep learning techniques, both traditional deep learning and adversarial deep learning [156].
- On the theoretical counterpart, a careful study of the notion of every abusive behavior online which is modeled with the purpose of automatic detection is important, to obtain a *clearer terminology and understanding of the abusive phenomena we want to capture in language*. The study by Vidgen et al. [137] proposed several possible solutions to address this issue, which can be considered for future works.

6 Conclusions

This survey provides a comprehensive overview of existing studies in developing robust models to detect abusive language across different domains and languages. First, we present the available datasets that could be exploited in this research direction, covering multiple platforms, abusive phenomena, and languages. We also review the approaches that have been proposed in this field, focusing on analyzing the specific methods to transfer knowledge between domains and languages. Finally, we also present the current challenges and opportunities related to this focus based on the existing works, providing further research development insights.

This study observe that most of the available abusive language datasets are gathered from social media platforms such as Twitter, Facebook, and Instagram. Twitter is the most exploited source of data, which may be due to the convenience of retrieving the samples using the Twitter public API and of the policies for sharing the data. We also notice that hate speech is the most studied phenomenon, compared to other abusive phenomena such as toxicity, offensiveness, and cyberbullying. A wide variety of methods have been proposed to deal with the cross-domain study of abusive language detection task. However, the most recent transformer-based architecture succeeds in obtaining the most promising result. Several studies also proposed specific approaches to coping with the domain shift in the cross-domain setting, such as merging datasets from different domains, modifying the input sample to minimize domain-shift, and proposing novel architectures to facilitate domain transfer, and using external resources as a domain-independent feature.

In the cross-lingual settings, we focus on non-English resources and observe that abusive language datasets are already available in 18 languages, but they are more centered on the Indo-European languages family. There are several underrepresented or even unavailable yet resources in some language families, including Afro-Asiatic, Austronesian, and Niger-Congo. Most datasets in languages other than English were also retrieved from the Twitter platform. Most studies in this direction focus on transferring knowledge from a resource-rich language to other lower resource languages. Like in cross-domain studies, most works in cross-lingual settings also exploit transformer-based architectures and use the available multilingual language representation models. Other studies also proposed several specific approaches to share information between languages, including machine translation to align training and testing data, infusing language-agnostic features as language-independent information, and offering novel architectures to facilitate knowledge transfer between languages.

Finally, we identify some recent challenges and opportunities in this research direction. Dataset bias is one of the main issues contributing to the difficulties of cross-domain and cross-lingual settings of abusive language detection tasks. This issue is an open problem, whereas the challenge is to develop novel resources that are less biased and cover different facets of abusive phenomena online. On the theoretical side, different concepts and terms were used across studies to describe similar abusive phenomena, which is problematic in the context of the cross-domain setting of abusive language detection task. Further exploration of every abusive phenomenon notion is also vital to obtain more precise terminology in the abusive language field. Overall, analysis of the relevant literature done in this study gives us a picture of cross-domain and cross-lingual abusive language detection as challenging tasks. Despite this research field still being in the early phase of development, the existing studies confirm the urgency of tackling this task and its further development opportunities.

Appendix

Table 5 List of URLs of available datasets

No.	Lang.	URL	Ref
1.	EN	https://github.com/t-davidson/hate-speech-and-offensive-language	[33]
2.	EN	https://github.com/mayelsherif/hate_speech_icwsm18	[36]
3.	EN	https://github.com/Vicomtech/hate-speech-dataset	[47]
4.	EN	https://github.com/jing-qian/A-Benchmark-Dataset-for-Learning-to-Intervene-in-Online-Hate-Speech	[105]
5.	EN	https://github.com/zeerakw/hatespeech	[144]
6.	EN	Email to valerio.basile@unito.it	[11]
7.	EN	https://github.com/marcoguerini/CONAN	[25]
8.	EN	https://github.com/manoelhortaribeiro/HatefulUsersTwitter	[114]
9.	EN	https://github.com/HKUST-KnowComp/MLMA_hate_speech	[89]
10.	EN	https://gombru.github.io/2019/10/09/MMHS/	[50]
11.	EN	https://hasocfire.github.io/hasoc/2019/dataset.html	[72]
12.	EN	https://figshare.com/articles/dataset/Wikipedia_Talk_Corpus/4264973	[150]
13.	EN	https://github.com/tapilab/icwsm-2020-toxic	[108]
14.	EN	https://www.kaggle.com/swetaagrawal/formspring-data-for-cyberbullying-detection	[116]
15.	EN	https://sites.google.com/site/offensevalsharedtask/olid	[152]
16.	EN	https://sites.google.com/site/offensevalsharedtask/solid	[154]
17.	EN	https://dataverse.mpi-sws.org/dataset.xhtml?persistentId=doi:10.5072/FK2/ZDTEMN	[45]
18.	EN	http://nlp.kiv.zcu.cz/projects/flame	[133]
19.	EN	https://github.com/Mrezvan94/Harassment-Corpus	[113]
20.	EN	Email to jgolbeck@umd.edu	[49]
21.	EN	https://amiibereval2018.wordpress.com/important-dates/data/	[41]
22.	EN	https://amievalita2018.wordpress.com/data/	[39]
23.	EN	https://amievalita2020.github.io/data/	[40]
24.	EN	https://github.com/AkshitaJha/NLP_CSS_2017	[61]
25.	EN	https://sites.google.com/view/trac1/shared-task	[67]
26.	EN	https://sites.google.com/view/trac2/shared-task	[68]
27.	AR	https://github.com/nuhaalbadi/Arabic_hatespeech	[4]
28.	AR	https://github.com/Hala-Mulki/T-HSAB-A-Tunisian-Hate-Speech-and-Abusive-Dataset	[53]
29.	AR	https://alt.qcri.org/~hmubarak/offensive/AJCommentsClassification-CF.xlsx	[80]
30.	AR	https://goo.gl/27EVbU	[3]
31.	AR	https://github.com/Hala-Mulki/L-HSAB-First-Arabic-Levantine-HateSpeech-Dataset	[81]
32.	AR	https://github.com/HKUST-KnowComp/MLMA_hate_speech	[89]
33.	AR	https://sites.google.com/site/offensevalsharedtask/multilingual	[154]
34.	BN	https://github.com/IshmamAlvi/Hate-Speech-for-Bengali-language	[60]
35.	BN	https://sites.google.com/view/trac2/shared-task	[68]
36.	CS	http://nlp.kiv.zcu.cz/projects/flame	[133]
37.	DA	https://figshare.com/articles/dataset/Danish_Hate_Speech_Abusive_Language_data/12220805	[128]
38.	DE	https://github.com/UCSM-DUE/IWG_hatespeech_public	[117]
39.	DE	http://nlp.kiv.zcu.cz/projects/flame	[133]
40.	DE	https://hasocfire.github.io/hasoc/2019/dataset.html	[72]
41.	DE	http://www.ub-web.de/research/	[16]
42.	DE	https://github.com/uds-lsv/GermEval-2018-Data	[148]
43.	EL	http://nlp.cs.aueb.gr/software.html	[97]
44.	EL	https://zpitenis.com/resources/ogtd/	[101]
45.	ES	https://amiibereval2018.wordpress.com/important-dates/data/	[41]
46.	ES	Email to valerio.basile@unito.it	[11]

Table 5 (continued)

