



**HAL**  
open science

## A survey on multi-lingual offensive language detection

Khoulood Mnassri, Reza Farahbakhsh, Razieh Chalehchaleh, Praboda Rajapaksha, Amir Reza Jafari, Guanlin Li, Noel Crespi

► **To cite this version:**

Khoulood Mnassri, Reza Farahbakhsh, Razieh Chalehchaleh, Praboda Rajapaksha, Amir Reza Jafari, et al.. A survey on multi-lingual offensive language detection. PeerJ Computer Science, 2024, 10.7717/peerj-cs.1934 . hal-04475626

**HAL Id: hal-04475626**

**<https://hal.science/hal-04475626>**

Submitted on 28 May 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# 1 A Survey on Multi-lingual Offensive 2 Language Detection

3 **Khouloud Mnassri, Reza Farahbakhsh, Razieh Chalehchaleh, Praboda  
4 Rajapaksha, Amir Reza Jafari, Guanlin Li, and Noel Crespi**

5 **Samovar, Telecom SudParis, Institut Polytechnique de Paris, 91120 Palaiseau, France.**

6 Corresponding author:

7 Khouloud Mnassri

8 Email address: khouloud.mnassri@telecom-sudparis.eu

## 9 **ABSTRACT**

10 The prevalence of offensive content on online communication and social media platforms is growing  
11 more and more common, which makes its detection difficult, especially in multilingual settings. The term  
12 “Offensive Language” encompasses a wide range of expressions, including various forms of hate speech  
13 and aggressive content. Therefore, exploring multilingual offensive content, that goes beyond a single  
14 language, focus and represents more linguistic diversities and cultural factors. By exploring multilingual  
15 offensive content, we can broaden our understanding and effectively combat the widespread global  
16 impact of offensive language. This survey examines the existing state of multilingual offensive language  
17 detection, including a comprehensive analysis on previous multilingual approaches, and existing datasets,  
18 as well as provides resources in the field. We also explore the related community challenges on this  
19 task, which include technical, cultural, and linguistic ones, as well as their limitations. Furthermore, in  
20 this survey we propose several potential future directions toward more efficient solutions for multilingual  
21 offensive language detection, enabling safer digital communication environment worldwide.

## 22 **INTRODUCTION**

23 Online offensive language has become an increasingly prevalent issue that has widespread concern among  
24 policymakers, civil society organizations, and even the general public. The extended use of social media  
25 platforms is an important facilitator to express offensive and hateful content, especially with the anonymity  
26 provided, and to disseminate them widely to a global audience. This has led to a rise in this content, which  
27 has had serious negative consequences for many individuals and communities from various demographics.  
28 Given the rise in the prevalence of offensive language on popular social media platforms such as Twitter  
29 and Facebook, there has been a growing number of proposed techniques for identifying them specifically  
30 in the monolingual (i.e., English) setting. Although research on the multilingual dimension of offensive  
31 language is a relatively recent area of research, numerous studies have attempted to address this issue  
32 comprehensively. The detection of this content helps to identify patterns and trends across different  
33 languages and cultures, allowing for a better understanding of the underlying factors that contribute to  
34 offensive language in different contexts. Knowledge gained through this can be used to develop targeted  
35 interventions to address hate speech more effectively.

36 This survey provides a comprehensive overview of the detection of offensive language and more  
37 specifically hate speech in a multilingual setting using various techniques. It examines early research  
38 and cutting-edge approaches, highlighting gaps and improvements in multilingual and cross-lingual  
39 existing models. The study carefully analyzes multilingual datasets, outlines global initiatives and projects  
40 combating offensive language, and underscores the importance of readily available open-source tools in  
41 fostering further research and practical applications. Finally, it assesses current challenges and potential  
42 solutions, aiming to set a future research direction in multilingual offensive language detection.

43 While there has been a number of survey papers on hate speech detection in general, our survey paper  
44 is among the first few studies to provide a comprehensive summary of how the problem is addressed  
45 in multilingual scenarios including an extensive summary of the datasets used. Moreover, this paper  
46 highlights significant resources such as projects, products, and APIs that are essential for analyzing

47 multilingual hatred and offensive content. Finally, we discuss the challenges and limitations of the existing  
48 techniques and potential future works.

49 **Data availability statement:** *All relevant of this manuscript, including the elements we mentioned*  
50 *in our survey (existing studies, datasets, and resources) are available (and will be completed and*  
51 *actively maintained): [https://github.com/KhouloudMN97/A-Survey-on-Multi-lingual-Offensive-Language-](https://github.com/KhouloudMN97/A-Survey-on-Multi-lingual-Offensive-Language-Detection/tree/main)*  
52 *Detection/tree/main.*

## 53 **Motivation and Impact**

54 Based on the latest research, there are various compelling reasons for investigating and scrutinizing hateful  
55 content on social media platforms, with the ultimate goal of fostering a secure, considerate, and diverse  
56 online community. Such motivations include i) Detection and the mitigation of harmful behavioural  
57 patterns by analysing the trends of hateful content (Al-Hassan and Al-Dossari (2019)) ii) Preserving a safe  
58 online community by analysing and diminishing offensive, insincere and unsafe content from the social  
59 media platforms (d'Sa et al. (2020)) iii) Analyzing and safeguarding marginalized user communities that  
60 are targeted by hate speech and abusive content due to their race, ethnicity, gender, sexuality, religion, or  
61 other identifiable traits (Al-Hassan and Al-Dossari (2019)) and most importantly iv) Ensuring adherence  
62 to legal and regulatory frameworks in multiple countries that prohibit the propagation of hate speech and  
63 other forms of harmful content Bakalis and Hornle (2021). As a result, analyzing hateful and abusive  
64 content can contribute to a more positive user experience on social media platforms by providing a safer  
65 and more respectful online community, as well as support efforts to combat online hate speech and provide  
66 insight into real user behaviours.

## 67 **Definition of Hate Speech and Offensive Language**

68 One primary challenge in recognizing content as offensive or not is that, up to date there is no widely  
69 acknowledged unique definition of offensive and hate speech. This is primarily due to the ambiguous,  
70 subjective, and personal interpretations of whether a speech is “hatred” or expresses “offensive” (Fortuna  
71 et al. (2020)). Unspecified definitions can result in increased subjectivity in annotations which then  
72 facilitates generating a biased model (Davidson et al. (2019)).

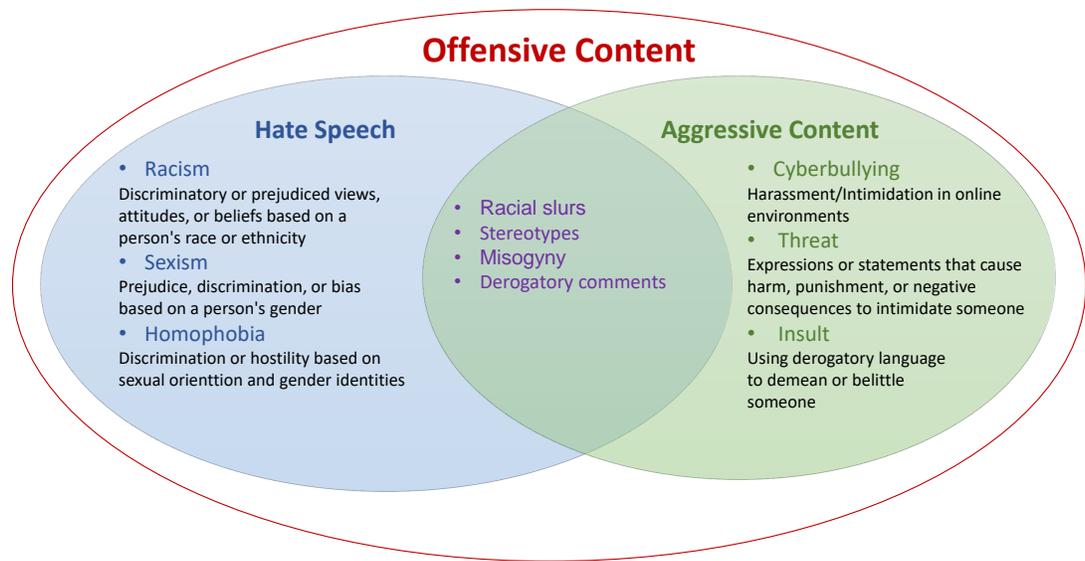
73 As shown in Fig. 1, due to the shared characteristics and effects, hate speech and aggressive content  
74 can be categorized as types of offensive content (Poletto et al. (2021)). Offensive language can encompass  
75 a broader range of content that may cause discomfort or offense, while hate speech is specifically targeted  
76 at marginalized communities or individuals, aiming to spread prejudice, hostility, or discrimination,  
77 and subject to legal and social consequences. Researchers primarily focus on analyzing hate speech  
78 rather than general offensive content detection due to its severity of harm and the significant legal and  
79 policy implications it carries in many countries. Compared with the law-enforcing based hate-speech  
80 definitions such as the descriptions provided by, the Europe union commission (Wigand and Voin (2017)),  
81 the European Union Agency for Fundamental Rights (<https://fra.europa.eu>) and the UN Human rights  
82 (<https://www.ohchr.org>), social media platforms should consider a broader definition for the hatred and  
83 offensive content.

84 Generally, hate speech in the context of social media commonly refers to any content that discriminates  
85 against or attacks individuals or groups based on several characteristics such as race, ethnicity, religion,  
86 sexual orientation, gender, or other identifiable features. This definition often aligns with community  
87 policies and content moderation guidelines of various social media platforms, drawing from assigned  
88 legal frameworks, academic research, and societal norms concerning discriminative behavior and hateful  
89 content online.

## 90 **Multilingual vs. Cross-Lingual**

91 In NLP, both multi-lingual and cross-lingual approaches deal with processing and understanding languages,  
92 which usually are different from English. However, there are differences in the way of using these terms:

93 **Multi-lingual NLP:** Multi-lingual NLP refers to the development and application of NLP models and  
94 techniques that can handle multiple languages simultaneously. This involves creating models that can  
95 process and understand different languages without the need for language-specific models. Multi-lingual  
96 models are trained on data from multiple languages and learn to capture shared linguistic patterns and  
97 representations across languages. These models can perform tasks such as text classification, named  
98 entity recognition, or machine translation across multiple languages. They are designed to provide a



**Figure 1.** Categories of Offensive Content: Offensive content, hate speech, and aggressive content are distinct but they are often overlapping categories that involve harmful or negative expressions. While there is some overlap between these categories, their distinctions lie in their specific intentions, targets, and impacts on individuals or communities.

99 generalized solution for various languages, but they may not achieve the same level of performance as  
100 language-specific models.

101 Cross-lingual NLP: Cross-lingual NLP focuses on enabling communication and understanding be-  
102 tween different languages. It involves developing techniques to transfer knowledge, resources, or models  
103 from a resource-rich language (often referred to as a source language) to resource-poor languages (target  
104 languages). The goal is to leverage the knowledge and resources available for one language to enhance  
105 NLP tasks in another language. Cross-lingual approaches can include tasks such as cross-lingual docu-  
106 ment classification, cross-lingual information retrieval, or cross-lingual word embeddings. Techniques  
107 used in cross-lingual NLP include machine translation, word alignment, parallel corpora, and bilingual  
108 dictionaries (Pikuliak et al. (2021)).

109 **Structure of this survey**

110 Our survey incorporates an extensive exploration in existing approaches and datasets of multilingual  
111 hate speech and offensive language detection. The structure of our review is as follows: In section  
112 ‘Background on Multilingual Hate Speech Phenomena’, we provide an examination of the previous  
113 surveys on hate speech detection. We focus on the studies on multilingual aspect of this task, where  
114 we carefully study their deficiencies in order to fill this gap in our survey. Next, section ‘Approaches on  
115 Multilingual Hate Speech Detection’ presents a thorough review of the existing approaches used for  
116 multilingual and cross-lingual offensive language detection. We give interpretations from our findings and  
117 summarize these studies in a comprehensive table. Following that, we examined the available resources in  
118 this field, starting in section ‘Datasets on Multilingual Hate Speech Detection’ with a detailed review of  
119 available multilingual datasets by a deep analysis of these corpora introducing their languages and main  
120 topics. Next, in section ‘Resources for Multilingual Hate Speech Detection’ we illustrate the different  
121 international collaborative projects provided for multilingual hate speech detection. We also present  
122 community challenges and competitions that focused this task. We, then, introduce a variety of publicly  
123 available source codes and APIs. Lastly, in section ‘Challenges and limitations’, we present the challenges  
124 encountered in the field of multilingual hate speech detection. We also emphasize the limitations, as well  
125 as a set of future directions that, we believe, will help to overcome these obstacles and to progress further  
126 in the task of multilingual hate speech detection.

127 **Who can benefit from this survey?** This survey aims to serve as a pivotal roadmap for both

128 the research community and business sector, delving into the current landscape and future direction  
129 of multilingual offensive language detection. For researchers in the academic field, it serves as a  
130 comprehensive synthesis, describing the ongoing advancements, evolving approaches, and available  
131 resources, including datasets, within this domain. It carefully discusses challenges, and limitations, and  
132 proposes promising directions for future exploration. In the business sector, this survey offers valuable  
133 understanding, serving as a guide for decision-makers. It helps in the assessment of the implementation  
134 of multilingual offensive language detection moderators for online textual content, particularly on social  
135 media platforms.

136 In order to give an in depth understanding of the state of the multilingual offensive language detection  
137 field, we provide an extensive review in this paper. We carefully explore a number of aspects: mainly  
138 current approaches and a wide range of datasets that include low-resource languages, and also, available  
139 resources like collaborative projects and tools, challenges faced and possible recommendations for future  
140 development. As a result, academics as well as professionals can benefit considerably from our survey.  
141 This study gives significant recommendations and defines solutions for future directions, especially  
142 proposing getting benefit from the recently released LLMs and generative pre-trained models, therefore,  
143 contributing to the compilation of existing knowledge and supporting ongoing progress in reducing hate  
144 speech across languages and groups.

## 145 **BACKGROUND ON MULTILINGUAL HATE SPEECH PHENOMENA**

### 146 **Previous Surveys**

147 Several previous studies have given comprehensive analysis of hate speech detection in different aspects,  
148 focusing on presenting to the community the related data, approaches and multilingual methods, and  
149 existing products. In this section, we aim to explore these surveys paying more attention to the multilingual  
150 aspect of the field. As displayed in Table 1, some of the existing survey studies focus on presenting  
151 definitions and notions related to the domain, as well as an examination of current approaches like in  
152 Poletto et al. (2021); Pamungkas et al. (2021b); Chhabra and Vishwakarma (2023). Meanwhile, other  
153 surveys focus on introducing the available sources of the topic, such as data and available source-code  
154 Vidgen and Derczynski (2020), (Poletto et al. (2021); Schmidt and Wiegand (2017)).

155 Table 1 illustrates previous literature reviews in the field of hate speech detection in multilingual and  
156 in general settings. We have carefully studied and analyzed these surveys, which enabled us to construct a  
157 comprehensive narrative review in order to cover the deficiencies in these surveys. More specifically, we  
158 aim to give an overview of the existing multilingual and cross-lingual approaches, similar to Schmidt and  
159 Wiegand (2017); Yin and Zubiaga (2021); Fortuna and Nunes (2018) (on English data), to Pamungkas  
160 et al. (2021b) focusing on cross-lingual methods, as well as Chhabra and Vishwakarma (2023), where they  
161 basically displayed monolingual approaches in some specific languages. Adding to that, some previous  
162 surveys have presented existing corpora in the domain, in some specific languages as in Poletto et al.  
163 (2021); Jahan and Oussalah (2021); Chhabra and Vishwakarma (2023), and more widely in Pamungkas  
164 et al. (2021b) presenting datasets in 18 different languages, and in Vidgen and Derczynski (2020), pro-  
165 viding an open source website to 63 datasets in multiple languages (<https://hatespeechdata.com/>). One  
166 metric aspect, to take into consideration to define our survey type is based on the forms of literature  
167 reviews: *Narrative* and *Systematic* reviews. In fact, when writing a narrative review from the literature  
168 findings, authors are utilizing a more subjective and qualitative research technique. However, they employ  
169 methodical/systematic research methodology using a more quantitative and objective approach when  
170 working on systematic reviews. These reviews are considered as a link between practice or policy-making  
171 and research. Narrative reviews are also noticed to have this linking function, however, they are usually  
172 utilized in order to handle more general and complex subjects (Hammersley (2001)).