No.	Lang.	URL	Ref
47.	ES	https://sites.google.com/view/mex-a3t/data-and-evaluation?authuser=0	[7]
48.	ES	https://zenodo.org/record/2592149#.YHfqa-gzY2w	[99]
49.	FR	https://github.com/marcoguerini/CONAN	[25]
50.	FR	http://nlp.kiv.zcu.cz/projects/flame	[133]
51.	FR	https://github.com/HKUST-KnowComp/MLMA_hate_speech	[89]
52.	FR	https://github.com/HKUST-KnowComp/MLMA_hate_speech	[89]
53.	HI	https://sites.google.com/view/trac1/shared-task	[67]
54.	HI	https://sites.google.com/view/trac2/shared-task	[68]
55.	HI-EN	https://github.com/deepanshu1995/HateSpeech-Hindi-English-Code-Mixed-Social-Media-Text	[14]
56.	HI-EN	https://hasocfire.github.io/hasoc/2019/dataset.html	[72]
57.	HR	https://www.clarin.si/repository/xmlui/handle/11356/1201	[70]
58.	ID	https://github.com/ialfina/id-hatespeech-detection	[5]
59.	ID	https://github.com/okkyibrohim/id-abusive-language-detection	[57]
60.	ID	https://github.com/okkyibrohim/id-multi-label-hate-speech-and-abusive-language-detection	[58]
61.	IT	https://github.com/marcoguerini/CONAN	[25]
62.	IT	http://nlp.kiv.zcu.cz/projects/flame	[133]
63.	IT	https://github.com/msang/hate-speech-corpus	[123]
64.	IT	https://github.com/msang/hate-speech-corpus	[123]
65.	IT	https://amievalita2018.wordpress.com/data/	[39]
66.	IT	https://amievalita2020.github.io/data/	[40]
67.	IT	http://www.di.unito.it/~tutreeb/haspeede-evalita18/data.html	[15]
68.	IT	https://github.com/msang/haspeede/tree/master/2020	[122]
69.	IT	https://dhsite.fbk.eu/2018/09/whatsapp-dataset-on-cyberbullying/	[131]
70.	PL	https://github.com/ptaszynski/cyberbullying-Polish	[104]
71.	PT	https://github.com/LaCAfe/Dataset-Hatespeech	[82]
72.	PT	http://www.inf.ufrgs.br/~rppelle/hatedetector/	[98]
73.	PT	https://github.com/paulafortuna/Portuguese-Hate-Speech-Dataset	[44]
74.	SL	https://www.spletno-oko.si/english	[42]
75.	SL	https://www.clarin.si/repository/xmlui/handle/11356/1201	[70]
76.	TR	https://coltekin.github.io/offensive-turkish/	[26]
77.	TR	https://sites.google.com/site/offensevalsharedtask/multilingual	[154]
78.	VI	https://github.com/vietnlp/vlsp2019_hatespeech_task/	[140]

Table 6 List of languages and the respective codes based on ISO 639-1

No.	Lang.	URL
1.	AM	Amharic
2.	AR	Arabic
3.	BN	Bengali
4.	CS	Czech
5.	DA	Danish
6.	DE	German
7.	EL	Greek
8.	ES	Spanish
9.	FR	French
10.	HI	Hindi
11.	HR	Croatian
12.	ID	Indonesian
13.	IT	Italian
14.	PL	Polish
15.	PT	Portuguese
16.	SL	Slovenian
17.	SV	Swedish
18.	SW	Swahili
19.	TR	Turkish
20.	VI	Vietnamese

Funding Open access funding provided by Università degli Studi di Torino within the CRUI-CARE Agreement. This work is partially funded by the project "Be Positive!" (under the 2019 "Google.org Impact Challenge on Safety" call).

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Ahn H, Sun J, Park CY, Seo J (2020) NLPDove at SemEval-2020 task 12: Improving offensive language detection with cross-lingual transfer. In: Proceedings of the fourteenth workshop on semantic evaluation. International Committee for Computational Linguistics, Barcelona, pp 1576–1586. https://www.aclweb.org/anthology/2020.semeval-1.206
2. Akhtar S, Basile V, Patti V (2019) A new measure of polarization in the annotation of hate speech. In: Alviano M, Greco G, Scarcello F (eds) AI*IA 2019 - advances in artificial intelligence - XVIIIth international conference of the italian association for artificial intelligence, Rende, Italy, November 19-22, 2019, Proceedings, Lecture Notes in Computer Science, vol 11946. Springer, pp 588–603. https://doi.org/10.1007/978-3-030-35166-3_41
3. Alakrot A, Murray L, Nikolov NS (2018) Dataset construction for the detection of anti-social behaviour in online communication in arabic. In: Shaalan K, El-Beltagy SR (eds) Fourth international conference on arabic computational linguistics, ACLING 2018, November 17-19, 2018, Dubai, United Arab Emirates, Procedia Computer Science, vol 142. Elsevier, pp 174–181. https://doi.org/10.1016/j.procs.2018.10.473
4. Albadi N, Kurdi M, Mishra S (2018) Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In: Brandes U, Reddy C, Tagarelli A (eds) IEEE/ACM 2018 international conference on advances in social networks analysis and mining, ASONAM 2018, Barcelona, Spain, August 28-31, 2018. IEEE Computer Society, pp 69–76. https://doi.org/10.1109/ASONAM.2018.8508247
5. Alfina I, Mulia R, Fanany MI, Ekanata Y (2017) Hate speech detection in the Indonesian language: A dataset and preliminary study. In: 2017 International conference on advanced computer science and information systems (ICACSIS). IEEE, pp 233–238
6. Aluru SS, Mathew B, Saha P, Mukherjee A (2020) Deep learning models for multilingual hate speech detection. arXiv:2004.06465
7. Álvarez-Carmona MÁ, Guzmán-Falcón E, Montes-y Gómez M, Escalante HJ, Villasenor-Pineda L, Reyes-Meza V, Rico-Sulayes A (2018) Overview of mex-a3t at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. In: Notebook papers of 3rd SEPLN workshop on evaluation of human language technologies for iberian languages (IBEREVAL), Seville, Spain, vol 6, p 23
8. 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). Inf Syst 101584
9. Badjatiya P, Gupta S, Gupta M, Varma V (2017) Deep learning for hate speech detection in tweets. In: Barrett R, Cummings R, Agichtein E, Gabrilovich E (eds) Proceedings of the 26th international conference on world wide web companion, Perth, Australia, April 3-7, 2017. ACM, pp 759–760. https://doi.org/10.1145/3041021.3054223
10. Basile A, Rubagotti C (2018) Crotonemilano for AMI at evalita2018. A performant, cross-lingual misogyny detection system. In: Caselli T, Novielli N, Patti V, Rosso P (eds) 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 Workshop Proceedings. CEUR-WS.org, vol 2263, pp 1–5 http://ceur-ws.org/Vol-2263/paper034.pdf
11. Basile V, Bosco C, Fersini E, Nozza D, Patti V, Rangel Pardo FM, Rosso P, Sanguinetti M (2019) SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In: Proceedings of the 13th international workshop on semantic evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA. pp 54–63. https://doi.org/10.18653/v1/S19-2007. https://www.aclweb.org/anthology/S19-2007
12. Bassignana E, Basile V, Patti V (2018) Hurtlex: A multilingual lexicon of words to hurt. In: Cabrio E, Mazzei A, Tamburini

F (eds) Proceedings of the fifth italian conference on computational linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018, CEUR Workshop Proceedings. CEUR-WS.org, vol 2253, pp 1–6. http://ceur-ws.org/Vol-2253/paper49.pdf

13. Bodapati S, Gella S, Bhattacharjee K, Al-Onaizan Y (2019) Neural word decomposition models for abusive language detection. In: Proceedings of the third workshop on abusive language online. Association for Computational Linguistics, Florence, pp 135–145. https://doi.org/10.18653/v1/W19-3515 https://www.aclweb.org/anthology/W19-3515
14. Bohra A, Vijay D, Singh V, Akhtar SS, Shrivastava M (2018) A dataset of Hindi-English code-mixed social media text for hate speech detection. In: Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media. Association for Computational Linguistics, New Orleans, pp 36–41. https://doi.org/10.18653/v1/W18-1105. https://www.aclweb.org/anthology/W18-1105
15. Bosco C, Dell'Orletta F, Poletto F, Sanguinetti M, Tesconi M (2018) Overview of the EVALITA 2018 hate speech detection task. In: Caselli T, Novielli N, Patti V, Rosso P (eds) 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 Workshop Proceedings. CEUR-WS.org, vol 2263, pp 1–9. http://ceur-ws.org/Vol-2263/paper010.pdf
16. Bretschneider U, Peters R (2016) Detecting cyberbullying in online communities. In: 24th European conference on information systems, ECIS 2016, Istanbul, Turkey, June 12-15, 2016, p. Research Paper 61 . http://aisel.aisnet.org/ecis2016_rp/61
17. Bretschneider U, Peters R (2017) Detecting offensive statements towards foreigners in social media. In: Bui T (ed) 50th Hawaii international conference on system sciences, HICSS 2017, Hilton Waikoloa Village, Hawaii, USA, January 4-7, 2017. ScholarSpace / AIS Electronic Library (AISeL) pp 1–10. http://hdl.handle.net/10125/41423
18. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D (2020) Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, NeurIPS 2020, December 6-12, 2020, virtual. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html
19. Caselli T, Basile V, Mitrovic J, Granitzer M (2020) Hatebert: Retraining BERT for abusive language detection in english. https://arxiv.org/abs/2010.12472
20. Casula C (2020) Transfer learning for multilingual offensive language detection with bert
21. Chatzakou D, Kourtellis N, Blackburn J, De Cristofaro E, Stringhini G, Vakali A (2017) Mean birds: Detecting aggression and bullying on twitter. In: Proceedings of the 2017 ACM on web science conference, WebSci '17, pp 13-22, Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3091478.3091487
22. Chen H, McKeever S, Delany SJ (2017) Presenting a labelled dataset for real-time detection of abusive user posts. In: Sheth AP, Ngonga A, Wang Y, Chang E, Slezak D, Franczyk B, Alt R, Tao X, Unland R (eds) Proceedings of the international conference on web intelligence, Leipzig, Germany, August 23-26, 2017. ACM, pp 884–890. https://doi.org/10.1145/3106426.3106456
23. Chiril P, Pamungkas EW, Benamara F, Moriceau V, Patti V (2021) Emotionally informed hate speech detection: a multi-target perspective. Cogn Comput 1–31
24. Chowdhury SA, Mubarak H, Abdelali A, Jung SG, Jansen BJ, Salminen J (2020) A multi-platform Arabic news comment dataset for offensive language detection. In: Proceedings of the 12th language resources and evaluation conference. European Language Resources Association, Marseille, France, pp 6203–6212 https://www.aclweb.org/anthology/2020.lrec-1.761
25. Chung YL, Kuzmenko E, Tekiroglu SS, Guerini M (2019) CONAN - counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In: Proceedings of the 57th annual meeting of the association for computational linguistics. Association for Computational Linguistics, pp 2819–2829
26. Çöltekin Ç (2020) A corpus of turkish offensive language on social media. In: Calzolari N, Béchet F, Blache P, Choukri K, Cieri C, Declerck T, Goggi S, Isahara H, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the 12th language resources and evaluation conference, LREC 2020, Marseille, France, May 11-16, 2020, European Language Resources Association. pp 6174–6184. https://www.aclweb.org/anthology/2020.lrec-1.758/
27. Corazza M, Menini S, Cabrio E, Tonelli S, Villata S (2019) Cross-platform evaluation for italian hate speech detection. In: Bernardi R, Navigli R, Semeraro G (eds) Proceedings of the Sixth italian conference on computational linguistics, Bari, Italy, November 13-15, 2019, CEUR Workshop Proceedings, vol 2481. CEUR-WS.org. http://ceur-ws.org/Vol-2481/paper22.pdf
28. Corazza M, Menini S, Cabrio E, Tonelli S, Villata S (2020) Hybrid emoji-based masked language models for zero-shot abusive language detection. In: Findings of the association for computational linguistics: EMNLP 2020, Association for Computational Linguistics, Online, pp 943–949. https://doi.org/10.18653/v1/2020.findings-emnlp.84. https://www.aclweb.org/anthology/2020.findings-emnlp.84
29. Corazza M, Menini S, Cabrio E, Tonelli S, Villata S (2020) A multilingual evaluation for online hate speech detection. ACM Trans Internet Techn 20(2):10:1–10:22. https://doi.org/10.1145/3377323
30. Dadu T, Pant K (2020) Team rouges at SemEval-2020 task 12: Cross-lingual inductive transfer to detect offensive language. In: Proceedings of the fourteenth workshop on semantic evaluation. International Committee for Computational Linguistics, Barcelona, pp 2183–2189. https://www.aclweb.org/anthology/2020.semeval-1.290
31. Dadu T, Pant K (2020) Towards code-switched classification exploiting constituent language resources. https://arxiv.org/abs/2011.01913
32. Davidson T, Bhattacharya D, Weber I (2019) Racial bias in hate speech and abusive language detection datasets. In: Proceedings of the third workshop on abusive language online. Association for Computational Linguistics, Florence, pp 25–35. https://doi.org/10.18653/v1/W19-3504. https://www.aclweb.org/anthology/W19-3504
33. Davidson T, Warmsley D, Macy MW, 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. AAAI Press, pp 512–515. https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665