### 173 174 **Surveys on Hate Speech - from general perspective:**

175 Several surveys have studied hate speech and offensive language detection. In 2017, Schmidt and  
176 Wiegand (2017) summarized the primary NLP aspects of automated hate speech detection: illustrating  
177 the various types of feature representation and, the existing supervised and semi-supervised techniques  
178 and their limitations. The authors, proposed as future direction, the urge to analyze hate speech detection  
179 from a multilingual viewpoint. Adding to that, Yin and Zubiaga (2021) described the most commonly  
180 implemented approaches in the field, such as dictionaries, bag-of-words, N-grams, among others. More-  
181 over, Fortuna and Nunes (2018) presented a systematic overview. They recap the various approaches

Title	Main focus of the survey	How to differentiate it with our survey	Year	Type*
<b>Surveys on Hate Speech in General</b>				
<b>A Survey on Hate Speech Detection using Natural Language Processing (Schmidt and Wiegand (2017))</b>	Investigating automated identification of hate speech through NLP, using linguistic and semantic features. Highlighting how identifying user profiles involved in spreading hateful content, and presenting supervised and semi-supervised approaches on English data.	No focus on multilingual aspect of hate speech detection, only on studies conducted on English datasets.	2017	Narrative
<b>Towards generalisable hate speech detection: a review on obstacles and solutions (Yin and Zubiaga (2021))</b>	Presenting NLP methods used for automated hate speech detection on online social media networks.	Not presenting multilingual hate speech detection.	2021	Narrative
<b>A Survey on Automatic Detection of Hate Speech in Text (Fortuna and Nunes (2018))</b>	Provide an overview of the researches conducted in hate speech detection, which includes describing available methods and resources.	General overview about hate speech detection, not focusing on the multilingual aspect of the subject.	2018	Systematic
<b>A systematic review of Hate Speech automatic detection using Natural Language Processing (Jahan and Oussalah (2021))</b>	Focusing on the use of deep learning technologies and architectures in hate speech detection, with emphasis the sequence of pipeline processing.	Although presenting some resources (datasets and some available Github projects) in different languages, but no detailed overview on multilinguality.	2023	Systematic
<b>Surveys on Multilingual Hate Speech</b>				
<b>Directions in abusive language training data, a systematic review: Garbage in, garbage out (Vidgen and Derczynski (2020))</b>	Comprehensive review of 63 abusive language datasets in several languages. It addresses the opportunities and problems of open science in this area and provides experts building new abusive content databases.	Only focus on datasets, not considering existing multilingual approaches in the field.	2020	Systematic
<b>Resources and benchmark corpora for hate speech detection: a systematic review (Poletto et al. (2021))</b>	Analyzing the annotated collections of texts released by the broader community, considering their method of creation, topic, language range, and other pertinent factors.	No analysis of the existing methods used in multilingual hate speech detection.	2021	Systematic
<b>Towards multidomain and multilingual abusive language detection: a survey (Pamungkas et al. (2021b))</b>	A study of existing researches about the available datasets and methods used in cross-domain and cross-lingual cases.	Focus on cross-lingual side only in the hate speech detection. No analysis on the available products or resources in the community, used and can be used in multilingual detection of hate speech.	2023	Narrative
<b>A literature survey on multimodal and multilingual automatic hate speech identification (Chhabra and Vishwakarma (2023))</b>	A survey of hate speech identification methods (strengths and weaknesses), and popular benchmark datasets.	Presenting approaches in several languages (monolingual), but no focus on multilingual nor cross-lingual approaches.	2023	Narrative

**Table 1. Key previous surveys on the topic of (multilingual) hate speech detection.**

\*A **narrative review** is a more subjective and qualitative study used to create a story from the literature in order to summarize the findings. In contrast, a **systematic review** is more objective and quantitative, used to discover and evaluate the available literature in order to address a certain research topic (Hammersley (2001)).

182 and resources available in this domain. They also critically examined valuable resources like datasets,  
183 illustrating some of the existings in different languages (English, Dutch and German). Another systematic  
184 survey has been conducted in 2021 (then updated in 2023), where Jahan and Oussalah (2021) addressed  
185 thorough aspects including the language used, the pipeline processing, and the techniques used, focusing  
186 primarily on deep learning approaches. They also presented many sources in some languages (data and  
187 GitHub projects).

### 188 **Surveys on Multilingual Hate Speech - multilingual perspective:**

189 Multilinguality is getting more popular in the task of hate speech detection and there are recently some  
190 surveys on this aspect, covering existing approaches and resources. These studies pay more attention to  
191 the considerable variations in the existing studies that aim to cover other languages (other than English),  
192 as well as more concepts related to offensive language (Racism, sexism, among others). This is required to  
193 build more generalized approaches. In 2021, Poletto et al. (2021) comprehensively studied the annotated  
194 datasets of hate speech, taking into account the creation process, subject case, language coverage, and other  
195 pertinent factors about the existing lexica and benchmark datasets in different languages. Moreover, few  
196 overviews have been conducted on the topic of multilingual offensive language detection, where the survey  
197 of Pamungkas et al. (2021b) presented the approaches and the available corpora employed in cross-domain  
198 and cross-language techniques. Moreover, the survey of Chhabra and Vishwakarma (2023) provided a  
199 comprehensive review of hate speech definitions, exploring the essential textual analysis procedures used.  
200 The survey also described the advantages and disadvantages of multimodal and cross-lingual approaches.

### 201 **Existing Gaps in the Previous Surveys**

202 Although several review studies have been written on the task of hate speech detection, but still there are  
203 several aspects that are not covered in those studies especially when it comes to presenting an in-depth

204 comprehension of the multilingual aspect of this area. Our study is a narrative review that seeks to close  
205 this gap by exploring larger number of characteristics of multilingual offensive language detection: from  
206 existing approaches and available datasets to related collaborative projects and resource products that  
207 include community challenges, source codes, and APIs. Adding to that, this study will also analyze  
208 the associated challenges and limitations in the field. Furthermore, we aim to get into future research  
209 directions providing a roadmap for the research progress in multilingual hate speech detection.

## 210 **APPROACHES ON MULTILINGUAL HATE SPEECH DETECTION**

### 211 **Existing Approaches**

212 The value of studying multilingual offensive language detection has earned attention in recent years.  
213 This increasing interest is a consequence of the linguistic variety within social media platforms. The  
214 availability of multilingual datasets, especially from social media platforms that are used worldwide, has  
215 made it possible to develop algorithms to detect this content in various languages. Some research studies  
216 used to focus on creating monolingual models, working on languages other than English, but they have  
217 later evolved into cross-lingual and multilingual approaches, utilizing rich resource languages in order  
218 to detect the offensive language in low-resource ones (totally unseen using zero-shot learning, or, using  
219 few-shot learning (Goodfellow et al. (2016))).

220 This section aims to provide an overview of existing approaches for this task. To that end, we organize  
221 the existing studies into eight distinct groups (as shown in the first column of Table 2): Traditional  
222 Machine Learning (where we found Logistic Regression (LR) models), Deep Neural Networks (DNN),  
223 Transfer Learning (TL), Machine Translation (MT), Ensemble Learning (EL), Meta Learning (Meta-L),  
224 Multitask Learning (Multitask-L), and Unsupervised Learning (UL). By categorizing these approaches,  
225 we make it possible to analyze them carefully and present an overview of the evolution of methods used  
226 to tackle multilingual and cross-lingual offensive language detection, as described in the next subsection  
227 ‘Analysis of the existing approaches’.

228 As for languages presentation, we will use ISO 639-2 codes ([https://www.loc.gov/standards/iso639-2/php/code\\_list.php](https://www.loc.gov/standards/iso639-2/php/code_list.php)).

230 **Methodology of research:** Our research of these existing approaches is based on information from  
231 earlier pertinent surveys (mentioned in the previous section). We also used specific keywords, such  
232 as “multilingual/cross-lingual offensive language detection”, “multilingual/cross-lingual hate speech  
233 detection”, “multilingual/cross-lingual abusive language detection, among others, to find studies that  
234 were published in IEEE Xplore, ACM Digital Library, Google Scholar, among others. By focusing  
235 on publications that were released in 2019 and beyond (until July 2023), we ensured that our survey  
236 included the most cutting-edge approaches for multilingual and cross-lingual offensive language detection.  
237 Moreover, we won’t cover monolingual approaches in our study about existing approaches since our focus  
238 is basically directed to multilingual approaches, however, we mention some of the most relevant ones we  
239 found in ‘Other Technologies’ subsection. A summary of the identified approaches are presented in Table  
240 2 and each of the eight techniques are detailed in the following part.

### 241 ***Logistic Regression (LR)***

242 There aren’t many machine learning-based approaches for detecting multilingual offensive language. In  
243 fact, deep neural networks and transfer learning-based methods have shown more effectiveness in this  
244 field, especially by utilizing pre-trained language models. Traditionally, the most widely used machine  
245 learning techniques included Naive Bayes, k-nearest neighbors, decision trees, random forests, and  
246 support vector machines. However, in recent years Deep Neural Networks have almost completely  
247 substituted or at least surpassed these methods, particularly in NLP and in sentiment analysis (Otter et al.  
248 (2018)). But traditional machine learning models might still be considered potential solutions to this issue  
249 because they are effective at identifying offensive language. On this scope, a multilingual hate speech  
250 and abusive language detection system was developed by Vashistha and Zubiaga (2021), and trained on a  
251 significant textual dataset of hate speech in English and Hindi. They demonstrated an online retraining  
252 capability for the system to identify new varieties of hate speech or linguistic patterns using LR. Moreover,  
253 cross-lingual (zero-shot and few-shot learning) experiments were executed by Aluru et al. (2020) on nine  
254 different languages. They analyzed different combinations of vector representations and machine learning  
255 algorithms, including MUSE and LASER embeddings. As a result, LASER and an LR model proved to  
256 be the most effective combined model. Adding to that, Bigoulaeva et al. (2021) utilized a Support Vector

Techniques	Ref. ◇/♣	Focused Languages	Approach (feature extraction methods)	Year
Logistic Regression (LR)	Vashistha and Zubiaga (2021) ◇	Hi, En and Code Mixed	Word embedding for feature extraction	2020
	Aluru et al. (2020) ◇	Ar, En, De, Id, It, Pl, Pt, Es and Fr	MUSE and LASER for feature extraction	2020
Deep Neural Network (DNN)	Vashistha and Zubiaga (2021) ◇	Hi, En and Code Mixed	CNN-LSTM (Word embedding)	2020
	Elouali et al. (2020) ◇	Ar, It, Pt, Id, En, De, Hi-En Code Mixed	CNN (Character-level representation)	2020
	Jiang and Zubiaga (2021) ♣	En, Es and It: 6 languages pairs	Bi-LSTM based capsule network (FastText)	2021
Transfer Learning (TL)	Vashistha and Zubiaga (2021) ◇	Hi, En and code mixed	BERT	2020
	Aluru et al. (2020) ◇	Ar, En, De, Id, It, Pl, Pt, Es, Fr	mBERT	2020
	Wang et al. (2020) ◇	En, Tr, Da, El and Ar	XML-R	2020
	Bhatia et al. (2021) ◇	En, Hi and Mr	XML-R, mBERT, DistilmBERT (emoji2vec)	2021
	Roy et al. (2021a) ◇	En, De, Hi	XML-R	2021
	Deshpande et al. (2022a) ◇	En, Ar, De, Id, It, Pt, Es, Fr, Tr, Da and Hi	mBERT (MUSE and LASER)	2022
	zahra El-Alami et al. (2022) ◇	En and Ar	BERT, mBERT and AraBERT	2022
	Ghadery and Moens (2020) ♣	En, Da, El, Ar and Tr	mBERT	2020
	Ranasinghe and Zampieri (2020) ♣	En, Hi, Bn and Es	XML-R	2020
	Dadu and Pant (2020) ♣	En, El, Da, Ar and Tr	XML-R	2020
	Stappen et al. (2020) ♣	En to Es	XML-R based AXEL	2020
	Ranasinghe and Zampieri (2021b) ♣	En, Ar, Bn, Da, El, Hi, Es, and Tr	XML-R	2021
	Pelicon et al. (2021a) ♣	En, Es, De, Id and Ar	mBERT, LASER	2021
	Tita and Zubiaga (2021) ♣	En, Fr	mBERT, XML-R	2021
	Ranasinghe and Zampieri (2021a) ♣	En and 6 Indian languages: Indo-Aryan (Bn, Hi-En, Ur-En) and Dravidian (Kn- En, Malayalam-En, Ta-En)	mBERT, XML-R	2021
	Pelicon et al. (2021b) ♣	Ar, Hr, De, En, and SI	mBERT, CseBERT	2021
Vitiugin et al. (2021) ♣	En and Es	MLIAN: Multilingual Interactive Attention Network (LASER, DistilmBERT)	2021	
Eronen et al. (2022) ♣	En, De, Da, Pl, Ru, Ja and Ko	mBERT, XML-R	2022	
Zia et al. (2022) ♣	En, Es, It, De, Ar, El and Tr	RoBERTa, BERT	2022	
Machine Translation (MT)	Ibrohim and Budi (2019b) ♣	Hi, En, and Id	Google Translate API to translate all data between source and target languages.	2019
	Aluru et al. (2020) ♣	Ar, En, De, Id, It, Pl, Pt, Es and Fr	Google Translate API to translate all the datasets in different languages to English = input to BERT	2020
	Jiang and Zubiaga (2021) ♣	En, Es and It: 6 languages pairs	Google Translate API to translate all data between source and target languages	2021
	Pamungkas et al. (2021a) ♣	En, Fr, De, Id, It, Pt and Es	Google Translate API to translate all datasets into En = input to BERT	2021
Ensemble Learning (EL)	Cohen et al. (2023) ◇	En, De, Fr, Es and No	Based DeBERTa: Simple averaging, weighted averaging based on AUC, and LightGBM using predictions as input.	2023
	Ahn et al. (2020a) ♣	En, El, Da, Ar and Tr	Majority Voting based mBERT	2020
	Bigoulaeva et al. (2021) ♣	En and De	Based Bilingual word embeddings: FastText then MUSE	2021
	Bigoulaeva et al. (2022) ♣	En and De	Based mBERT, CNN and LSTM (Cross-Lingual Word Embeddings)	2022
Bigoulaeva et al. (2023) ♣	En and Es	Based mBERT	2023	
Meta Learning (Meta-L)	Vadakkera Suresh et al. (2022) ♣	Ta-English and Malayalam-English code-mixed	MAML and Proto-MAML, based XML-R	2021
	Mozafari et al. (2022) ♣	Hate speech: En, Ar, Es, De, Id, It, Pt, Fr and Offensive lang. Ar, Da, En, El, Fa and Tr	MAML and Proto-MAML, based XML-R	2022
	Awal et al. (2023) ♣	En, Es, Ar, Da, El, Tr, Hi, De, It	HateMAML: domain-adaptive MAML based mBERT and XML-R	2023
Multitask-L - Joint training	Chiril et al. (2019) ◇	Fr and En	Based Bi-LSTM (Glove bilingual word embeddings)	2019
	Pamungkas and Patti (2019) ♣	En, It, Es, and De	Based MUSE	2019
	Pamungkas et al. (2021a) ♣	En, Fr, De, Id, It, Pt and Es	Based MUSE, LASER, mBERT	2021
Multitask-L - Auxiliary task	Riabi et al. (2022) ♣	En, It and Es	Based XML-R, XML-T	2022
	Montariol et al. (2022) ♣	En, It and Es	Based mBERT, XML-R, XML-T	2022
UL - GAE	De la Peña Sarracén and Rosso (2022) ◇	En, De, Ru, Tr, Hr and Sq	Based mBERT, XML-R (TFIDF)	2022
UL - Adversarial	Shi et al. (2022) ♣	En, Da, Ar, El and Tr	Based mBERT	2022

Abbreviation: LR:Logistic Regression, TL:Transfer Learning, MT:Machine Translation, EL:Ensemble Learning, Meta-L:Meta Learning, Multitask-L:Multitask Learning, UL:Unsupervised Learning, GAE:Graph Auto-Encoders. Languages abbreviations are based on ISO 639 language codes list.

◇/♣: ◇ refers to **Multilingual** methods & ♣ refers to **cross-lingual** methods. Feature extraction methods are put between ().

**Table 2. Overview of approaches on Multilingual and Cross-Lingual hate speech detection.**

(Note: only one representative study of each approach is cited in the table due to space limitation.)

257 Machine (SVM), as a baseline classifier, along with Bilingual Word Embeddings (BWE) to detect hate  
258 speech in English and German.

### 259 **Deep Neural Networks (DNN)**

260 Deep neural networks have been widely used for multilingual offensive language recognition because of  
261 their ability to acquire complex representations of text across different languages. There was an extensive  
262 use of Convolutional Neural Networks (CNNs) (Elouali et al. (2020); Bigoulaeva et al. (2023, 2021,  
263 2022)), CNN-GRU (Gated Recurrent Unit) (Deshpande et al. (2022a); Aluru et al. (2020)) and Recurrent  
264 Neural Networks (RNNs) like Long Short-Term Memory (LSTM), where Vashistha and Zubiaga (2021)  
265 used CNN-LSTM model, and Pamungkas et al. (2021a) utilized LSTM along with MUSE and mBERT.  
266 Adding to that, Bigoulaeva et al. (2022), and Vitiugin et al. (2021) used LSTM for word embeddings. As  
267 for Bidirectional LSTM (BiLSTM), it was implemented by Chiril et al. (2019); Bigoulaeva et al. (2023)  
268 and Bigoulaeva et al. (2021). Moreover, authors in Jiang and Zubiaga (2021) proposed a hate speech  
269 detection model called CCNL-Ex that includes additional hate-related semantic features. The model uses  
270 a Cross-lingual Capsule Network Learning approach CCNL with two parallel architectures for source  
271 and target languages. They used BiLSTM to extract contextual features, and Capsule Network to capture  
272 hierarchically positional relationships.

### 273 **Transfer Learning (TL)**

274 Transfer learning has emerged as an effective approach for identifying multilingual hate speech because  
275 it enables systems to use the information obtained from data in the source domain to perform better on  
276 data in other domains. As a result, many studies employed Pre-trained multilingual word embeddings  
277 like FastText (Bigoulaeva et al. (2021)), MUSE (Pamungkas and Patti (2019); Deshpande et al. (2022a);  
278 Aluru et al. (2020); Bigoulaeva et al. (2021)), or LASER (Deshpande et al. (2022a); Aluru et al. (2020);  
279 Pelicon et al. (2021a)), and (Vitiugin et al. (2021)). Moreover, most of the research studies has focused on  
280 the use of pre-trained language models LLMs (basically as classifiers): BERT (Vashistha and Zubiaga  
281 (2021); zahra El-Alami et al. (2022); Zia et al. (2022); Pamungkas et al. (2021a)), AraBERT (for Arabic  
282 data) (zahra El-Alami et al. (2022)), CseBERT (for English, Croatian and Slovenian data) (Pelicon et al.  
283 (2021b)), as well as multilingual BERT models: (Shi et al. (2022); Bhatia et al. (2021); Deshpande et al.  
284 (2022a); Aluru et al. (2020); zahra El-Alami et al. (2022); De la Peña Sarracén and Rosso (2022); Tita  
285 and Zubiaga (2021); Eronen et al. (2022); Ranasinghe and Zampieri (2021a); Ghadery and Moens (2020);  
286 Pelicon et al. (2021b); Awal et al. (2023); Montariol et al. (2022); Ahn et al. (2020a); Bigoulaeva et al.  
287 (2022, 2023); Pamungkas et al. (2021a); Pelicon et al. (2021a)), DistilmBERT model (Vitiugin et al.  
288 (2021)), and RoBERTa (Zia et al. (2022)).

289 On the other hand, cross-lingual language models like XLM were also widely employed, where we  
290 found implementation of XLM-RoBERTa (XLM-R) (Roy et al. (2021a); Bhatia et al. (2021); Wang et al.  
291 (2020); De la Peña Sarracén and Rosso (2022); Zia et al. (2022); Tita and Zubiaga (2021); Ranasinghe and  
292 Zampieri (2021b); Dadu and Pant (2020); Eronen et al. (2022); Ranasinghe and Zampieri (2021a, 2020);  
293 Mozafari et al. (2022); Barbieri et al. (2022); Awal et al. (2023); Stappen et al. (2020)), and both XLM-R  
294 and XLM-T (Montariol et al. (2022); Riabi et al. (2022)). These approaches have all been shown to  
295 improve performance on tasks involving multilingual/cross-lingual hate speech detection because they are  
296 more likely able to capture semantic and syntactic features across languages thanks to their pre-training  
297 on multilingual large volumes of texts. Therefore, transfer learning is expected to play an even greater  
298 part in improving the accuracy of multilingual hate speech detection algorithms.

### 299 **Machine Translation (MT)**

300 Using Machine Translation enables multilingual classification with monolingual models, where different  
301 languages are translated into the training language. Moreover, machine translation can be used as data  
302 augmentation to improve model performance. In this domain, Jiang and Zubiaga (2021) proposed a  
303 capsule network for cross-lingual hate speech detection. The network relies on source language and its  
304 translated counterpart in target language. Aluru et al. (2020) employed machine translation method for  
305 cross-lingual hate speech detection and compared the performance of LASER embedding and mBERT on  
306 datasets in 9 different languages. They found that simply adopting the machine translation method has  
307 comparative performance with multilingual models. Pamungkas et al. (2021a) proposed a joint-learning  
308 architecture utilizing multilingual language representations, and evaluated several competitive baseline  
309 systems including using machine translation to augment training data. The authors further investigated  
310 the impact of integrating a multilingual hate lexicon as an external source of knowledge into their joint-  
311 learning models. They found that a simple model relying on automatic machine translation and an English  
312 BERT pre-trained model achieved competitive results in their tasks. Ibrohim and Budi (2019b) discussed

313 the challenges of identifying hate speech in a multilingual setting and presented a comparison between  
314 two methods for multilingual text classification, translated and non-translated. The authors experimented  
315 with Support Vector Machine, Naive Bayes, and Random Forest Decision Tree classifiers with word  
316 n-grams and char n-grams as feature extraction. The experiment results suggested that the non-translated  
317 method performs better, but it is more costly due to data collection and annotation. On the other hand, the  
318 translated method without language identification gives poor results. To address this issue, the authors  
319 proposed combining the translated method with monolingual hate speech identification, which improved  
320 multilingual hate speech identification performance.

### 321 ***Ensemble Learning (EL)***

322 In multilingual and cross-lingual settings, ensemble learning has shown promise as an approach for  
323 increasing the performance of offensive language detection systems. Recent studies have shown that  
324 researchers used a variety of methods to use ensemble learning in this domain. For instance, Bigoulaeva  
325 et al. (2022) used a bootstrapping ensemble of several models for unlabeled German datasets and then  
326 fine-tuned English-trained models using this bootstrapped data. Also, Bigoulaeva et al. (2021) built a  
327 transferred system that used an ensemble-based approach to train on unlabeled data and included newly  
328 labeled data to improve performance on the target language. Adding to that, Bigoulaeva et al. (2023)  
329 provided a method for bootstrap labels using a variety of model structures and including unlabeled  
330 targeted language data for further advancements. Moreover, Ahn et al. (2020a) employed an ensembling  
331 procedure on multiple mBERT models to adjust hyperparameters (using the Translation Embedding  
332 Distance metric) and they improve the performance of both cross-lingual transfer and semi-supervised  
333 annotation labels. The model’s performance was improved compared to the baselines (which were  
334 trained only on manually annotated data) after using the semi-supervised dataset. Finally, Cohen et al.  
335 (2023) utilized DeBERTa-based ensemble learning method, including both back-translation and GPT-3  
336 augmentation.

### 337 ***Meta Learning (Meta-L)***

338 Meta-learning, also known as “learning to learn” is a burgeoning field of machine learning that is concerned  
339 with developing algorithms that enable an agent to learn how to learn. The objective of meta-learning  
340 is to design models that can rapidly adapt to new tasks with limited training data by leveraging prior  
341 experience. Meta-learning has shown significant promise in the field of NLP, where it can be used to  
342 improve the performance of language models by leveraging knowledge gained from solving one task to  
343 improve performance on another related task, which is particularly useful in scenarios where there is  
344 limited labeled data available for a given task. Current meta-learning methods include optimization-based  
345 methods and metric-based methods. Optimization-based methods aim to learn better model parameter  
346 initialization (Finn et al. (2017)), model architecture (Zoph and Le (2016)), or more efficient optimization  
347 strategy (Andrychowicz et al. (2016)). Metric-based methods aim to learn better distance metrics (Vinyals  
348 et al. (2016)) or representations (Snell et al. (2017)) to enable more efficient data contrastiveness in the  
349 metric space.

350 Compared to the more general NLP tasks, there are much fewer works focusing on the use of meta-  
351 learning in hate speech detection, especially in the cross-lingual setting. Vadakkekara Suresh et al.  
352 (2022) proposed a two-step strategy using meta-learning algorithms to identify offensive text in Tamil-  
353 English and Malayalam-English code-mixed texts. The authors introduced a weighted data sampling  
354 approach to enable better convergence in the meta-training phase compared to conventional methods.  
355 Their experimental results demonstrated that the meta-learning approach improves the performance of  
356 models significantly in low-resource (few-shot learning) tasks.

357 Mozafari et al. (2022) proposed a meta-learning-based approach to detect hate speech and offensive  
358 language in low-resource languages with limited labeled data. The methodology leverages two meta-  
359 learning models, MAML and Proto-MAML, to perform cross-lingual few-shot detection. The authors  
360 curated two diverse collections of publicly available datasets and compared the performance of their  
361 approach with transfer-learning-based models. The authors demonstrated by experiments that meta-  
362 learning-based models, particularly Proto-MAML, outperform transfer-learning-based models in most  
363 cases, and proposed that the meta-learning approach shows promise in identifying hateful or offensive  
364 content in low-resource languages with only a few labeled data points. Awal et al. (2023) proposed  
365 HateMAML to detect hate speech in low-resource languages. They proposed a self-supervision strategy to  
366 adapt the model to unseen target languages and evaluated the framework on five datasets across eight low-

367 resource languages, which showed that the proposed meta-learning method outperformed state-of-the-art  
368 baselines in cross-domain multilingual transfer settings.

### 369 ***Multitask Learning (Multitask-L)***

370 Using recent developments in NLP, multitask learning has come to be a promising strategy for enhancing  
371 multilingual and cross-lingual offensive language identification. This approach involves simultaneous  
372 training of models on multiple tasks in order to enhance their performance through the features sharing  
373 among the tasks data. Due to its capacity to enhance model performance by utilizing the knowledge  
374 gained from one or several tasks to improve the performance of another task, multitask learning has  
375 recently drawn more attention. Hate speech and offensive language detection models can benefit from this  
376 approach by training them on numerous related tasks such as sentiment analysis, dependency parsing, and  
377 named entity recognition, among others, allowing them to better detect offensive content. And by training  
378 models on multiple tasks in multiple languages, multitask learning can enable these systems to better  
379 understand the nuances of different languages and improve their performance in multilingual offensive  
380 language (Chen et al. (2021)).

381 In this context, Chiril et al. (2019) used joint training on both English and 30% of the French dataset,  
382 then they tested the trained model on the rest of French corpus. In addition, Pamungkas and Patti (2019)  
383 used a domain-independent lexicon called “HurtLex” (Bassignana et al. (2018)) in cross-domain and  
384 cross-language methods within a joint learning approach, which leads to improving the performance in  
385 detecting abusive content. Moreover, Montariol et al. (2022) studied used multilingual model pre-trained  
386 models (XLM-R and XLM-T) within a multitask architecture. They analyzed the impact of auxiliary tasks,  
387 like Name Entity Recognition NER, Part Of Speech POS tagging, dependency parsing, and sentiment  
388 analysis, and found that some hate speech labels were more susceptible to cross-lingual transfer learning.  
389 Furthermore, Riabi et al. (2022) used XLM-R and XLM-t based multitask learning models, as well as the  
390 MACHAMP strategy to fine-tune on various auxiliary tasks. Lastly, Pamungkas et al. (2021a) proposed  
391 two joint-learning approaches using diverse multilingual language features to transfer knowledge between  
392 pairs of languages. According to their study, their models performed the best across the board.

### 393 ***Unsupervised learning (UL)***

394 The issue of labeled data insufficiency and data annotation can be handled by using methods of unsuper-  
395 vised learning that do not rely on labeled training data to detect multilingual hate speech and offensive  
396 language. In fact, as indicated in De la Peña Sarracén and Rosso (2022), the authors use a label-free  
397 approach by encoding texts as graph nodes using Graph Auto-Encoders (GAE). This method utilizes a  
398 combination of transformers (mBERT and XLM-R) with convolutional neural layers, to encode the texts.  
399 Moreover, Shi et al. (2022), proposed an unsupervised model that employs cross-lingual mapping, sample  
400 generation, and transfer learning. Their model employs a novel training methodology that combines  
401 adversarial learning, transfer learning, and agreement regularization to detect offensive language in many  
402 low-resource languages. In multilingual and cross-lingual environments, each of the aforementioned  
403 methodologies suggest interesting new directions for unsupervised hate speech identification.

### 404 ***Other Technologies***

405 Several research studies have explored various advanced methods for offensive language detection and  
406 classification in different languages (especially in low-resource languages). In fact, a Transformer-based  
407 architecture called TIF-DNN was created for code-mixed Hindi and English (Biradar et al. (2021))  
408 using translation and transliteration techniques for hate speech detection. Furthermore, AraBERT and  
409 MarBERT based multi-task learning models were presented in Aldjanabi et al. (2021), these models  
410 perform offensive and hate speech detection tasks in modern standard Arabic language and in several  
411 dialects of Arabic tweets. Moreover, using polarity and emotions datasets, another multi-task learning  
412 technique Plaza-DeI-Arco et al. (2021) proved its success in the detection of hate speech in Spanish  
413 tweets. Additionally, BERT models were pre-trained on Hindi and Marathi: tweetsHindTweetBERT and  
414 MahaTweetBERT, illustrating state-of-the-art performance for hate speech detection in Gokhale et al.  
415 (2022).

416 **Since 2023**, there have been many new approaches presented in the field of multilingual hate speech de-  
417 tection, emphasizing the necessity of more learning and more use of the new technological advancements.  
418 In fact, Ghosal and Jain (2023) introduced a new unsupervised approach in Hindi and Bengali languages,  
419 incorporating detection of hateful content, classification of tweets, and preparation of code-switch data.

420 Furthermore, Goldzycher et al. (2023) utilizes intermediate English data fine-tuning along with Natural  
421 Language Inference (NLI) in Arabic, Portuguese, Spanish, and Italian hate speech detection. Adding  
422 to that, many BERT-based data augmentation methods are successfully incorporated to generate more  
423 data in various languages as illustrates in Takawane et al. (2023) where researchers, here, managed to  
424 enhance these models' performance on Code-Mixed Hindi-English hate speech data. Moreover, Kar and  
425 Debbarma (2023) explored a system using hybrid Diagonal Gated Recurrent Neural Networks DGRNN  
426 within an optimal feature extraction technique in multilingual code-mixed texts in English, Hindi, and  
427 German. Also, Das et al. (2023) emphasized both the strengths and limitations of the ChatGPT model  
428 in hate speech detection in eleven languages. Lastly, Roychowdhury and Gupta (2023) presented many  
429 data-efficient techniques like task reformulation and data augmentation in French, Spanish, Arabic, and  
430 Portuguese in hate speech detection.

### 431 **Analysis of the existing approaches**

432 Table 2 provides an overview of the previous research studies, we collected during our research analysis.  
433 It illustrate the different approaches implemented in multilingual and cross-lingual offensive language  
434 detection, we can see how these approaches have evolved over the years as well as the complexity, and  
435 the diversity of the languages studied in the field.