34. De Mauro T (2016) Le parole per ferire. Internazionale, 27 settembre 2016
35. Du C, Sun H, Wang J, Qi Q, Liao J (2020) Adversarial and domain-aware BERT for cross-domain sentiment analysis. In: Proceedings of the 58th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Online, pp 4019–4028. https://doi.org/10.18653/v1/2020.acl-main.370. https://www.aclweb.org/anthology/2020.acl-main.370
36. ElSherief M, Nilizadeh S, Nguyen D, Vigna G, Belding EM (2018) Peer to peer hate: Hate speech instigators and their targets. In: Proceedings of the twelfth international conference on web and social media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018. AAAI Press, pp 52–61. https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17905
37. EU Commission (2016) Code of conduct on countering illegal hate speech online. https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/countering-illegal-hate-speech-online_en#theeucodeofconduct
38. Fernquist J, Lindholm O, Kaati L, Akrami N (2019) A study on the feasibility to detect hate speech in swedish. In: 2019 IEEE International conference on big data (big data), los angeles, CA, USA, December 9-12, 2019. IEEE, pp 4724–4729. https://doi.org/10.1109/BigData47090.2019.9005534
39. Fersini E, Nozza D, Rosso P (2018) Overview of the evalita 2018 task on automatic misogyny identification (AMI). In: Caselli T, Novielli N, Patti V, Rosso P (eds) 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 Workshop Proceedings, vol 2263, pp 1–9. CEUR-WS.org. http://ceur-ws.org/Vol-2263/paper009.pdf
40. Fersini E, Nozza D, Rosso P (2020) AMI @ EVALITA2020: automatic misogyny identification. In: Basile V, Croce D, Maro MD, Passaro LC (eds) Proceedings of the seventh evaluation campaign of natural language processing and speech tools for italian. final workshop (EVALITA 2020), Online event, December 17th, 2020, CEUR Workshop Proceedings, vol 2765. CEUR-WS.org. http://ceur-ws.org/Vol-2765/paper161.pdf
41. Fersini E, Rosso P, Anzovino M (2018) Overview of the task on automatic misogyny identification at ibereval 2018. In: Rosso P, Gonzalo J, Martínez R, Montalvo S, de Albornoz JC (eds) 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, CEUR Workshop Proceedings, vol 2150, pp 214–228. CEUR-WS.org. http://ceur-ws.org/Vol-2150/overview-AMI.pdf
42. Fiser D, Erjavec T, Ljubesic N (2017) Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in slovene. In: Waseem Z, Chung WHK, Hovy D, Tetreault JR (eds) Proceedings of the first workshop on abusive language online, ALW@ACL 2017, Vancouver, BC, Canada, August 4, 2017. Association for Computational Linguistics, pp 46–51. https://doi.org/10.18653/v1/w17-3007
43. Fortuna P, Nunes S (2018) A survey on automatic detection of hate speech in text. ACM Comput Surv 51(4):85:1–85:30. https://doi.org/10.1145/3232676
44. 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. Association for Computational Linguistics, Florence, pp 94–104. https://doi.org/10.18653/v1/W19-3510. https://www.aclweb.org/anthology/W19-3510
45. Founta A, Djouvas C, Chatzakou D, Leontiadis I, Blackburn J, Stringhini G, Vakali A, Sirivianos M, Kourtellis N (2018) Large scale crowdsourcing and characterization of twitter abusive behavior. In: Proceedings of the twelfth international conference on web and social media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018. AAAI Press, pp 491–500. https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17909
46. Gao L, Kuppersmith A, Huang R (2017) Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In: Kondrak G, Watanabe T (eds) Proceedings of the eighth international joint conference on natural language processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers. Asian Federation of Natural Language Processing, pp 774–782. https://www.aclweb.org/anthology/I17-1078/
47. de Gibert O, Pérez N., Pablos AG, Cuadros M (2018) Hate speech dataset from a white supremacy forum. In: Fiser D, Huang R, Prabhakaran V, Voigt R, Waseem Z, Wernimont J (eds) Proceedings of the 2nd workshop on abusive language online, ALW@EMNLP 2018, Brussels, Belgium, October 31, 2018. Association for Computational Linguistics, pp 11–20. https://doi.org/10.18653/v1/w18-5102
48. Glavaš G., Karan M, Vulić I (2020) XHate-999: Analyzing and detecting abusive language across domains and languages. In: Proceedings of the 28th international conference on computational linguistics. International Committee on Computational Linguistics, Barcelona, pp 6350–6365. https://www.aclweb.org/anthology/2020.coling-main.559
49. Golbeck J, Ashktorab Z, Banjo RO, Berlinger A, Bhagwan S, Buntain C, Cheakalos P, Geller AA, Gergory Q, Gnanasekaran RK, Gunasekaran RR, Hoffman KM, Hottle J, Jienjitlert V, Khare S, Lau R, Martindale MJ, Naik S, Nixon HL, Ramachandran P, Rogers KM, Rogers L, Sarin MS, Shahane G, Thanki J, Vengataraman P, Wan Z, Wu DM (2017) A large labeled corpus for online harassment research. In: Fox P, McGuinness DL, Poirier L, Boldi P, Kinder-Kurlanda K (eds) Proceedings of the 2017 ACM on Web Science Conference, WebSci 2017, Troy, NY, USA, June 25 - 28, 2017. ACM, pp 229–233. https://doi.org/10.1145/3091478.3091509
50. Gomez R, Gibert J, Gómez L, Karatzas D (2020) Exploring hate speech detection in multimodal publications. In: IEEE Winter conference on applications of computer vision, WACV 2020, snowmass village, CO, USA, March 1-5, 2020. IEEE, pp 1459–1467. https://doi.org/10.1109/WACV45572.2020.9093414
51. Goodfellow IJ, Bengio Y, Courville AC (2016) Deep Learning. Adaptive computation and machine learning. MIT Press, Cambridge. http://www.deeplearningbook.org/
52. van der Goot R, Ljubesic N, Matroos I, Nissim M, Plank B (2018) Bleaching text: Abstract features for cross-lingual gender prediction. In: Gurevych I, Miyao Y (eds) Proceedings of the 56th annual meeting of the association for computational linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers. Association for Computational Linguistics, pp 383–389. https://aclanthology.info/papers/P18-2061/p18-2061
53. Haddad H, Mulki H, Oueslati A (2019) T-HSAB: A tunisian hate speech and abusive dataset. In: Smaïli K (ed) Arabic language processing: from theory to practice - 7th international conference, ICALP 2019, Nancy, France, October 16-17, 2019, proceedings, communications in computer and information science, vol 1108. Springer, pp 251–263. https://doi.org/10.1007/978-3-030-32959-4_18