436 Around 2020, several approaches like Vashistha and Zubiaga (2021) and Aluru et al. (2020) con-  
437 centrated on word embeddings for feature representation and they managed to use small-sized datasets  
438 in different languages such as English, Hindi, and some code-mixed languages. There is a progression  
439 towards more complex deep learning architectures like CNN-LSTM in Vashistha and Zubiaga (2021), and  
440 diverse language sets in Aluru et al. (2020), that were using more low resource languages like Arabic,  
441 Indonesian, Italian, German, Portuguese, Polish, French, and Spanish. Later on, 2021 witnessed the use  
442 of more complex approaches in research in this domain, such as the use of bi-directional pre-trained  
443 transformers (like mBERT) and the use of XLM-R, which displays more direction toward implementing  
444 pre-trained language models ( especially the ones trained on significant volumes of multilingual data).  
445 More recently, since 2022, we observe more use of several sophisticated approaches, such as multitask  
446 learning, meta-learning, and ensemble learning. In particular, ensemble learning and meta-learning are  
447 gaining more attention, combining the predictions of multiple models or building and training a meta-  
448 learner able to adapt to new tasks with a few training examples, which could be mainly practical to use in  
449 the low-resource language data. Starting with the feature extraction techniques, word embeddings were  
450 among the widely used techniques, especially with the frequent implementation of several transformers  
451 like BERT and XLM-R, along with the use of other methods such as character-level representation and  
452 FastText but less often.

453 English language was the mostly learned compared to other languages, along with some other  
454 European languages such as Spanish, German, and French. Other languages were very little studied  
455 such as Japanese (Ja) and Norwegian (No), which reveals a serious challenge in analyzing offensive  
456 language and hate speech in these languages along with other non accessible ones (languages used by  
457 small communities that don't have ready data to work on). Therefore, we urge the need to conduct more  
458 research in order to be able to create performant services and tools for the detection of such content in  
459 these low-resource languages.

460 As shown in Table 2, the period of publication considered for the existing studies, ranges from  
461 2019 and till the time of this study (July 2023). This implies that researching in multilingual text  
462 classification task is still relatively new area of study, with a lot of ongoing research. Moreover, for the  
463 multilingual approaches, the analyses cover a wide scope of languages, such as English, Bengali, Arabic,  
464 Danish, Croatian, French and more. As for the cross-lingual approaches, the focus was mostly on fewer  
465 languages, typically, with English being the most studied as a source language (used for training), then  
466 some other rich-resource languages like Spanish, Italian, German, among others. Which indicates that  
467 multilingual text classification still requires much research, especially in low-resource languages. Lastly,  
468 we discovered an increasing tendency toward employing pretrained Large Language Models (LLMs) like  
469 BERT and XLM-R and their variations (mBERT, XLM-T, among others). These LLMs prove their ability  
470 to remarkably enhance the performance of multilingual hate speech detection tasks.

## 471 DATASETS ON MULTILINGUAL HATE SPEECH DETECTION

472 Most of the classification models for offensive language detection rely on supervised learning, therefore,  
473 access to high-quality and well-labeled data is necessary for training effective models. However, preparing  
474 such data is a very difficult and challenging task due to the enormous volume of information generated on  
475 social media platforms and other online sources. The task of curating and annotating this vast amount  
476 of data is both time-consuming and costly. In addition, a lot of effort is required to ensure the work is  
477 accurate and reliable. Thus, the process of data preparation continues to be a crucial and challenging  
478 aspect of training effective models.

479 The methodology of collecting the datasets analyzed in this section included first, selecting a set of  
480 chosen English keywords including terms like offensive language, hate speech, aggressive, multilingual,  
481 and low-resource datasets. Then utilizing these keywords we conducted searches on Google Scholar.  
482 Additionally, relevant workshops and shared tasks websites were also explored, as well as Hate Speech  
483 Dataset catalogue hatespeechdata, that presented many of these shared tasks datasets. The search process  
484 took place between March and April 2023. In the initial search round, we did not include any time or other  
485 filters and only considered the most relevant papers. Then to have also more recent datasets analyzed  
486 we applied a time filter to have more focus on the publications from 2020 to 2023. After the collection  
487 step, we created a table including the publication year, number of citations, and language(s) of each paper.  
488 We used this table to obtain and report the general statistics regarding the languages and citations of the  
489 datasets. Then due to time constraints, for more detailed analysis, we gave priority to papers representing  
490 datasets with more than one language and languages that had less than five datasets dedicated to them. For  
491 the remaining papers, we used the top ones with the most citations to ensure the inclusion of influential  
492 works. Furthermore, we also considered, in case of availability, the links provided in the papers, which  
493 were mostly from their Github repositories (URL to the repos are included in this study). This more  
494 detailed analysis led to the creation of Table 3 which we will explain in more detail.

495 Many of the analyzed hate speech datasets relied on Twitter as their primary data source. One  
496 key reason can be the availability of Twitter’s public API (Application Programming Interface). This  
497 API allows researchers to retrieve relevant tweets based on specific criteria and keywords, including  
498 those related to offensive content, events, and target groups. After Twitter, Facebook pages were the  
499 next prominent source of offensive language collected corpora. Offensive language datasets have also  
500 incorporated data from platforms like YouTube and Reddit, alongside various other sources and websites.

501 The collected datasets encompassed a variety of subjects and used different terms to describe the  
502 types of offensive content they gathered, highlighting the different aspects of negative language prevalent  
503 in online discourse. Many of the datasets specifically focused on hate speech, offensive language, and  
504 aggressiveness. Others explored misogyny, cyberbullying, abusive language, socially unacceptable dis-  
505 course, moderated news comments, stereotypes, among others. Datasets focusing on offensive content  
506 have also made efforts to encompass a diverse range of populations, taking into account various charac-  
507 teristics such as race, gender, ethnicity, religion, and sexual orientation. Women, individuals of African  
508 ancestry, LGBTQ+ individuals, immigrants, and members of various religious organizations, including  
509 Hindus, Christians, Jews, and Muslims are some of the highly targeted groups. By including such diverse  
510 populations, these datasets aim to provide a comprehensive understanding of offensive language and its  
511 impact on different communities.

512 Offensive language datasets come with diverse labeling schemes, capturing the multifaceted nature  
513 of the content analyzed. Table 4 provides an overview of datasets that encompass various types of  
514 labeling schemes, including binary labels, intensity levels, variations in hateful speech categorization, and  
515 multiple labels for themes and target groups. Most of the datasets rely on binary labels like hate/non-hate,  
516 offensive/non-offensive, aggressive/non-aggressive, among others. A number of datasets also annotated  
517 the levels of hate showing how weak or strong the hate and offensiveness is significantly providing a more  
518 detailed understanding of the intensity of harmful speech. Some datasets exhibit variations in labeling  
519 strategies particularly in distinguishing between different forms of negative speech. These variations can  
520 include the presence of distinct labels for hate speech, offensive language, abusive speech, among others.  
521 To address the context of hate speech, some datasets introduce multiple labels that capture themes and  
522 target groups of hate speech or whether they are aimed at individuals or groups.

523 Table 3 presents a comprehensive and structured analysis of hate speech detection datasets analyzed  
524 in this study. The table provides essential information for each dataset, including the year of publication,  
525 languages covered, and the main subject, encompassing hate speech, offensive language, aggressiveness,

Ref	Year	Language(s)	Main Subject	Source	Size	Cit.	Av.
Carvalho et al. (2022)	2022	Pt	HS	Twitter	63450	<10	N
Wang et al. (2022)	2022	Zh	HS	LINE Today	47844	<10	N
Ollagnier et al. (2022)	2022	Fr	Agr.	Aggressive multiparty chats collected through a role-playing game	19 conversations	<10	Y
Beyhan et al. (2022)	2022	Tr	HS	Twitter	IstanbulConv:1206 Refugee:1278	<10	Y
Madhu et al. (2023)	2023	Hi-En	HS and OL	Twitter	7088	<10	Y
Mohapatra et al. (2021)	2021	Or/Or-En	HS	Facebook	5000	<10	N
Steinberger et al. (2017)	2017	Cs, En, Fr, It, De	Flames	User-generated news article discussions	Cs:1812 De:1122, En:1007 Fr:487 It:649	<10	Y
Fernquist et al. (2019)	2019	Sv	HS	A Swedish discussion forum	3056	<10	N
Rahman et al. (2021)	2021	En	HS	Twitter	9667	<50	Y
Zampieri et al. (2022)	2022	Mr	OL	Twitter	MOLD2.0: 3611 SeMold: 8000	<50	Y
Nascimento et al. (2019)	2019	Pt-BR	OL	Twitter and Brazilian 55chan imageboard	7672	<50	Y
Akhtar et al. (2021)	2021	En	HS, Agr., OL, and Stereotype	Twitter	4480	<50	N
Mubarak et al. (2022)	2022	Ar	HS and OL	Twitter	12698	<50	N
Ombui et al. (2019)	2019	En, Sw, Other East African Languages	HS and OL	Twitter	260k	<50	N
Evkoski et al. (2022)	2022	Sl	HS	Twitter	12961136	<50	Y
Satapara et al. (2021)	2021	Hi-En	HS	Twitter	7088	<50	Y
Luu et al. (2021)	2021	Vi	HS and OL	Facebook and YouTube	33400	<50	Y
Fanton et al. (2021)	2021	En	HS / CN		5000 HS/CN pairs	<50	Y
Ali et al. (2022a)	2022	Ur	HS and OL	Twitter	10526	<50	N
Ljubešić et al. (2018)	2018	Sl, Hr	Moderated News Comments	The Slovene RTV MCC and Croatian 24sata News Portals	24639651	<50	Y(1) Y(2)
Vu et al. (2020)	2020	Vi	HS and OL	Facebook	5431	<50	Y
Ptaszynski et al. (2019)	2019	Pl	HS and Cyberbullying	Twitter	11041	<50	Y
Gaikwad et al. (2021)	2021	Mr	OL	Twitter	MOLD 1.0: 2499	<50	Y
Haddad et al. (2019)	2019	Tunisian Ar	HS and Abusive	Different social media platforms	6075	<50	Y
Das et al. (2021)	2021	Bn	HS	Facebook	7425	<50	N
Rizwan et al. (2020)	2020	Roman Ur	HS	Twitter	10012	<50	Y
Guest et al. (2021)	2021	En	Misogyny	Reddit	6567	<50	Y
Leite et al. (2020)	2020	Pt-BR	Toxic Speech	Twitter	21K	<50	Y
Ishmam and Sharmin (2019)	2019	Bn	HS	Facebook	5126	<50	Y
Moon et al. (2020)	2020	Ko	Toxic Speech (HS and OL)	A popular domestic entertainment news aggregation platform	9381	<100	Y
Mandl et al. (2021)	2021	En, Hi, Mr	HS and OL	Twitter	En:3843 Mr:1874 Hi:4594	<100	Y
de Pelle and Moreira (2017)	2016	Pt-BR	OL	Brazilian Web (g1.globo.com)	OFFCOMBR-2: 1250, OFFCOMBR-3: 1033	<100	Y
Fortuna et al. (2019)	2019	Pt	HS	Twitter	5668	<100	Y
Álvarez-Carmona et al. (2018)	2018	Es-MX	Agr.	Twitter	10856	<100	Y
Fišer et al. (2017)	2017	Sl	Socially Unacceptable Discourse	Spletno Oko1 (Web Eye) hotline service	13000	<100	N
Mossie and Wang (2020)	2020	Am	HS and Vulnerable Community	Facebook	491424	<100	N
Kumar et al. (2020)	2020	Bn, Hi, En	Agr.	YouTube	Approx. 6000 per lang.	<100	Y
Ibrohim and Budi (2018)	2018	Id	HS	Twitter	2016	<100	Y
Mulki et al. (2019)	2019	Levantine Ar	HS and Abusive	Twitter	5846	<150	N
Çöltekin (2020)	2020	Tr	OL	Twitter	36232	<150	Y
Mathur et al. (2018)	2018	Hi-En	HS and OL	Twitter	3679	<150	N
Pitenis et al. (2020)	2020	El	OL	Twitter	OGTD 1.0: 4779, OGTD 2.0: 10287	<150	Y
Chung et al. (2019)	2019	En, Fr, It	HS / CN	Generated by experts	4078 HS/CN pairs	<150	Y
Sigurbjergsson and Derczynski (2019)	2019	Da	HS and OL	Reddit and Facebook	3600	<150	Y
Pavlopoulos et al. (2017)	2017	El	User Comment Moderation	A Greek news portal ( <a href="http://www.gazzetta.gr/">http://www.gazzetta.gr/</a> )	Approx. 1.6M	<150	Y
Ibrohim and Budi (2019a)	2019	Id	HS and Abusive	Twitter	13169	<150	Y
Pereira-Kohatsu et al. (2019)	2019	Es	HS	Twitter	6000	<150	Y
Kumar et al. (2018b)	2018	Hi-En	Agr.	Facebook and Twitter	Approx. 18k tweets and 21k Facebook comments	<150	N
Alfina et al. (2017)	2017	Id	HS	Twitter	520	<200	Y
Ousidhoum et al. (2019)	2019	En, Fr, Ar	HS	Twitter	En:5647 Fr:4014 Ar:3353	<200	Y
Bohra et al. (2018)	2018	Hi-En	HS	Twitter	4575	<200	Y
Sanguinetti et al. (2018)	2018	It	HS	Twitter	6009	<200	Y
Fersini et al. (2018)	2018	Es, En	Misogyny	Twitter	En:3977 Es:4138	<250	Y
Mandl et al. (2019)	2019	En, Hi, De	HS and OL	Twitter and Facebook	En:5852, Hi:4665, De:3819	<350	Y
Zampieri et al. (2020)	2020	Ar, Da, En, El, Tr	OL	Twitter, Facebook, Reddit, a local newspaper: Ekstra Bladet	En:1448861, Ar:1589, Da:384, El:2486, Tr:6131	<400	Y
Del Vigna12 et al. (2017)	2017	It	HS	Facebook	?	<400	N
Basile et al. (2019)	2019	Es, En	HS	Twitter	En:13000, Es:6600	<750	Y

**Table 3. A summary of the available datasets for hate speech detection.**

Abbreviations notes: For the column names, Ref.:Reference, Cit.:Citation by May-2023, and Av.:Available. Language names have been shortened using the ISO 639-1 standardized nomenclature. Under the “Main Subject” column: HS:Hate Speech, OL:Offensive Language, CN:Counter-Narrative, and Agr.:Aggressiveness. In “Size” column: Approx.:Approximately. In “Av.” column: Y:Yes and N:No. *Note:* Links to the resources may not be shown in the hard copy.

Dataset Labeling Schemes	Datasets
Only Binary Labels	Zampieri et al. (2020); Bohra et al. (2018); Alfina et al. (2017); Pitenis et al. (2020); Álvarez-Carmona et al. (2018); Pereira-Kohatsu et al. (2019); Pavlopoulos et al. (2017); Ptaszynski et al. (2019); de Pelle and Moreira (2017); Evkoski et al. (2022); Steinberger et al. (2017); Rahman et al. (2021); Nascimento et al. (2019); Wang et al. (2022); Yang et al. (2022); Ranasinghe et al. (2022); Madhu et al. (2023); Satapara et al. (2021); Aliyu et al. (2022); Gaikwad et al. (2021)
Contains Intensity Levels	Del Vigna12 et al. (2017); Sanguinetti et al. (2018); Ibrohim and Budi (2019a); Kumar et al. (2020)
Contains Different Categorizations of Negative Speech	Sanguinetti et al. (2018); Mandl et al. (2019); Ousidhoum et al. (2019); Ibrohim and Budi (2019a); Mulki et al. (2019); Basile et al. (2019); Ibrohim and Budi (2018); Fersini et al. (2018); Mandl et al. (2021); Haddad et al. (2019); Moon et al. (2020); Mathur et al. (2018); Vu et al. (2020); Luu et al. (2021); Ombui et al. (2019); Das et al. (2021); Mohapatra et al. (2021); Beyhan et al. (2022); Ali et al. (2022a); Fernquist et al. (2019); Mubarak et al. (2022); Mazari and Kheddar (2023); Akhtar et al. (2021); Rizwan et al. (2020)
Contains Themes/Target Groups	Del Vigna12 et al. (2017); Ibrohim and Budi (2019a); Sigurbergsson and Derczynski (2019); Çöltekin (2020); Basile et al. (2019); Mossie and Wang (2020); Fortuna et al. (2019); Kumar et al. (2018b); Fersini et al. (2018); Ishmam and Sharmin (2019); Das et al. (2021); Beyhan et al. (2022); Guest et al. (2021); Carvalho et al. (2022); Yadav et al. (2023); Akram et al. (2023); Zampieri et al. (2022); Rizwan et al. (2020)

**Table 4. Type of available labels in the studied Datasets.**

526 and more. Additionally, the table includes details on the dataset source, indicating the platform from  
527 which the data was collected, such as Twitter or Facebook. Dataset size is also included, representing  
528 the number of samples within each dataset, while the citation count provides a measure of the usage and  
529 recognition of the datasets. Furthermore, the table indicates the availability of each dataset along with  
530 the link where they can be found, the publicly accessible ones are marked as “Y” with a hyperlink, and  
531 those that are not are marked as “N”. This comprehensive overview serves as a valuable resource, offering  
532 researchers a consolidated reference for hate speech detection datasets, their attributes, and accessibility.

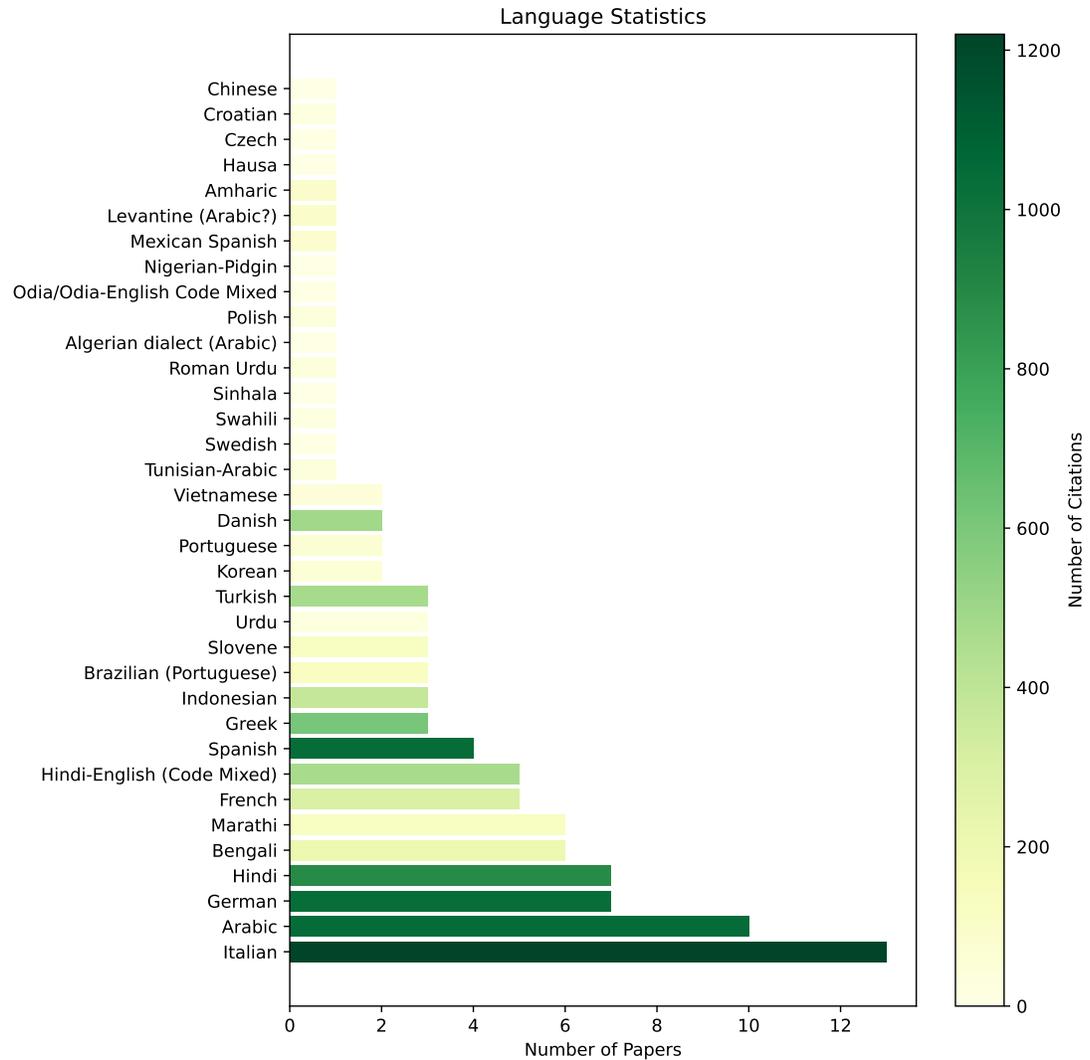
533 We constructed an informative figure to depict the availability of datasets for different languages  
534 and the corresponding citations, enabling us to assess the corpora distribution and utilization patterns  
535 across various languages. Fig. 2 showcases the languages for which datasets are available, alongside the  
536 number of citations received by each dataset’s paper. This visual representation offers valuable insights  
537 into the high and low-resource languages, emphasizing the significance of both categories in research  
538 endeavors. It allows us to identify languages that receive substantial attention and recognition, regardless  
539 of their resource availability. The figure underscores the importance of supporting research efforts for  
540 low-resource languages, as their impact and usage transcend their limited resources. The horizontal bar  
541 chart displays the languages on the y-axis, while the x-axis represents the number of papers available for  
542 each language. Each bar’s color is determined by the corresponding number of citations received (by the  
543 time of this study: May 2023), with a color bar provided to indicate the intensity of citation impact. Please  
544 note that English has been intentionally omitted from the chart to maintain visual clarity. With 47 papers  
545 and approximately 11,000 citations, English’s strong presence would have overshadowed the statistics of  
546 other languages, limiting the informative value of the figure. During the analysis of the collected datasets,  
547 it was observed that English, Italian, and Arabic were the most prevalent languages, with a relatively high  
548 number of dedicated datasets available for each. Conversely, several languages were represented by only a  
549 single dataset. Notably, some languages such as Spanish, Greek, Indonesian, Turkish, and Danish, despite  
550 having more limited dedicated datasets, received a considerable number of citations. This observation  
551 points to a noteworthy level of interest and utilization within the research community for these languages.  
552 It emphasizes the importance and necessity of addressing the needs of low-resource languages.

## 553 RESOURCES FOR MULTILINGUAL HATE SPEECH DETECTION

554 In this section, we aim to explore the existing resources on the multilingual offensive language field, we  
555 divide this section into two major parts: available collaborative projects, and related products (e.g. open  
556 community challenges, source codes, and APIs). We are strengthening these resources’ influence by  
557 bringing them to the awareness of a wide audience, including developers, academics, students, among  
558 others, who can employ these tools to work on offensive language detection in multiple languages.

### 559 Collaborative International Projects

560 A summary of the projects we are mentioning in this section is presented briefly in Table 5.



**Figure 2. Datasets distribution based on languages.** Number of papers studies in this survey and their impact (based on citations @May 2023).  
*Note:* English (47 papers with 11k citations) is excluded from the figure to have a better visualization in the distribution.

561 **European project:** Several European projects have been undertaken to address the detection and  
 562 mitigation of online hate speech. The *DARE* project, <https://cordis.europa.eu/project/id/725349> (2017-  
 563 2021), funded under the EU Horizon 2020 Framework Programme, aims to develop technologies and  
 564 methodologies for combating hate speech, including radicalization and extremist content. It involves  
 565 partners from 13 countries, including Belgium, Croatia, Germany, Greece, France, Malta, Poland, Russia,  
 566 Turkey, Tunisia, The Netherlands, and the UK. Another project, *Hatemeter* (2018-2020), focuses on  
 567 monitoring, analyzing, and tackling anti-Muslim hatred online at the EU level. It takes a multidimensional  
 568 approach to identify red flags of hate speech, understand patterns of Islamophobia, develop tactical and  
 569 strategic responses, and produce counter-narratives. The project partners include Italy, France, and the  
 570 UK. In the realm of cybersecurity, the *PANACEA* (2019-2022) project aims to improve cybersecurity and  
 571 privacy/data protection in hospital and health infrastructures. It provides toolkits to enhance security and  
 572 data protection for various stakeholders in the healthcare sector, including hospitals, software/system  
 573 developers, medical device manufacturers, and digital service providers. The project involves partners  
 574 from Italy, UK, Greece, France, Belgium, Netherlands, Germany, and Ireland. Moreover, the *sCAN* project  
 575 (2018-2020), coordinated by LICRA (International League against Racism and Antisemitism), focuses

Project	Partners	Year
DARE: Dialogue About Radicalisation and Equality	Belgium, Croatia, France, Germany, Greece, Malta, Norway, Poland, Russian Federation, The Netherlands, Tunisia, Turkey and the UK	2017-2020
MANDOLA: Monitoring and Detecting OnLine Hate Speech	Greece, Ireland, France, Spain, Bulgaria and Cyprus	2017
PRO2HATERS: PROactive PROfiling of HATE speech spreadeRs	Germany	2017
Hatometer: Hate speech tool for monitoring, analyzing and tackling Anti-Muslim hatred online	Italy, France and the UK	2018-2020
sCAN: specialised Cyber-Activists Network	France, Germany, Italy, Belgium, Czech Republic, Austria, Slovenia, Croatia and Latvia	2018-2020
PANACEA: Protection and privAcY of hospital and health iN-frastructures with smArt Cyber sEcurity and cyber threat toolkit for dAta and people	Italy, UK, Greece, France, Belgium, Netherlands, Germany and Ireland	2019-2022
DTCT: Detect Then Act	Belgium, Germany, the UK and the Netherlands	2019-2021
Stand By Me	Italy, Poland and Hungary	2020
EOOH: The European Observatory of Online Hate	Belgium, Slovakia and the Netherlands	2021
Identricis	Bulgaria	2023
OHI: Online Hate Index	USA	Released in 2018
ProPublica's Documenting Hate Project	USA	Started in 2017

**Table 5. Collaborative International Projects on Hate speech.**

576 on gathering expertise, tools, methodology, and knowledge to identify, analyze, report, and counteract  
577 online hate speech. It involves partners from France, Germany, Italy, Belgium, Czech Republic, Austria,  
578 Slovenia, Croatia, and Latvia.

579 While explicit information about language support may not be mentioned for some of the mentioned  
580 projects above, it is reasonable to assume that their deliverables and solutions would likely focus on  
581 the languages spoken in the countries of their partner organizations. Given the diverse range of partner  
582 countries involved in these projects, it is possible that they would consider the languages relevant to those  
583 countries. This would imply that their solutions and products could potentially support languages beyond  
584 English, depending on the specific project's objectives and target regions.

585 Other projects are explicitly mentioning their focus on various languages. Among them, we found  
586 *Detect Then Act (DTCT)* (2019-2021), a European collaboration that aims to monitor and tackle online  
587 hate speech. It utilizes Explainable AI to assist users in deflating toxic discussions. The project reports  
588 illegal hate speech cases in accordance with the EU's Code of Conduct and local legislation. Moreover, it  
589 provides master training for hate speech detection in multiple languages, including English, French, Dutch,  
590 German, and Hungarian. The project partners involved are from Belgium, Germany, the UK, and the  
591 Netherlands. The project *Stand By Me* (2020) was created to moderate online violence against women in  
592 Europe. The project aims to help addressing this issue by enhancing individuals' awareness and capability  
593 to recognize such content. The project utilizes a diverse approach, including a combined learning program  
594 and educational resources. In addition, The European Observatory of Online Hate (EOOH) project (2021)  
595 was released by Textgain (in the lead), as a Multi-platform for monitoring hate speech in more than 20  
596 social media platforms and covering 24 different languages. This project has made significant progress  
597 in understanding the complexities of online hate speech. This incorporates comprehending its various  
598 forms, the relationships among corresponding users, and the strategies involving disinformation. Lastly,  
599 the *MANDOLA* project (2017) aims to improve our understanding of online hate speech prevalence and  
600 empower ordinary citizens to report it. It utilizes big-data approaches to monitor this content, provide  
601 policymakers with actionable information, and transfer best practices among Member States. The project  
602 partners include Greece, Ireland, France, Spain, Bulgaria, and Cyprus. While the specific languages  
603 supported by the project may not be explicitly mentioned in the provided information, the project's  
604 objectives display that it seeks to monitor and analyze hate-related speech across multiple languages,  
605 including languages other than English.

606 On the other hand, several "industrial projects" aim to tackle the issue of hate speech and polarization  
607 in society. One such project is *PRO2HATERS: PROactive PROfiling of HATE speech spreadeRs* (2017)  
608 by Symanto, (in Germany). *PRO2HATERS* focuses on addressing hate speech and polarization, with a  
609 particular emphasis on languages beyond English, such as German and Spanish, and their dialects. The  
610 project proposes language resources, network analysis, methods, and tools as key components to combat  
611 hate speech. It envisions various application scenarios, including cyber-security, where government

612 agencies can detect and counter hate speech, and social media companies can automate hate speech  
613 detection. Another project in this domain is *Identric*s (2023, Bulgaria). Identric offers a cutting-edge  
614 Hate Speech Detection service that helps eliminate hate speech in the comments sections of websites.  
615 Leveraging machine learning models, this project continuously learns to identify and flag hate speech,  
616 providing alerts to potential occurrences. By utilizing Identric's service, platforms can foster meaningful  
617 conversations without the concern of offensive or abusive language spreading throughout their community.

618 **Non-European projects:** Fighting hate speech is undoubtedly an international effort that cuts across  
619 national boundaries. The multifaceted projects use a wide range of tactics, including constructing  
620 machine learning models, crafting legislation, starting public education programs, and starting awareness  
621 campaigns. Despite the complexity of the problem, these international initiatives show a shared dedication  
622 to creating safer and more inclusive online and physical settings. The *Online Hate Index* (OHI), a tool  
623 employing machine learning to identify and quantify hate speech targeting marginalized groups on digital  
624 platforms in English, it was created by the Anti-Defamation League ADL's Center for Technology and  
625 Society (CTS). The program is made to identify linguistic trends and continuously advance in antisemitic  
626 content detection, giving an objective way to gauge the incidence of hate speech and assess the success  
627 of digital businesses' anti-hate measures. This project is made by the USA and released in 2018 (the  
628 project was developed domestically but has international implications to be used worldwide). Moreover,  
629 *ProPublica's Documenting Hate Project*, a well-known American endeavor that was started in 2017,  
630 aimed to compile an extensive database of hate crimes committed throughout the nation. The project  
631 teamed up with newsrooms, educational institutions, and independent journalists to assist in reporting and  
632 documenting instances of bias and hatred in response to the dearth of accurate statistics on hate crimes.

### 633 Available Products

634 We examine, in this part, various facets of the resources available in hate speech detection, more specifi-  
635 cally in multilingual hate speech detection. We'll start by highlighting the community challenges and  
636 competitions. We will next move on to talking about the accessible open-source codes. These represent  
637 concrete instruments that are open to learning and discovering the developed solutions. In order to wrap  
638 off this analysis, we will look at the APIs, including multilingual APIs.

639 **Community Challenges & Datasets Provided:** Detecting multilingual hate speech and offensive  
640 language is a paramount challenge, especially for social media platforms where a myriad of cultures and  
641 languages interact daily. This complexity arises due to the nuanced, context-specific, and often indirect  
642 nature of hate speech and offensive language. Over the years, several competitions and hackathons  
643 have been aimed at addressing this issue. In 2018, the *TRAC-1: Aggression Identification* (Kumar  
644 et al. (2018a)), the first Workshop on Trolling, Aggression, and Cyberbullying, honed in on identifying  
645 aggression in social media posts in English and Hindi, both in Roman and Devanagari scripts. Then,  
646 2019 saw many important challenges, such as: the *SemEval-2019 Task 5: Multilingual Detection of*  
647 *Hate Speech Against Immigrants and Women in Twitter* (Basile et al. (2019)). It centered around hate  
648 speech detection in a multilingual setting, focusing on English and Spanish tweets. In 2020, several key  
649 events occurred. *Kaggle's Jigsaw Multilingual Toxic Comment Classification competition* encouraged  
650 participants to build models that identify rudeness, disrespect, or any conversation-derailing toxicity  
651 in multilingual online discussions using English-only training data. The same year, the *Hate Speech*  
652 *Detection (HASOC) Competition*, hosted by FIRE (Forum for Information Retrieval Evaluation), focused  
653 on identifying hate speech and offensive content in English, German, and Hindi. Additionally, the *TRAC*  
654 *2020*, the second Workshop on Trolling, Aggression, and Cyberbullying, provided two shared tasks,  
655 one on Aggression Identification and another on Misogynistic Aggression Identification in Bangla (in  
656 both Roman and Bangla script), Hindi (in both Roman and Devanagari script) and English. Adding to  
657 that, in 2021, two significant challenges emerged: the *Kaggle IIT-D Multilingual Abusive Comment*  
658 *Identification* focused on identifying abusive comments across various Indic languages, and the *PAN*  
659 *shared task of Profiling Hate Speech Spreaders on Twitter 2021* involved profiling hate speech spreaders  
660 on Twitter in English and Spanish. Apart from these targeted events, numerous other hackathons and  
661 competitions contribute indirectly to the field of multilingual hate speech detection. Events centered  
662 around cross-lingual or multilingual text classification, sentiment analysis, or broad NLP problems offer  
663 valuable platforms for devising innovative solutions to detect hate speech and offensive language across  
664 languages.