54. Hammer HL (2016) Automatic detection of hateful comments in online discussion. In: Maglaras LA, Janicke H, Jones KI (eds) Industrial networks and intelligent systems - second international conference, INISCOM 2016, Leicester, UK, October 31 - November 1, 2016, revised selected papers, lecture notes of the institute for computer sciences, social informatics and telecommunications engineering, vol 188, pp 164–173. https://doi.org/10.1007/978-3-319-52569-3_15
55. Hee CV, Lefever E, Verhoeven B, Mennes J, Desmet B, Pauw GD, Daelemans W, Hoste V (2015) Detection and fine-grained classification of cyberbullying events. In: Angelova G, Bontcheva K, Mitkov R (eds) Recent advances in natural language processing, RANLP 2015, 7-9 September, 2015, Hissar, Bulgaria, pp 672–680. RANLP 2015 Organising Committee / ACL. https://www.aclweb.org/anthology/R15-1086/
56. Hosseinmardi H, Mattson SA, Rafiq RI, Han R, Lv Q, Mishra S (2015) Analyzing labeled cyberbullying incidents on the instagram social network. In: Liu T, Scollon CN, Zhu W (eds) Social informatics - 7th international conference, SocInfo 2015, Beijing, China, December 9-12, 2015, Proceedings, Lecture Notes in Computer Science, vol 9471. Springer, pp 49–66. https://doi.org/10.1007/978-3-319-27433-1_4
57. Ibrohim MO, Budi I (2018) A dataset and preliminaries study for abusive language detection in indonesian social media. Procedia Comput Sci 135:222–229
58. Ibrohim MO, Budi I (2019) Multi-label hate speech and abusive language detection in Indonesian Twitter. In: Proceedings of the third workshop on abusive language online. Association for Computational Linguistics, Florence, pp 46–57. https://doi.org/10.18653/v1/W19-3506 https://www.aclweb.org/anthology/W19-3506
59. Ibrohim MO, Budi I (2019) Translated vs non-translated method for multilingual hate speech identification in twitter. Int J Adv Sci Eng Inf Technol 9(4):1116–1123
60. Ishmam AM, Sharmin S (2019) Hateful speech detection in public facebook pages for the bengali language. In: Wani MA, Khoshgoftaar TM, Wang D, Wang H, Seliya N (eds) 18th IEEE international conference on machine learning and applications, ICMLA 2019, Boca Raton, FL, USA, December 16-19, 2019. IEEE, pp 555–560. https://doi.org/10.1109/ICMLA.2019.00104
61. Jha A, Mamidi R (2017) When does a compliment become sexist? analysis and classification of ambivalent sexism using Twitter data. In: Proceedings of the Second Workshop on NLP and Computational Social Science. Association for Computational Linguistics, Vancouver, pp 7–16. https://doi.org/10.18653/v1/W17-2902. https://www.aclweb.org/anthology/W17-2902
62. Joshi P, Santy S, Budhiraja A, Bali K, Choudhury M (2020) The state and fate of linguistic diversity and inclusion in the NLP world. In: Proceedings of the 58th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Online, pp 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560. https://www.aclweb.org/anthology/2020.acl-main.560
63. Jurgens D, Chandrasekharan E, Hemphill L (2019) A just and comprehensive strategy for using NLP to address online abuse. In: Proceedings of the 57th annual meeting of the association for computational linguistics. Association for Computational Linguistics (ACL), pp 3658–3666
64. Karan M, Šnajder J (2018) Cross-domain detection of abusive language online. In: Proceedings of the 2nd workshop on abusive language online (ALW2). Association for Computational Linguistics, Brussels, pp 132–137. https://doi.org/10.18653/v1/W18-5117. https://www.aclweb.org/anthology/W18-5117
65. Kolhatkar V, Wu H, Cavasso L, Francis E, Shukla K, Taboada M (2019) The sfu opinion and comments corpus: a corpus for the analysis of online news comments. Corpus Pragmatics 4(2):1–36
66. Koufakou A, Pamungkas EW, Basile V, Patti V (2020) HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In: Proceedings of the fourth workshop on online abuse and harms. Association for Computational Linguistics, Online, pp 34–43. https://doi.org/10.18653/v1/2020.alw-1.5. https://www.aclweb.org/anthology/2020.alw-1.5
67. Kumar R, Ojha AK, Malmasi S, Zampieri M (2018) Benchmarking aggression identification in social media. In: Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018), Association for Computational Linguistics, Santa Fe, New Mexico, USA. pp 1–11. https://www.aclweb.org/anthology/W18-4401
68. Kumar R, Ojha AK, Malmasi S, Zampieri M (2020) Evaluating aggression identification in social media. In: Proceedings of the second workshop on trolling, aggression and cyberbullying. European Language Resources Association (ELRA), Marseille, pp 1–5. https://www.aclweb.org/anthology/2020.trac2-1.1
69. Lin YH, Chen CY, Lee J, Li Z, Zhang Y, Xia M, Rijhwani S, He J, Zhang Z, Ma X, Anastasopoulos A, Littell P, Neubig G (2019) Choosing transfer languages for cross-lingual learning. In: Proceedings of the 57th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Florence, pp 3125–3135. https://doi.org/10.18653/v1/P19-1301. https://www.aclweb.org/anthology/P19-1301
70. Ljubešić N., Erjavec T, Fišer D (2018) Datasets of Slovene and Croatian moderated news comments. In: Proceedings of the 2nd workshop on abusive language online (ALW2). Association for Computational Linguistics, Brussels, pp 124–131. https://doi.org/10.18653/v1/W18-5116. https://www.aclweb.org/anthology/W18-5116
71. MacAvaney S, Yao HR, Yang E, Russell K, Goharian N, Frieder O (2019) Hate speech detection: Challenges and solutions. Plos One 14(8):e0221152
72. Mandl T, Modha S, Majumder P, Patel D, Dave M, Mandalia C, Patel A (2019) Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in indo-european languages. In: Majumder P, Mitra M, Gangopadhyay S, Mehta P (eds) FIRE '19: Forum for information retrieval evaluation, Kolkata, India, December, 2019. ACM, pp 14–17. https://doi.org/10.1145/3368567.3368584
73. Mandl T, Modha S, Majumder P, Patel D, Dave M, Mandlia C, Patel A (2019) Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In: Proceedings of the 11th forum for information retrieval evaluation. ACM, pp 14–17
74. Mathur P, Shah RR, Sawhney R, Mahata D (2018) Detecting offensive tweets in hindi-english code-switched language. In: Ku L, Li C (eds) Proceedings of the sixth international workshop on natural language processing for social media, SocialNLP@ACL 2018, Melbourne, Australia, July 20, 2018. Association for Computational Linguistics, pp 18–26. https://doi.org/10.18653/v1/w18-3504
75. Meyer JS, Gambäck B (2019) A platform agnostic dual-strand hate speech detector. In: Proceedings of the third workshop on abusive language online. Association for Computational Linguistics, Florence, pp 146–156. https://doi.org/10.18653/v1/W19-3516. https://www.aclweb.org/anthology/W19-3516
76. Mishra P, Del Tredici M, Yannakoudakis H, Shutova E (2019) Author profiling for hate speech detection. arXiv:1902.06734

77. Mossie Z, Wang JH (2018) Social network hate speech detection for amharic language. Comput Sci In Technol 8:41–55
78. Mozafari M, Farahbakhsh R, Crespi N (2019) A bert-based transfer learning approach for hate speech detection in online social media. In: Cherifi H, Gaito S, Mendes JF, Moro E, Rocha LM (eds) Complex networks and their applications VIII - volume 1 proceedings of the eighth international conference on complex networks and their applications complex networks 2019, Lisbon, Portugal, December 10-12, 2019, Studies in Computational Intelligence, vol 881. Springer, pp 928–940. https://doi.org/10.1007/978-3-030-36687-2_77
79. Mozafari M, Farahbakhsh R, Crespi N (2020) Hate speech detection and racial bias mitigation in social media based on bert model. Plos one 15(8):e0237861
80. Mubarak H, Darwish K, Magdy W (2017) Abusive language detection on arabic social media. In: Waseem Z, Chung WHK, Hovy D, Tetreault JR (eds) Proceedings of the first workshop on abusive language online, ALW@ACL 2017, Vancouver, BC, Canada, August 4, 2017. Association for Computational Linguistics, pp 52–56. https://doi.org/10.18653/v1/w17-3008
81. Mulki H, Haddad H, Bechikh Ali C, Alshabani H (2019) L-HSAB: A levantine twitter dataset for hate speech and abusive language. In: Proceedings of the third workshop on abusive language online. Association for Computational Linguistics, Florence, pp 111–118. https://doi.org/10.18653/v1/W19-3512. https://aclanthology.org/W19-3512/
82. Nascimento G, Carvalho F, da Cunha AM, Viana CR, Guedes GP (2019) Hate speech detection using brazilian imageboards. In: dos Santos JAF, Muchaluat-Saade DC (eds) Proceedings of the 25th Brazillian Symposium on Multimedia and the Web, WebMedia 2019, Rio de Janeiro, Brazil, October 29 - November 01, 2019. ACM, pp 325–328. https://doi.org/10.1145/3323503.3360619
83. Nejadgholi I, Kiritchenko S (2020) On cross-dataset generalization in automatic detection of online abuse. https://arxiv.org/abs/2010.07414
84. Nithyanand R, Schaffner B, Gill P (2017) Measuring offensive speech in online political discourse. In: Penney J., Weaver N. (eds) 7th USENIX workshop on free and open communications on the internet, FOCI 2017, Vancouver, BC, Canada, August 14, 2017. USENIX Association. https://www.usenix.org/conference/foci17/workshop-program/presentation/nithyanand
85. Nobata C, Tetreault J, Thomas A, Mehdad Y, Chang Y (2016) Abusive language detection in online user content. In: Proceedings of the 25th international conference on world wide web, pp 145–153
86. Olteanu A, Castillo C, Boy J, Varshney KR (2018) The effect of extremist violence on hateful speech online. In: Proceedings of the twelfth international conference on web and social media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018. AAAI Press, pp 221–230. https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17908
87. Ombui E, Muchemi L, Wagacha P (2019) Hate speech detection in code-switched text messages. In: 2019 3Rd international symposium on multidisciplinary studies and innovative technologies (ISMSIT). IEEE, pp 1–6
88. Oriola O, Kotzé E (2020) Evaluating machine learning techniques for detecting offensive and hate speech in south african tweets. IEEE Access 8:21496–21509. https://doi.org/10.1109/ACCESS.2020.2968173
89. Ousidhoum N, Lin Z, Zhang H, Song Y, Yeung DY (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). Association for Computational Linguistics, Hong Kong, pp 4675–4684. https://doi.org/10.18653/v1/D19-1474 https://www.aclweb.org/anthology/D19-1474
90. Ozler KB, Kenski K, Rains S, Shmargad Y, Coe K, Bethard S (2020) Fine-tuning for multi-domain and multi-label uncivil language detection. In: Proceedings of the fourth workshop on online abuse and harms. Association for Computational Linguistics, Online, pp 28–33. https://doi.org/10.18653/v1/2020.alw-1.4. https://www.aclweb.org/anthology/2020.alw-1.4
91. Pamungkas EW, Basile V, Patti V (2020) Do you really want to hurt me? predicting abusive swearing in social media. In: Proceedings of the 12th language resources and evaluation conference, pp 6237–6246
92. Pamungkas EW, Basile V, Patti V (2020) Misogyny detection in twitter: a multilingual and cross-domain study. Inf Process Manag 57(6):102360. https://www.sciencedirect.com/science/article/pii/S0306457320308554
93. Pamungkas EW, Basile V, Patti V (2021) A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection, vol 58, p 102544
94. Pamungkas EW, Patti V (2019) Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In: Alva-Manchego F, Choi E, Khashabi D (eds) Proceedings of the 57th Conference of the association for computational linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 2: Student Research Workshop. Association for Computational Linguistics, pp 363–370. https://www.aclweb.org/anthology/P19-2051/
95. Pan SJ, Ni X, Sun J, Yang Q, Chen Z (2010) Cross-domain sentiment classification via spectral feature alignment. In: Rappa M, Jones P, Freire J, Chakrabarti S (eds) Proceedings of the 19th international conference on world wide web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010. ACM, pp 751–760. https://doi.org/10.1145/1772690.1772767
96. Park JH, Shin J, Fung P (2018) Reducing gender bias in abusive language detection. In: Proceedings of the 2018 conference on empirical methods in natural language processing. Association for Computational Linguistics, Brussels, pp 2799–2804. https://doi.org/10.18653/v1/D18-1302. https://www.aclweb.org/anthology/D18-1302
97. Pavlopoulos J, Malakasiotis P, Bakagianni J, Androutsopoulos I (2017) Improved abusive comment moderation with user embeddings. In: Popescu O, Strapparava C (eds) Proceedings of the 2017 workshop: natural language processing meets journalism, NLPmJ@EMNLP, Copenhagen, Denmark, September 7, 2017, pp 51–55. Association for Computational Linguistics. https://doi.org/10.18653/v1/w17-4209
98. de Pelle RP, Moreira VP (2017) Offensive comments in the brazilian web: a dataset and baseline results. In: Anais do VI brazilian workshop on social network analysis and mining. SBC, p 10
99. Pereira-Kohatsu JC, Sánchez L. Q., Liberatore F, Camacho-Collados M (2019) Detecting and monitoring hate speech in twitter. Sensors 19(21):4654. https://doi.org/10.3390/s19214654
100. Pérez J. M., Arango A, Luque F (2020) ANDES at SemEval-2020 task 12: A jointly-trained BERT multilingual model for offensive language detection. In: Proceedings of the fourteenth workshop on semantic evaluation. International Committee for Computational Linguistics, Barcelona, pp 1524–1531. https://www.aclweb.org/anthology/2020.semeval-1.199
101. Pitenis Z, Zampieri M, Ranasinghe T (2020) Offensive language identification in greek. In: Calzolari N, Béchet F, Blache P, Choukri K, Cieri C, Declerck T, Goggi S, Isahara H, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the 12th language resources and evaluation conference, LREC 2020, Marseille, France, May 11-16, 2020.