665 **Datasets provided of the community challenges:** Overall, the above-mentioned challenges are pre-

666 sented in Table 6, they were instrumental in providing a broad spectrum of datasets that cater to different  
 667 languages and aspects of hate speech detection. They are sources of rich and diverse information, accessi-  
 668 ble to researchers worldwide. By compiling and making these datasets available, they have fundamentally  
 669 contributed to the field. These datasets can be accessed by registering or filling out the appropriate forms.  
 670 One prominent resource comes from the SemEval 2019 Task 5, a Shared Task focused on the Multilingual  
 671 Detection of Hate. Furthermore, the IIT-D Multilingual Abusive Comment Identification challenge has  
 672 provided a dataset unique in its capacity for massively multilingual abusive comment identification across  
 673 a variety of Indic languages. Another dataset worth mentioning revolves around Profiling Hate Speech  
 674 Spreaders on Twitter, and is available in both English and Spanish. Complementing this, the HASOC  
 675 2020 challenge provides a dataset for Hate Speech and Offensive Content Identification. Additionally, the  
 676 dataset from the TRAC - 2020, caters to Bangla, Hindi, and English.

Challenge Name	Languages	Year
TRAC-1: Aggression Identification	English, Hindi (Roman and Devanagari scripts)	2018
SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter	English, Spanish	2019
Kaggle’s Jigsaw Multilingual Toxic Comment Classification	Multilingual (trained on English data)	2020
Hate Speech Detection (HASOC) Competition	English, German, Hindi	2020
TRAC 2020: Trolling, Aggression, and Cyberbullying	English, Hindi (Roman and Devanagari scripts), Bangla (Roman and Bangla scripts)	2020
Kaggle IIT-D Multilingual Abusive Comment Identification	Multiple Indic languages	2021
PAN shared task of Profiling Hate Speech Spreaders on Twitter	English, Spanish	2021

**Table 6. Summary of Community Challenges on Multilingual Hate Speech Detection.**

677 **Available Source Codes:** Table 7 provides a general overview of different solutions that have been  
 678 developed in multilingual hate speech and offensive language detection. The source codes presented here  
 679 are across 2020 to 2023, dealing with a variety of languages. Recently, Cohen et al. (2023) offered a  
 680 source code along with live demonstrations to execute it. This code especially helps to further study and  
 681 implement ensemble models based on RoBERTa or DeBERTa. It also gives a practical tool for researchers  
 682 studying back translation and GPT-3 data augmentation techniques. Also, Deshpande et al. (2022b)  
 683 provided the source code of a model for detecting hate speech across ten languages. Another Github  
 684 project is the “Multilingual-Abuse-Comment-Detection”, which focuses on identifying abusive comments  
 685 in seventeen Indian languages using MuRIL-based models (BERT based Multilingual Representations for  
 686 Indian Languages). Moreover, Röttger et al. (2021) evaluated the performance of hate speech detection  
 687 models across ten languages, giving code to test across multilingual provided datasets. Adding to  
 688 that, Aluru et al. (2021) worked across nine languages on sixteen datasets for the classification of hate  
 689 speech data, presenting a source code to train and fine-tune several models, including mBERT-based,  
 690 Translation+BERT, CNN+GRU and LASER+LR. Besides, each of Aluru et al. (2021) and Sharif et al.  
 691 (2021) gave solutions for offensive language detection in three different Dravidian languages. In 2020,  
 692 the “Detoxify” project by Hanu and Unitary team (2020), established on three Jigsaw challenges, studied  
 693 multilingual toxic comment classification across seventeen languages using XLM-R based models, as well  
 694 as Ahn et al. (2020b) which deals with offensive language detection in five languages, within Semeval  
 695 2020 task, being among the first ten places in each of Greek, Danish, and Turkish languages datasets.  
 696 Overall, these source codes describe the recent research actions toward producing more practical and  
 697 effective solutions for multilingual hate speech and offensive language detection. They underscore the  
 698 increasing direction toward low-resource languages. Even though some of these codes didn’t have research  
 699 study associated, they present detailed descriptions of their source codes in Github repositories.

700 **Available APIs:** The landscape of hate speech detection is rich with an array of tools that harness  
 701 the power of AI to identify and counteract such harmful discourse. A comprehensive collection of  
 702 tools and services designed to counteract and analyze hate speech is available online. Among the many  
 703 prominent tools, illustrated in Table 8, we have the *HateLab*, an international center for studying hate  
 704 speech founded by the Economic and Social Research Council (ESRC), the *RapidAPI Hate Speech*  
 705 *Detection* that enables effective detection of offensive language, and *iSpotHate*, a freely available API  
 706 dedicated to eradicating hate speech. Moreover, the Python library *HateSonar* (2020) offers simple and  
 707 efficient hate speech detection without any need for user training, and *Profanity-check* (2019), another one,  
 708 swiftly checks for profanity or offensive language in strings. Furthermore, *StopPropagHate* by INESC  
 709 TEC, utilizes machine learning techniques to help news organizations automatically identify hate speech.  
 710 *Cohere* offers a text moderation API that can efficiently filter out harmful or inappropriate content

Name	Languages	Link	Study if available	Year
Enhancing social network hate detection using back translation and GPT-3 augmentations during training and test-time	En, Fr, De, Es and No	code	Cohen et al. (2023)	2023
Highly Generalizable Models for Multilingual Hate Speech Detection	En, Ar, De, Id, It, Pt, Es, Fr, Tr, Da and Hi	code	Deshpande et al. (2022b)	2022
Multilingual-Abuse-Comment-Detection	17 Hi languages	code	Non available	2022
HateCheck: Functional Tests for Hate Speech Detection Models	Ar, NI, Fr, De, Hi, It, Zh, Pl, Pt and Es	code	Röttger et al. (2021)	2022
Deep Learning Models for Multilingual Hate Speech Detection	Ar, En, De, Id, It, Pl, Pt, Es and Fr	code	Aluru et al. (2021)	2021
EACL 2021 OffensEval in Dravidian Languages	Kn, MI and Ta	code	Jayanthi and Gupta (2021)	2021
Leveraging Multilingual Transformers for Hate Speech Detection	En, De and Hi	code	Roy et al. (2021b)	2021
Offensive Language Detection from Multilingual Code-Mixed Text using Transformers	Kn, MI and Ta	code	Sharif et al. (2021)	2021
Multi-oli	Da, Ko and En	code	The language-adversarial training pipeline inspired from Keung et al. (2019)	2021
Indonesian Text Classification Multilingual	En and Id	code	Putra and Purwarianti (2020)	2021
Detoxify: Toxic Comment Classification with Pytorch Lightning and Transformers	En, Fr, Es, It, Pt, Tr and Ru	code	Non available	2020
NLPDove at SemEval-2020 Task 12: Improving Offensive Language Detection with Cross-lingual Transfer	En, El, Da, Ar and Tr	code	Ahn et al. (2020b)	2020
Multilingual Fairness LREC	En, It, Pl, Pt, Es	code	Huang et al. (2020)	2020

**Table 7. Github Repositories for Multilingual source code projects of Hate Speech.**

711 in real-time, while *Hive.ai* is a high-speed content moderation API with extensive training data (with  
712 results returned in under 200ms). Additionally, *MODERATION API* can detect and hide a wide range of  
713 data entities, including sensitive information and inappropriate content. Similarly, *Openai* contributes to  
714 comprehend linguistic context, precisely identifying subtle instances of abusive language, including hate  
715 speech, cyberbullying, and content that promotes self-harm, as well as detecting probable instances of  
716 misinformation or disinformation.

API Name	Description	Multilingual Support
HateLab	International center for studying hate speech	–
RapidAPI Hate Speech Detection	Detection of offensive language	–
iSpotHate	Detection and elimination of hate speech	–
HateSonar	Python library - Hate speech detection	–
Profanity-check	Python library - Profanity detection	–
StopPropagHate	Hate speech detection and prediction of a news potential to provoke such comments.	–
Cohere	Filter out harmful or inappropriate content	–
Hive.ai	High-speed content moderation API	–
MODERATION API	Detect and hide sensitive and inappropriate content	–
Openai	Detection of abusive language, hate speech, and misinformation.	–
Sightengine	Detection of hateful, sexual and toxic content	English, Chinese, Dutch, French, German, Portuguese, Italian, Swedish, Spanish, Tagalog/Filipino and Turkish.
Spectrum Labs' Guardian	Employs Natural Language Understanding AI to detect harmful behaviors.	Arabic, French, Hindi, Korean, and many others
Microsoft Azure	Detection of profanity	Over 110 languages
Membrane	Filter out various types of hateful content	English, Spanish, German, French, Polish, Turkish, Dutch, Italian, Swedish, Arabic, Chinese, Portuguese, Japanese, and Russian.
Alibaba Cloud's Text Moderation 2.0	Content review, and custom configurations	Up to 20 languages
Huawei Cloud	Content moderation API	Chinese

**Table 8. Summary of APIs and their Language Capabilities in Hate Speech Detection.**

717 **Multilingual APIs:** The relevance of multilingual tools in hate speech detection is paramount, as they  
718 aid in breaking language barriers to ensure the internet remains a safe space for all. Among these powerful  
719 tools, *Sightengine* stands out with its capability to detect hateful, sexual and toxic content across multiple  
720 languages, which include not just English, but also Chinese, French, Italian, Dutch, German, Portuguese,  
721 Swedish, Turkey, Filipino, and Spanish. Similarly, *Spectrum Labs' Guardian* elevates the standard of

722 multilingual content moderation. Unlike conventional tools that largely depend on keyword-based filters,  
723 Guardian employs true Natural Language Understanding (NLU) AI. This advanced technique allows  
724 the system to detect harmful behaviors, such as bullying, hate speech, spam, extremism, among others,  
725 across languages including Arabic, French, Hindi, Korean, among many others. *Microsoft Azure* is an  
726 API developed within Cognitive Services, it helps in detecting profanity in more than 110 languages.  
727 Also, *Membrane* can filter out various types of content, including spam, clickbait, offensiveness, among  
728 others. It covers multiple languages: English, German, French, Polish, Spanish, Turkish, Italian, Dutch,  
729 Swedish, Chinese, Arabic, Chinese, Russian and Japanese. Moreover, *Alibaba Cloud's Text Moderation*  
730 *2.0 API*, which supports up to 20 languages, offers a potent suite of features. These include content review,  
731 and custom configurations. Adding to that, *Huawei Cloud* contributes to this language-inclusive trend  
732 with its content moderation API. Although it currently supports only Chinese, its presence underlines  
733 the importance of multilingual tools and the continuous strides being made towards expanding language  
734 support in the field of hate speech detection.

## 735 CHALLENGES AND LIMITATIONS

736 In this section first, we review the main challenges that have been faced during the detection of offensive  
737 language in NLP, with a focus on the challenges in multilingual and cross-lingual corresponding tasks.  
738 Next, we examine limitations, and lastly, we propose some future directions.

### 739 Challenges

#### 740 *Technical Challenges*

741 **Lack of labeled data across languages:** One of the main issues lies in the scarcity of annotated data,  
742 or the non-accessible ones (non-public ones). Thus, getting this data proves to be a difficult and time-  
743 consuming task. The restricted availability of such data, and in certain languages acts as a significant  
744 obstacle. Building highly accurate and performant models requires often a large amount of annotated  
745 data, especially in the target languages. Nevertheless, the process of data annotation is still challenging; it  
746 is costly, time-consuming, and requires a lot of experts in the domain to do the job (Kovács et al. (2021);  
747 Röttger et al. (2022a)), especially with the granularity of this content (Vidgen et al. (2019)). Moreover, the  
748 problem of imbalanced datasets still persists, usually making the offensive labeled data in all its categories  
749 a minor class. Therefore, traditional machine learning techniques often perform badly on these minority  
750 class samples, especially in binary datasets. Researchers have suggested a number of oversampling and  
751 undersampling strategies to solve this issue as in Khairy et al. (2023).

752 Adding to that, even non-English datasets present a substantial challenge, as there are still limited  
753 annotated datasets in the domain. For example, a dataset could include tweets in Persian and Arabic  
754 while creating a dataset of hate speech in Urdu Ali et al. (2022b). As a result, taking these challenges into  
755 consideration, it becomes obvious that the creation of annotated datasets in multiple languages remains an  
756 important step for advancing approaches, especially in low-resource languages like Arabic (Omar et al.  
757 (2020)), among others.

758 **Cross-lingual Transfer Learning:** Applying knowledge learned from one language to another,  
759 is considered a difficult task. Despite the recent significant progress to build accurate pre-trained  
760 multilingual language models, their cross-lingual ability for offensive language detection remains limited.  
761 This limitation is evident when working on swear words of specific cultures, which often vary among  
762 languages, and cannot even be easily translatable with the current machine translation tools. For instance,  
763 researchers have found important linguistic problems when employing Google Translate in their models,  
764 more specifically, they identified errors (Pamungkas and Patti (2019)). Another crucial problem is the  
765 unstable performance of some approaches across distinct target languages. In fact, Glavaš et al. (2020)  
766 indicates that rich-resource languages manage to give better results compared to low-resource ones.

767 **Language and Topic Inequality:** The dominance of the English language in the current datasets  
768 has led to another significant challenge, such as anglophone bias outcomes in non-English data. This  
769 issue affects prominent companies such as Facebook, whose capacities were limited in 2020 to detect  
770 hate speech in Spanish, and Mandarin (Aluru et al. (2020)). Adding to that, datasets in other languages  
771 are not only insufficient, but also tend to be small-sized, restricting the performance of offensive content  
772 detection in these languages (Aluru et al. (2020)) which explains the restricted number of studies in these  
773 low-resource languages, like the Arabic language Khairy et al. (2021). Another crucial factor is the  
774 dynamics of language and topic, as some datasets just cover one topic (misogyny, racism, among others)

775 in many languages (Arango et al. (2019)), which adds more complexity to the generalisability of this  
776 detection task in multilingualism.

777 **Bias:** Bias can appear during data collection, labeling, or training. It is one of the main problems  
778 that makes it hard to identify offensive language in multiple languages. A number of biases were found:  
779 racial bias, author bias, and subject bias. However, topic bias is still the most important one, as shown  
780 by some studies in this field (Arango et al. (2019)). As one of the vital solutions to this issue, many  
781 studies have introduced several functional tests, such as Röttger et al. (2022b), which have presented  
782 functional tests for hate speech detection models, introducing Multilingual HateCheck (MHC). Their work  
783 offered a various set of tests across ten different languages and aimed to improve the assessment of hate  
784 speech detection models, revealing crucial weaknesses in both monolingual and cross-lingual applications.  
785 Another crucial solution that was introduced in the detection of offensive and abusive language in Dutch  
786 is Caselli and Van Der Veen (2023), which is a comprehensive study of fine-tuned models. The study  
787 also examines the use of data cartography to determine high-quality training data. These two mentioned  
788 studies are not only restricted to solving data bias issues, but also to evaluate pre-trained language models  
789 and LLMs, and to identify precisely the quality of datasets.

790 **Hallucination of LLMs:** Bang et al. (2023) shows that multitasking, multimodal, and multilingual  
791 use cases have profited from the usage of Large Language Models (LLMs), such as ChatGPT. Yet, in the  
792 multilingual domain, they often have issues with hallucinations. For example, the user confidence rate can  
793 be decreased by low-performant translation tools, resulting in safety problems (Guerreiro et al. (2023)).

#### 794 ***Non-Technical Challenges***

795 **Language Cultures and Dialects:** Offensive language tasks can be embedded in cultural issues. Any  
796 cultural background could impact whether a word or expression is considered offensive or not (Schmidt  
797 and Wiegand (2017)), thus, even utilizing the same language, this content and the capacity of offensiveness  
798 could be varied among regions and populations. Another factor is to consider the language’s various  
799 dialects. For instance, the Arabic language is associated with lots of different dialects utilized by Arabic  
800 speakers on Twitter. As a result, learning and comprehending Arabic is a difficult task, especially for  
801 offensive language detection (Al-Hassan and Al-Dossari (2019)).