European language resources association, pp 5113–5119 https://www.aclweb.org/anthology/2020.lrec-1.629/
102. Poletto F, Basile V, Bosco C, Patti V, Stranisci M (2019) Annotating hate speech: Three schemes at comparison. In: Bernardi R, Navigli R, Semeraro G (eds) Proceedings of the sixth italian conference on computational linguistics, Bari, Italy, November 13-15, 2019, CEUR Workshop Proceedings, vol 2481. CEUR-WS.org. http://ceur-ws.org/Vol-2481/paper56.pdf
103. Poletto F, Basile V, Sanguinetti M, Bosco C, Patti V (2020) Resources and benchmark corpora for hate speech detection: A systematic review. Language resources and evaluation. https://link.springer.com/article/10.1007/s10579-020-09502-8
104. Ptaszynski M, Pieciukiewicz A, Dybała P (2019) Results of the poleval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in polish twitter. Proceedings of the PolEval 2019 Workshop pp 89
105. Qian J, Bethke A, Liu Y, Belding EM, Wang WY (2019) A benchmark dataset for learning to intervene in online hate speech. In: Inui K, Jiang J, Ng V, Wan X (eds) Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. Association for Computational Linguistics, pp 4754–476. https://doi.org/10.18653/v1/D19-1482
106. Qian J, ElSherief M, Belding EM, Wang WY (2018) Hierarchical CVAE for fine-grained hate speech classification. In: Riloff E, Chiang D, Hockenmaier J, Tsujii J (eds) Proceedings of the 2018 conference on empirical methods in natural language processing, Brussels, Belgium, October 31 - November 4, 2018, pp 3550–3559. Association for computational linguistics. https://doi.org/10.18653/v1/d18-1391
107. Qian J, ElSherief M, Belding EM, Wang WY (2019) Learning to decipher hate symbols. In: Burstein J, Doran C, Solorio T (eds) Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). Association for Computational Linguistics, pp 3006–3015. https://doi.org/10.18653/v1/n19-1305
108. Radfar B, Shivaram K, Culotta A (2020) Characterizing variation in toxic language by social context. In: Choudhury MD, Chunara R, Culotta A, Welles BF (eds) Proceedings of the fourteenth international AAAI conference on web and social media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020. AAAI Press, pp 959–963. https://aaai.org/ojs/index.php/ICWSM/article/view/7366
109. Rajamanickam S, Mishra P, Yannakoudakis H, Shutova E (2020) Joint modelling of emotion and abusive language detection. In: Jurafsky D, Chai J, Schluter N, Tetreault JR (eds) Proceedings of the 58th annual meeting of the association for computational linguistics, ACL 2020, Online, July 5-10, 2020, pp 4270–4279. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.394
110. Ranasinghe T, Zampieri M (2020) Multilingual offensive language identification with cross-lingual embeddings. In: Webber B, Cohn T, He Y, Liu Y (eds) Proceedings of the 2020 conference on empirical methods in natural language processing, EMNLP 2020, Online, November 16-20, 2020, pp 5838–5844. Association for computational linguistics. https://www.aclweb.org/anthology/2020.emnlp-main.470/
111. Rani P, Suryawanshi S, Goswami K, Chakravarthi BR, Fransen T, McCrae JP (2020) A comparative study of different state-of-the-art hate speech detection methods in Hindi-English code-mixed data. In: Proceedings of the second workshop on trolling, Aggression and Cyberbullying. European Language Resources Association (ELRA), Marseille. https://www.aclweb.org/anthology/2020.trac-1.7
112. Razo D, Kübler S (2020) Investigating sampling bias in abusive language detection. In: Proceedings of the fourth workshop on online abuse and harms, pp 70–78. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.alw-1.9, https://www.aclweb.org/anthology/2020.alw-1.9
113. Rezvan M, Shekarpour S, Balasuriya L, Thirunarayan K, Shalin VL, Sheth AP (2018) A quality type-aware annotated corpus and lexicon for harassment research. In: Akkermans H, Fontaine K, Vermeulen I, Houben G, Weber MS (eds) Proceedings of the 10th ACM conference on web science, WebSci 2018, Amsterdam, The Netherlands, May 27-30, 2018. ACM, pp 33–36. https://doi.org/10.1145/3201064.3201103
114. Ribeiro MH, Calais PH, Santos YA, Almeida VA, Meira Jr., W (2018) Characterizing and detecting hateful users on twitter. In: Proceedings of the twelfth international conference on web and social media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018. AAAI Press, pp 676–679. https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17837
115. Rizoiu M, Wang T, Ferraro G, Suominen H (2019) Transfer learning for hate speech detection in social media. arXiv:1906.03829
116. Rosa H, Carvalho JP, Calado P, Martins B, Ribeiro R, Coheur L (2018) Using fuzzy fingerprints for cyberbullying detection in social networks. In: 2018 IEEE International conference on fuzzy systems, FUZZ-IEEE 2018, Rio de Janeiro, Brazil, july 8-13, 2018. IEEE, pp 1–7. https://doi.org/10.1109/FUZZ-IEEE.2018.8491557
117. Ross B, Rist M, Carbonell G, Cabrera B, Kurowsky N, Wojatzki M (2017) Measuring the reliability of hate speech annotations: The case of the european refugee crisis. arXiv:1701.08118
118. Ruder S, Bingel J, Augenstein I, Søgaard A (2017) Sluice networks: Learning what to share between loosely related tasks. arXiv:1705.08142
119. Safi Samghabadi N, Hatami A, Shafaei M, Kar S, Solorio T (2020) Attending the emotions to detect online abusive language. In: Proceedings of the fourth workshop on online abuse and harms. Association for Computational Linguistics, pp 79–88 Online. https://doi.org/10.18653/v1/2020.alw-1.10 https://www.aclweb.org/anthology/2020.alw-1.10
120. Saha P, Mathew B, Goyal P, Mukherjee A (2019) Hatemonitors: Language agnostic abuse detection in social media. In: Working notes of FIRE 2019 - forum for information retrieval evaluation. pp 246–253, Kolkata, India
121. Salminen J, Hopf M, Chowdhury SA, Jung S. g., Almerekhi H, Jansen BJ (2020) Developing an online hate classifier for multiple social media platforms. Hum-Centric Comput Inf Sci 10(1):1
122. Sanguinetti M, Comandini G, Nuovo ED, Frenda S, Stranisci M, Bosco C, Caselli T, Patti V, Russo I (2020) Haspeede 2 @ EVALITA2020: overview of the EVALITA 2020 hate speech detection task. In: Basile V, Croce D, Maro MD, Passaro LC (eds) Proceedings of the seventh evaluation campaign of natural language processing and speech tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020, CEUR Workshop Proceedings, vol 2765. CEUR-WS.org. http://ceur-ws.org/Vol-2765/paper162.pdf
123. Sanguinetti M, Poletto F, Bosco C, Patti V, Stranisci M (2018) An italian Twitter corpus of hate speech against immigrants. In: Calzolari N, Choukri K, Cieri C, Declerck T, Goggi S, Hasida K, Isahara H, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S, Tokunaga T (eds) Proceedings of the eleventh international conference on language resources and evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018. European

Language Resources Association (ELRA), pp 2798–2805. http://www.lrec-conf.org/proceedings/lrec2018/summaries/710.html
124. Schäfer J, Burtenshaw B (2019) Offence in dialogues: A corpus-based study. In: Mitkov R, Angelova G (eds) Proceedings of the international conference on recent advances in natural language processing, RANLP 2019, Varna, Bulgaria, September 2-4, 2019. INCOMA Ltd. pp 1085–1093. https://doi.org/10.26615/978-954-452-056-4_125
125. Schmidt A, Wiegand M (2017) A survey on hate speech detection using natural language processing. In: Ku L, Li C (eds) Proceedings of the fifth international workshop on natural language processing for social media, SocialNLP@EACL 2017, Valencia, Spain, April 3, 2017. Association for Computational Linguistics, pp 1–10. https://doi.org/10.18653/v1/w17-1101
126. Schuster S, Gupta S, Shah R, Lewis M (2019) Cross-lingual transfer learning for multilingual task oriented dialog. In: Burstein J, Doran C, Solorio T (eds) Proceedings of the 2019 conference of the north american chapter of the association for computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). Association for Computational Linguistics, pp 3795–3805. https://doi.org/10.18653/v1/n19-1380
127. Sharma HK, Kshitiz K et al (2018) Nlp and machine learning techniques for detecting insulting comments on social networking platforms. In: 2018 International conference on advances in computing and communication engineering (ICACCE). IEEE, pp 265–272
128. Sigurbergsson GI, Derczynski L (2020) Offensive language and hate speech detection for danish. In: Calzolari N, Béchet F, Blache P, Choukri K, Cieri C, Declerck T, Goggi S, Isahara H, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk A, Piperidis S (eds) Proceedings of The 12th language resources and evaluation conference, LREC 2020, Marseille, France, May 11-16, 2020. European language resources association, pp 3498–3508. https://www.aclweb.org/anthology/2020.lrec-1.430/
129. Smith SL, Turban DHP, Hamblin S, Hammerla NY (2017) Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In: 5th International conference on learning representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference track proceedings. OpenReview.net. https://openreview.net/forum?id=r1Aab85gg
130. Sohn H, Lee H (2019) MC-BERT4HATE: Hate speech detection using multi-channel BERT for different languages and translations. In: Papapetrou P, Cheng X, He Q (eds) 2019 International conference on data mining workshops, ICDM Workshops 2019, Beijing, China, November 8-11, 2019. IEEE, pp 551–559. https://doi.org/10.1109/ICDMW.2019.00084
131. Sprugnoli R, Menini S, Tonelli S, Oncini F, Piras E (2018) Creating a WhatsApp dataset to study pre-teen cyberbullying. In: Proceedings of the 2nd workshop on abusive language online (ALW2). Association for computational linguistics, Brussels, pp 51–59. https://doi.org/10.18653/v1/W18-5107, https://www.aclweb.org/anthology/W18-5107
132. Stappen L, Brunn F, Schuller BW (2020) Cross-lingual zero- and few-shot hate speech detection utilising frozen transformer language models and AXEL. arXiv:2004.13850
133. Steinberger J, Brychcín T, Hercig T, Krejzl P (2017) Cross-lingual flames detection in news discussions. In: Mitkov R, Angelova G (eds) Proceedings of the international conference recent advances in natural language processing, RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017, pp. 694–700. INCOMA Ltd. https://doi.org/10.26615/978-954-452-049-6_089
134. Swamy SD, Jamatia A, Gambäck B (2019) Studying generalisability across abusive language detection datasets. In: Proceedings of the 23rd conference on computational natural language learning (CoNLL). Association for Computational Linguistics, Hong Kong, pp 940–950. https://doi.org/10.18653/v1/K19-1088, https://www.aclweb.org/anthology/K19-1088
135. Vashistha N, Zubiaga A (2021) Online multilingual hate speech detection: experimenting with hindi and english social media. Information 12(1):5
136. Vidgen B, Derczynski L (2020) Directions in abusive language training data: Garbage in, garbage out. arXiv:2004.01670
137. 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. Association for Computational Linguistics, Florence, pp 80–93. https://doi.org/10.18653/v1/W19-3509, https://www.aclweb.org/anthology/W19-3509
138. Vidgen B, Yasseri T (2018) Detecting weak and strong islamophobic hate speech on social media. arXiv:1812.10400
139. Vigna FD, Cimino A, Dell'Orletta F, Petrocchi M, Tesconi M (2017) Hate me, hate me not: Hate speech detection on facebook. In: Armando A, Baldoni R, Focardi R (eds) Proceedings of the first italian conference on cybersecurity (ITASEC17), Venice, Italy, January 17-20, 2017, CEUR Workshop Proceedings, vol 1816, pp 86–95. CEUR-WS.org. http://ceur-ws.org/Vol-1816/paper-09.pdf
140. Vu X, Vu T, Tran M, Le-Cong T, Nguyen HTM (2020) HSD shared task in VLSP campaign 2019: Hate speech detection for social good. arXiv:2007.06493
141. Wang K, Lu D, Han SC, Long S, Poon J (2020) Detect all abuse! toward universal abusive language detection models. In: Scott D, Bel N, Zong C (eds) Proceedings of the 28th international conference on computational linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020. International Committee on Computational Linguistics, pp 6366–6376. https://www.aclweb.org/anthology/2020.coling-main.560/
142. Wang Z, K K, Mayhew S, Roth D (2020) Extending multilingual BERT to low-resource languages. In: Findings of the association for computational linguistics: EMNLP 2020. Association for Computational Linguistics, Online, pp 2649–2656. https://doi.org/10.18653/v1/2020.findings-emnlp.240, https://www.aclweb.org/anthology/2020.findings-emnlp.240
143. Waseem Z, Davidson T, Warmsley D, Weber I (2017) Understanding abuse: A typology of abusive language detection subtasks. In: Proceedings of the first workshop on abusive language online, pp 78–84. Association for Computational Linguistics, Vancouver, BC, Canada. https://doi.org/10.18653/v1/W17-3012, https://www.aclweb.org/anthology/W17-3012
144. 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. Association for Computational Linguistics, San Diego, pp 88–93. https://doi.org/10.18653/v1/N16-2013, https://www.aclweb.org/anthology/N16-2013
145. Waseem Z, Thorne J, Bingel J (2018) Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In: Online harassment. Springer, pp 29–55
146. Wiegand M, Ruppenhofer J, Kleinbauer T (2019) Detection of abusive language: The problem of biased datasets. In: Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, pp 602–608. https://doi.org/10.18653/v1/N19-1060, https://www.aclweb.org/anthology/N19-1060
147. Wiegand M, Ruppenhofer J, Schmidt A, Greenberg C (2018) Inducing a lexicon of abusive words – a feature-based approach. In: Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human

language technologies, Volume 1 (Long Papers). Association for Computational Linguistics, New Orleans, pp 1046–1056. https://doi.org/10.18653/v1/N18-1095, https://www.aclweb.org/anthology/N18-1095
148. Wiegand M, Siegel M, Ruppenhofer J (2018) Overview of the germeval 2018 shared task on the identification of offensive language. In: 14Th conference on natural language processing KONVENS 2018, p 1
149. Wu S, Dredze M (2020) Are all languages created equal in multilingual BERT? In: Proceedings of the 5th workshop on representation learning for NLP, pp 120–130. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.repl4nlp-1.16, https://www.aclweb.org/anthology/2020.repl4nlp-1.16
150. Wulczyn E, Thain N, Dixon L (2017) Ex machina: Personal attacks seen at scale. In: Barrett R, Cummings R, Agichtein E, Gabrilovich E (eds) Proceedings of the 26th international conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017. ACM, pp 1391–1399. https://doi.org/10.1145/3038912.3052591
151. Yuan Z, Wu S, Wu F, Liu J, Huang Y (2018) Domain attention model for multi-domain sentiment classification. Knowl Based Syst 155:1–10. https://doi.org/10.1016/j.knosys.2018.05.004
152. Zampieri M, Malmasi S, Nakov P, Rosenthal S, Farra N, Kumar R (2019) Predicting the type and target of offensive posts in social media. In: Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, pp 1415–1420. https://doi.org/10.18653/v1/N19-1144, https://www.aclweb.org/anthology/N19-1144
153. 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. Association for Computational Linguistics, Minneapolis, Minnesota, USA. https://doi.org/10.18653/v1/S19-2010, https://www.aclweb.org/anthology/S19-2010
154. Zampieri M, Nakov P, Rosenthal S, Atanasova P, Karadzhov G, Mubarak H, Derczynski L, Pitenis Z, Çöltekin Ç (2020) SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In: Proceedings of the fourteenth workshop on semantic evaluation, pp 1425–1447. International Committee for Computational Linguistics, Barcelona (online). https://www.aclweb.org/anthology/2020.semeval-1.188
155. Zhang X, Tong J, Vishwamitra N, Whittaker E, Mazer JP, Kowalski R, Hu H, Luo F, Macbeth J, Dillon E (2016) Cyberbullying detection with a pronunciation based convolutional neural network. In: 15Th IEEE international conference on machine learning and applications, ICMLA 2016, anaheim, CA, USA, December 18-20, 2016. IEEE Computer Society, pp 740–745. https://doi.org/10.1109/ICMLA.2016.0132
156. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, Xiong H, He Q (2021) A comprehensive survey on transfer learning. Proc IEEE 109(1):43–76. https://doi.org/10.1109/JPROC.2020.3004555