802 **Definition of hate speech:** As described in Section “Definition of Hate Speech”, various jurisdictions  
803 give different definitions of offensive language, thus, resulting in a non-standard general definition. This  
804 becomes more complex in multilingual scenarios considering the cultural aspects and dialects.

805 **Annotation problem of ‘foreign language effect’:** The “foreign language effect” is one of the major  
806 issues of offensive data labeling, it yields people (annotators) to adopt different moral stances and usually  
807 consider this content to be less harsh in their second languages, thus affecting the multilingual annotation  
808 stability of this content. This has been studied in Abercrombie et al. (2023), which finds out a lower  
809 annotation agreement on hateful English and German labeling tasks.

#### 810 **Limitations**

811 **Computational Limitations:** Multilingual text classification task requires extensive computational  
812 resources due to large data volume from many languages and the complexity of the models employed  
813 (since multilingual models are usually bigger in size with more weights). In fact, multilingual embeddings  
814 or pre-trained transformers (mBERT, XLM-R, mT5, among others.) require more computing resources.  
815 Moreover, the process of fine-tuning these models usually needs extensive training. While some resources  
816 provided, like Google Colab, VastAI, and cloud platforms such as AWS, Google Cloud, Microsoft Azure,  
817 and Baidu offer essential computational resources, they may be costly, especially for users and researchers  
818 with limited budgets. For further details, a brief description of the resource providers, mentioned above, is  
819 presented in Table 9, we illustrated the most affordable offers delivered.

820 **Multilingual Pre-trained Large Language models (Multilingual LLMs):** Implementing and  
821 training pre-trained language models is not an easy task due to their limitations (Nozza (2021)). An  
822 example of these crucial limitations, presented by Conneau et al. (2020), is the “curse of multilinguality”.  
823 They indicate that training multilingual LLMs in more languages shows declines in performance despite  
824 keeping the number of update steps. Furthermore, performance usually declines when supporting more  
825 languages and providing optimal performance on a more limited language set. This ‘curse’ basically  
826 involves determining whether to work on a small number of languages for more accurate performance or  
827 to distribute resources across multiple languages but with reduced performance (Pfeiffer et al. (2022)).

Platform	Resources (The cheapest offers)	Limitations
Google Colab	Free tier provides an NVIDIA Tesla K80 GPU with 12 GB of RAM.	The session length is capped at 12 hours. After this, all data will be deleted, including any trained models unless they've been saved elsewhere.
VastAI	Depending on demand, one can rent an NVIDIA GTX 1080 Ti with 11 GB of GPU memory for as low as around \$0.10/hour.	Although cost-effective, the availability of cheap resources is highly dependent on demand and can be unreliable.
AWS	The EC2 Spot Instances allow for cheap access to powerful resources. For instance, a g4dn.xlarge instance with an NVIDIA T4 GPU (16 GB of GPU memory) can be rented for around \$0.30/hour, but the exact rate varies.	Spot Instances can be interrupted by AWS with a 2-minute notification. They are best for flexible applications that aren't sensitive to sudden interruptions.
Google Cloud	Preemptible VMs provide affordable access to powerful resources. For example, a preemptible instance with an NVIDIA Tesla T4 GPU can be rented for approximately \$0.30/hour, but the exact rate varies.	Preemptible VMs can be stopped by Google at any time if resources are required elsewhere, and they automatically shut down after 24 hours.
Microsoft Azure	Azure Spot Instances provide cheaper access to resources. An example is the Standard_NV4as_v4 Spot instance with a portion (1/8) of an NVIDIA Tesla V100 GPU available for approximately \$0.17/hour, though exact rates vary.	Like other spot or preemptible instances, Azure Spot Instances can be interrupted by Microsoft at any time if the resources are required elsewhere.
Baidu Cloud Compute (BCC)	Offers an array of hardware resources, like the NVIDIA deep learning development card and NVIDIA Tesla K40 as cost-effective GPU for beginners and those with lower training requirements. It also offers discounts that can be checked directly in the website.	Potential language barriers given Baidu's primary focus on the Chinese market.

**Table 9. Comparative Analysis of Affordable Computational Resources for Machine Learning Training.**

828 **Limitations on Machine translation tools:** The performance of offensive language detection models  
829 can be impacted by the quality and precision of the machine translation tools. Although multilingual  
830 Neural Machine Translation (multilingual NMT) displays significant performance, the degree to which  
831 it can handle many languages remains limited (Aharoni et al. (2019)). Recently, these tools have made  
832 important results in bilingual translation (Cho et al. (2014); Vaswani et al. (2017)). However, there remain  
833 considerable barriers when it comes to implementing NMT in low-resource languages (Dabre et al. (2020);  
834 Wang et al. (2021)). Therefore, recent research studies have emphasized enclosing many translation data  
835 inside a single model to improve their performance (Aharoni et al. (2019)). However, it has been observed  
836 that these models frequently give low performance compared to the bilingual ones (Arivazhagan et al.  
837 (2019); Pham et al. (2019)). Moreover, the majority of earlier studies have been focused on English  
838 language translation, which lead to non-English ones to be low performing (Aharoni et al. (2019); Zhang  
839 et al. (2020)).

#### 840 **Future Directions**

841 The field of multilingual and cross-lingual offensive language detection offers many promising recom-  
842 mendations for future research. especially in low resource languages, as well as in different topics of  
843 offensive language. For instance, since social media users generate one-third of the poor-quality Arabic  
844 content, Koshiry et al. (2023) built and annotated a standardized toxic Arabic dataset from Twitter, which  
845 would facilitate and improve toxicity analysis in Arabic language.

#### 846 **on DataSet**

847 Future studies could concentrate on developing more diverse, and balanced datasets in multiple languages  
848 and dialects, as well as in different topics of offensive language.

849 **Generating Data:** With the problem of data scarcity, especially in multilingual settings, some  
850 research studies are directed into providing more efficient solutions for data augmentation, by leveraging  
851 generated samples in order to gradually train their detection models and enhance the performance of their  
852 classification capabilities. Several approaches have been already released on English samples, that may  
853 be used to work on generating multilingual data, using different methods like adversarial auto-regressive  
854 models (Ocampo et al. (2023)), generative GPT3 PLM-based models (Hartvigsen et al. (2022)), or  
855 generative GPT-Neo based model (Muti et al. (2023)).

856 **External Features:** Multiple research studies have highlighted the incorporation of features extracted  
857 from domain-agnostic or language-independent resources in cross-lingual aspects (also in cross-domain).  
858 Moreover, certain studies have underlined the vital role that emotional information has in detecting  
859 offensive language (Rajamanickam et al. (2020); Safi Samghabadi et al. (2020)). Therefore, it would  
860 be helpful to examine the inclusion of emotional or sentiment data for boosting knowledge transfer.

861 Similarly, a study performed by Pamungkas et al. (2021a) confirmed the effectiveness of external features  
862 extracted from the multilingual lexicon HurtLex. They highlight its importance in assisting the knowledge  
863 transfer process in the detection of multilingual offensive content especially when dealing with metaphors,  
864 metonymy, among others, as well as non-formal expressions that are highly sensitive to geographical, and  
865 cultural deviations.

866 **Advanced Annotation:** Using Generative Pretrained Transformer models could increase training  
867 data (data augmentation), and despite showing promising performance in generating data in high-resource  
868 languages, this still requires to be enhanced more to generate data in low-resource languages (Ahuja et al.  
869 (2023)). Additionally, semi-supervised learning and unsupervised methods could effectively use both  
870 labeled and unlabeled data, and reduce the challenge of annotated data scarcity.

### 871 ***on Modeling***

872 Besides working on data, future studies should also focus on how to get benefit from the new innovative  
873 methods, in order to build more effective models in the field. For instance, using Federated Learning for  
874 decentralized and privacy-preserving training may assist in understanding local dialects and slang (Weller  
875 et al. (2022)). Adding to that, Explainable AI (XAI) could guarantee more transparent model decisions,  
876 promoting trust in their classification performance (Kumar et al. (2021)). Moreover, Reinforcement  
877 learning can make models determine optimal measures for the classification task (Fang et al. (2017)),  
878 and active learning can enable demanding labels for the most informative instances (Hajmohammadi  
879 et al. (2015)). Another future aspect to be considered is the “*Teacher and Student*”, it’s used to transfer  
880 knowledge from LLMs to create smaller pre-trained models. For example, the study Ranasinghe and  
881 Zampieri (2023), worked on creating lightweight offensive language models (with fewer numbers of  
882 parameters and with less computational consumption resources), which can be among the initial steps to  
883 create multilingual models specialized more in this domain. Besides machine learning field, quantum  
884 computing has also proved to be a competitive method, faster and promising high performance in low  
885 resource languages like Arabic (Omar and Abd El-Hafeez (2023))

886 **Study the Impact of New Generative Models:** Future studies could explore the impact of the new  
887 generative models, such as GPT-3, GPT-4, and ChatGPT, in order to improve multilingual offensive  
888 language detection, using their ability in cross/multi-lingual understanding. They can also handle data  
889 scarcity problems by generating synthetic data in low-resource languages, much more similar to human-  
890 written data. For example, Hartvigsen et al. (2022) released ‘ToxiGen’: an English machine-generated  
891 dataset, that could be a start to create datasets in other languages in the field.

### 892 ***on Low Resource Languages***

893 There is an increasing necessity to focus on low-resource languages, ensuring more general language  
894 coverage worldwide. For example, there are multiple research models released that worked on African  
895 languages such as Wolof and Swahili (Jacobs et al. (2023)).

## 896 **CONCLUSION**

897 While monolingual resources and approaches are important, the significance of multilingual efforts are  
898 highly crucial. Multilingual solutions not only broaden the scope of understanding but also enable the  
899 development of more robust models. By addressing various languages simultaneously, we can bridge  
900 communication gaps and cultural diversity. In fact, leveraging multilingual resources promotes innovation  
901 and technological advancements, leading to more effective and universally applicable solutions. Moreover,  
902 encouraging multilingualism efforts in offensive language detection enables greater understanding, and  
903 effectiveness in safeguarding online communication. In this survey, we conduct a thorough investigation of  
904 multilingual offensive language identification in fast-globalizing social media platforms where hundreds  
905 of languages and dialects are used to communicate. Our work is motivated by the difficulty of detecting  
906 offensive content within the increasing use of non-English and low-resource languages. Our study draws  
907 inspiration from previous surveys in the field, outlining the gaps we managed to address. Specifically, our  
908 survey distinguishes itself by comprehensively presenting both multilingual and cross-lingual offensive  
909 language detection approaches, organizing findings across various machine learning classes, ranging from  
910 traditional to more advanced approaches. This inclusive strategy aims to offer readers a comprehensive  
911 understanding of existing approaches while encouraging for the adoption of more progressive techniques,  
912 detailed in the ‘Other Technologies’ and ‘Future Directions’ subsections. Moreover, a crucial aspect that

distinguished our research is the expansive coverage of resources and tools. We prioritize the presentation of datasets, more specifically enclosing a significant number of corpora within the field, and covering a greater number of low-resource languages. We also tried to give a wider view of the other resources, getting deeply into the projects, source codes, APIs, among others. Finally, our study underlines critical challenges in the multilingual landscape of offensive language detection, attributing limitations to these issues and providing clear solutions to be considered as future directions. Overall, our survey aims to serve as a comprehensive guideline for both industry and academic practitioners, offering a significant and rich understandings into various aspects of multilingual offensive language detection while advocating for progressive advancements in the field.

## REFERENCES

- Abercrombie, G., Hovy, D., and Prabhakaran, V. (2023). Temporal and second language influence on intra-annotator agreement and stability in hate speech labelling. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 96–103, Toronto, Canada. Association for Computational Linguistics.
- Aharoni, R., Johnson, M., and Firat, O. (2019). Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.
- Ahn, H., Sun, J., Park, C. Y., and Seo, J. (2020a). NLPDove at SemEval-2020 task 12: Improving offensive language detection with cross-lingual transfer. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1576–1586, Barcelona (online). International Committee for Computational Linguistics.
- Ahn, H., Sun, J., Park, C. Y., and Seo, J. (2020b). Nlpdove at semeval-2020 task 12: Improving offensive language detection with cross-lingual transfer. *CoRR*, abs/2008.01354.
- Ahuja, K., Diddee, H., Hada, R., Ochieng, M., Ramesh, K., Jain, P., Nambi, A., Ganu, T., Segal, S., Axmed, M., Bali, K., and Sitaram, S. (2023). Mega: Multilingual evaluation of generative ai.
- Akhtar, S., Basile, V., and Patti, V. (2021). Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *arXiv preprint arXiv:2106.15896*.
- Akram, M. H., Shahzad, K., and Bashir, M. (2023). Ise-hate: A benchmark corpus for inter-faith, sectarian, and ethnic hatred detection on social media in urdu. *Information Processing & Management*, 60(3):103270.
- Al-Hassan, A. and Al-Dossari, H. (2019). Detection of hate speech in social networks: a survey on multilingual corpus. In *6th international conference on computer science and information technology*, volume 10, pages 10–5121.
- Aldjanabi, W., Dahou, A., Al-qaness, M. A. A., Elaziz, M. A., Helmi, A. M., and Damaševičius, R. (2021). Arabic offensive and hate speech detection using a cross-corpora multi-task learning model. *Informatics*, 8(4).
- Alfina, I., Mulia, R., Fanany, M. I., and 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)*, pages 233–238.
- Ali, R., Farooq, U., Arshad, U., Shahzad, W., and Beg, M. O. (2022a). Hate speech detection on twitter using transfer learning. *Computer Speech & Language*, 74:101365.
- Ali, R., Farooq, U., Arshad, U., Shahzad, W., and Beg, M. O. (2022b). Hate speech detection on twitter using transfer learning. *Computer Speech Language*, 74:101365.
- Aliyu, S. M., Wajiga, G. M., Murtala, M., Muhammad, S. H., Abdulmumin, I., and Ahmad, I. S. (2022). Herdphobia: A dataset for hate speech against fulani in nigeria. *arXiv preprint arXiv:2211.15262*.
- Aluru, S. S., Mathew, B., Saha, P., and Mukherjee, A. (2020). Deep learning models for multilingual hate speech detection. *CoRR*, abs/2004.06465.
- Aluru, S. S., Mathew, B., Saha, P., and Mukherjee, A. (2021). A deep dive into multilingual hate speech classification. In Dong, Y., Ifrim, G., Mladenčić, D., Saunders, C., and Van Hoecke, S., editors, *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track*, pages 423–439, Cham. Springer International Publishing.
- Álvarez-Carmona, M. Á., Guzmán-Falcón, E., Montes-y Gómez, M., Escalante, H. J., Villasenor-Pineda, L., Reyes-Meza, V., and Rico-Sulayes, A. (2018). Overview of mex-a3t at ibereval 2018: Authorship

967 and aggressiveness analysis in mexican spanish tweets. In *Notebook papers of 3rd sepln workshop on*  
968 *evaluation of human language technologies for iberian languages (ibereval), seville, spain*, volume 6.

969 Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M. W., Pfau, D., Schaul, T., Shillingford, B., and  
970 De Freitas, N. (2016). Learning to learn by gradient descent by gradient descent. *Advances in neural*  
971 *information processing systems*, 29.

972 Arango, A., Pérez, J., and Poblete, B. (2019). Hate speech detection is not as easy as you may think: A  
973 closer look at model validation. *Proceedings of the 42nd International ACM SIGIR Conference on*  
974 *Research and Development in Information Retrieval*.

975 Arivazhagan, N., Bapna, A., Firat, O., Aharoni, R., Johnson, M., and Macherey, W. (2019). The missing  
976 ingredient in zero-shot neural machine translation. *CoRR*, abs/1903.07091.

977 Awal, M. R., Lee, R. K.-W., Tanwar, E., Garg, T., and Chakraborty, T. (2023). Model-agnostic meta-  
978 learning for multilingual hate speech detection.

979 Bakalis, C. and Hornle, J. (2021). The role of social media companies in the regulation of online hate  
980 speech. In *Studies in Law, Politics, and Society*. Emerald Publishing Limited.

981 Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do,  
982 Q. V., Xu, Y., and Fung, P. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on  
983 reasoning, hallucination, and interactivity.

984 Barbieri, F., Espinosa Anke, L., and Camacho-Collados, J. (2022). XLM-T: Multilingual language models  
985 in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and*  
986 *Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

987 Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., and Sanguinetti, M.  
988 (2019). SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women  
989 in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63,  
990 Minneapolis, Minnesota, USA. Association for Computational Linguistics.

991 Bassignana, E., Basile, V., and Patti, V. (2018). Hurltex: A multilingual lexicon of words to hurt. In  
992 *CEUR Workshop proceedings*, volume 2253, pages 1–6. CEUR-WS.

993 Beyhan, F., Çarık, B., Arın, İ., Terzioğlu, A., Yanikoglu, B., and Yeniterzi, R. (2022). A turkish hate  
994 speech dataset and detection system. In *Proceedings of the Thirteenth Language Resources and*  
995 *Evaluation Conference*, pages 4177–4185.

996 Bhatia, M., Bhotia, T. S., Agarwal, A., Ramesh, P., Gupta, S., Shridhar, K., Laumann, F., and Dash, A.  
997 (2021). One to rule them all: Towards joint indic language hate speech detection. In *Fire*.

998 Bigoulaeva, I., Hangya, V., and Fraser, A. (2021). Cross-lingual transfer learning for hate speech detection.  
999 In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*,  
1000 pages 15–25, Kyiv. Association for Computational Linguistics.

1001 Bigoulaeva, I., Hangya, V., Gurevych, I., and Fraser, A. (2022). Addressing the challenges of cross-lingual  
1002 hate speech detection. *CoRR*, abs/2201.05922.

1003 Bigoulaeva, I., Hangya, V., Gurevych, I., and Fraser, A. (2023). Label modification and bootstrapping for  
1004 zero-shot cross-lingual hate speech detection. *Language Resources and Evaluation*, pages 1–32.

1005 Biradar, S., Saumya, S., and Chauhan, A. (2021). Hate or non-hate: Translation based hate speech  
1006 identification in code-mixed hinglish data set. In *2021 IEEE International Conference on Big Data*  
1007 *(Big Data)*, pages 2470–2475.

1008 Bohra, A., Vijay, D., Singh, V., Akhtar, S. S., and Shrivastava, M. (2018). A dataset of hindi-english  
1009 code-mixed social media text for hate speech detection. In *Proceedings of the second workshop on*  
1010 *computational modeling of people’s opinions, personality, and emotions in social media*, pages 36–41.

1011 Carvalho, P., Cunha, B., Santos, R., Batista, F., and Ribeiro, R. (2022). Hate speech dynamics against  
1012 african descent, roma and lgbtqi communities in portugal. In *Proceedings of the Thirteenth Language*  
1013 *Resources and Evaluation Conference*, pages 2362–2370.

1014 Caselli, T. and Van Der Veen, H. (2023). Benchmarking offensive and abusive language in Dutch tweets.  
1015 In Chung, Y.-l., Rottger, P., Nozza, D., Talat, Z., and Mostafazadeh Davani, A., editors, *The 7th*  
1016 *Workshop on Online Abuse and Harms (WOAH)*, pages 69–84, Toronto, Canada. Association for  
1017 Computational Linguistics.

1018 Chen, S., Zhang, Y., and Yang, Q. (2021). Multi-task learning in natural language processing: An  
1019 overview. *CoRR*, abs/2109.09138.

1020 Chhabra, A. and Vishwakarma, D. K. (2023). A literature survey on multimodal and multilingual  
1021 automatic hate speech identification. *Multimedia Systems*, pages 1–28.

- 1022 Chiril, P., Benamara Zitoune, F., Moriceau, V., Coulomb-Gully, M., and Kumar, A. (2019). Multilingual  
1023 and multitarget hate speech detection in tweets. In *Actes de la Conférence sur le Traitement Automatique*  
1024 *des Langues Naturelles (TALN) PFIA 2019. Volume II : Articles courts*, pages 351–360, Toulouse,  
1025 France. ATALA.
- 1026 Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y.  
1027 (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation.  
1028 In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*  
1029 *(EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.
- 1030 Chung, Y.-L., Kuzmenko, E., Tekiroglu, S. S., and Guerini, M. (2019). Conan–counter narratives  
1031 through nichesourcing: a multilingual dataset of responses to fight online hate speech. *arXiv preprint*  
1032 *arXiv:1910.03270*.
- 1033 Cohen, S., Presil, D., Katz, O., Arbili, O., Messica, S., and Rokach, L. (2023). Enhancing social network  
1034 hate detection using back translation and gpt-3 augmentations during training and test-time. *Information*  
1035 *Fusion*, 99:101887.
- 1036 Çöltekin, Ç. (2020). A corpus of turkish offensive language on social media. In *Proceedings of the*  
1037 *Twelfth Language Resources and Evaluation Conference*, pages 6174–6184.
- 1038 Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M.,  
1039 Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale.  
1040 In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages  
1041 8440–8451, Online. Association for Computational Linguistics.
- 1042 Dabre, R., Chu, C., and Kunchukuttan, A. (2020). A survey of multilingual neural machine translation.  
1043 *ACM Comput. Surv.*, 53(5).
- 1044 Dadu, T. and Pant, K. (2020). Team rouges at SemEval-2020 task 12: Cross-lingual inductive transfer to  
1045 detect offensive language. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages  
1046 2183–2189, Barcelona (online). International Committee for Computational Linguistics.
- 1047 Das, A. K., Al Asif, A., Paul, A., and Hossain, M. N. (2021). Bangla hate speech detection on social  
1048 media using attention-based recurrent neural network. *Journal of Intelligent Systems*, 30(1):578–591.
- 1049 Das, M., Pandey, S. K., and Mukherjee, A. (2023). Evaluating chatgpt’s performance for multilingual and  
1050 emoji-based hate speech detection.
- 1051 Davidson, T., Bhattacharya, D., and Weber, I. (2019). Racial bias in hate speech and abusive language  
1052 detection datasets. *arXiv preprint arXiv:1905.12516*.
- 1053 De la Peña Sarracén, G. L. and Rosso, P. (2022). Unsupervised embeddings with graph auto-encoders  
1054 for multi-domain and multilingual hate speech detection. In *Proceedings of the Thirteenth Language*  
1055 *Resources and Evaluation Conference*, pages 2196–2204, Marseille, France. European Language  
1056 Resources Association.
- 1057 de Pelle, R. P. and Moreira, V. P. (2017). Offensive comments in the brazilian web: a dataset and baseline  
1058 results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*. SBC.
- 1059 Del Vigna<sup>12</sup>, F., Cimino<sup>23</sup>, A., Dell’Orletta, F., Petrocchi, M., and Tesconi, M. (2017). Hate me, hate me  
1060 not: Hate speech detection on facebook. In *Proceedings of the first Italian conference on cybersecurity*  
1061 *(ITASEC17)*, pages 86–95.
- 1062 Deshpande, N., Farris, N., and Kumar, V. (2022a). Highly generalizable models for multilingual hate  
1063 speech detection. *ArXiv*, abs/2201.11294.
- 1064 Deshpande, N., Farris, N., and Kumar, V. (2022b). Highly generalizable models for multilingual hate  
1065 speech detection. *CoRR*, abs/2201.11294.
- 1066 d’Sa, A. G., Illina, I., and Fohr, D. (2020). Bert and fasttext embeddings for automatic detection of  
1067 toxic speech. In *2020 International Multi-Conference on: “Organization of Knowledge and Advanced*  
1068 *Technologies”(OCTA)*, pages 1–5. IEEE.
- 1069 Elouali, A., Elberrichi, Z., and Elouali, N. (2020). Hate speech detection on multilingual twitter using  
1070 convolutional neural networks. *Rev. d’Intelligence Artif.*, 34(1):81–88.
- 1071 Eronen, J., Ptaszynski, M., Masui, F., Arata, M., Leliwa, G., and Wroczynski, M. (2022). Transfer  
1072 language selection for zero-shot cross-lingual abusive language detection. *Information Processing*  
1073 *Management*, 59(4):102981.
- 1074 Evkoski, B., Pelicon, A., Mozetič, I., Ljubešić, N., and Kralj Novak, P. (2022). Retweet communities  
1075 reveal the main sources of hate speech. *Plos one*, 17(3):e0265602.
- 1076 Fang, M., Li, Y., and Cohn, T. (2017). Learning how to active learn: A deep reinforcement learning ap-

- 1077       proach. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*,  
1078       pages 595–605, Copenhagen, Denmark. Association for Computational Linguistics.
- 1079 Fanton, M., Bonaldi, H., Tekiroglu, S. S., and Guerini, M. (2021). Human-in-the-loop for data collection:  
1080       a multi-target counter narrative dataset to fight online hate speech. *arXiv preprint arXiv:2107.08720*.
- 1081 Fernquist, J., Lindholm, O., Kaati, L., and Akrami, N. (2019). A study on the feasibility to detect hate  
1082       speech in swedish. In *2019 IEEE international conference on big data (Big Data)*, pages 4724–4729.  
1083       IEEE.
- 1084 Fersini, E., Rosso, P., and Anzovino, M. (2018). Overview of the task on automatic misogyny identification  
1085       at ibereval 2018. *Iberval@ sepln*, 2150:214–228.
- 1086 Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep  
1087       networks. In *International conference on machine learning*, pages 1126–1135. PMLR.
- 1088 Fišer, D., Erjavec, T., and Ljubešić, N. (2017). Legal framework, dataset and annotation schema for  
1089       socially unacceptable online discourse practices in slovene. In *Proceedings of the first workshop on*  
1090       *abusive language online*, pages 46–51.
- 1091 Fortuna, P., da Silva, J. R., Wanner, L., and Nunes, S. (2019). A hierarchically-labeled portuguese hate  
1092       speech dataset. In *Proceedings of the third workshop on abusive language online*, pages 94–104.
- 1093 Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Comput.*  
1094       *Surv.*, 51(4).
- 1095 Fortuna, P., Soler, J., and Wanner, L. (2020). Toxic, hateful, offensive or abusive? what are we really  
1096       classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th language*  
1097       *resources and evaluation conference*, pages 6786–6794.
- 1098 Gaikwad, S. S., Ranasinghe, T., Zampieri, M., and Homan, C. (2021). Cross-lingual offensive language  
1099       identification for low resource languages: The case of Marathi. In *Proceedings of the International*  
1100       *Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 437–443, Held  
1101       Online. INCOMA Ltd.
- 1102 Ghadery, E. and Moens, M.-F. (2020). LIIR at SemEval-2020 task 12: A cross-lingual augmentation  
1103       approach for multilingual offensive language identification. In *Proceedings of the Fourteenth Work-*  
1104       *shop on Semantic Evaluation*, pages 2073–2079, Barcelona (online). International Committee for  
1105       Computational Linguistics.
- 1106 Ghosal, S. and Jain, A. (2023). Hatecircle and unsupervised hate speech detection incorporating emotion  
1107       and contextual semantics. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(4).
- 1108 Glavaš, G., Karan, M., and Vulić, I. (2020). XHate-999: Analyzing and detecting abusive language  
1109       across domains and languages. In *Proceedings of the 28th International Conference on Computational*  
1110       *Linguistics*, pages 6350–6365, Barcelona, Spain (Online). International Committee on Computational  
1111       Linguistics.
- 1112 Gokhale, O., Kane, A., Patankar, S., Chavan, T., and Joshi, R. (2022). Spread love not hate: Undermining  
1113       the importance of hateful pre-training for hate speech detection.
- 1114 Goldzycher, J., Preisig, M., Amrhein, C., and Schneider, G. (2023). Evaluating the effectiveness of natural  
1115       language inference for hate speech detection in languages with limited labeled data.
- 1116 Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. <http://www.deeplearningbook.org>.
- 1117
- 1118 Guerreiro, N. M., Alves, D., Waldendorf, J., Haddow, B., Birch, A., Colombo, P., and Martins, A. F. T.  
1119       (2023). Hallucinations in large multilingual translation models.
- 1120 Guest, E., Vidgen, B., Mittos, A., Sastry, N., Tyson, G., and Margetts, H. (2021). An expert annotated  
1121       dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European*  
1122       *Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350.
- 1123 Haddad, H., Mulki, H., and Oueslati, A. (2019). T-hsab: A tunisian hate speech and abusive dataset. In  
1124       *Arabic Language Processing: From Theory to Practice: 7th International Conference, ICALP 2019,*  
1125       *Nancy, France, October 16–17, 2019, Proceedings 7*, pages 251–263. Springer.
- 1126 Hajmohammadi, M. S., Ibrahim, R., Selamat, A., and Fujita, H. (2015). Combination of active learning  
1127       and self-training for cross-lingual sentiment classification with density analysis of unlabelled samples.  
1128       *Information Sciences*, 317:67–77.
- 1129 Hammersley, M. (2001). On ‘systematic’ reviews of research literatures: A ‘narrative’ response to evans  
1130       & benefield. *British Educational Research Journal*, 27(5):543–554.
- 1131 Hanu, L. and Unitary team (2020). Detoxify. Github. <https://github.com/unitaryai/detoxify>.

- 1132 Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., and Kamar, E. (2022). ToxiGen: A large-scale  
1133 machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the*  
1134 *60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages  
1135 3309–3326, Dublin, Ireland. Association for Computational Linguistics.
- 1136 Huang, X., Xing, L., Deroncourt, F., and Paul, M. J. (2020). Multilingual Twitter corpus and baselines  
1137 for evaluating demographic bias in hate speech recognition. In *Proceedings of the Twelfth Language*  
1138 *Resources and Evaluation Conference*, pages 1440–1448, Marseille, France. European Language  
1139 Resources Association.
- 1140 Ibrohim, M. O. and Budi, I. (2018). A dataset and preliminaries study for abusive language detection in  
1141 indonesian social media. *Procedia Computer Science*, 135:222–229.
- 1142 Ibrohim, M. O. and Budi, I. (2019a). Multi-label hate speech and abusive language detection in indonesian  
1143 twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57.
- 1144 Ibrohim, M. O. and Budi, I. (2019b). Translated vs non-translated method for multilingual hate speech  
1145 identification in twitter. *Int. J. Adv. Sci. Eng. Inf. Technol*, 9(4):1116–1123.
- 1146 Ishmam, A. M. and Sharmin, S. (2019). Hateful speech detection in public facebook pages for the bengali  
1147 language. In *2019 18th IEEE international conference on machine learning and applications (ICMLA)*,  
1148 pages 555–560. IEEE.
- 1149 Jacobs, C., Rakotonirina, N. C., Chimoto, E. A., Bassett, B. A., and Kamper, H. (2023). Towards hate  
1150 speech detection in low-resource languages: Comparing asr to acoustic word embeddings on wolof and  
1151 swahili.
- 1152 Jahan, M. S. and Oussalah, M. (2021). A systematic review of hate speech automatic detection using  
1153 natural language processing. *CoRR*, abs/2106.00742.
- 1154 Jayanthi, S. M. and Gupta, A. (2021). SJ\_AJ@DravidianLangTech-EACL2021: Task-adaptive pre-  
1155 training of multilingual BERT models for offensive language identification. In *Proceedings of the*  
1156 *First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 307–312, Kyiv.  
1157 Association for Computational Linguistics.
- 1158 Jiang, A. and Zubiaga, A. (2021). Cross-lingual capsule network for hate speech detection in social media.  
1159 In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, HT '21, page 217–223,  
1160 New York, NY, USA. Association for Computing Machinery.
- 1161 Kar, P. and Debbarma, S. (2023). Multilingual hate speech detection sentimental analysis on social media  
1162 platforms using optimal feature extraction and hybrid diagonal gated recurrent neural network. *The*  
1163 *Journal of Supercomputing*, pages 1–32.
- 1164 Keung, P., Lu, Y., and Bhardwaj, V. (2019). Adversarial learning with contextual embeddings for zero-  
1165 resource cross-lingual classification and NER. In *Proceedings of the 2019 Conference on Empirical*  
1166 *Methods in Natural Language Processing and the 9th International Joint Conference on Natural*  
1167 *Language Processing (EMNLP-IJCNLP)*, pages 1355–1360, Hong Kong, China. Association for  
1168 Computational Linguistics.
- 1169 Khairy, M., Mahmoud, T. M., and Abd-El-Hafeez, T. (2021). Automatic detection of cyberbullying  
1170 and abusive language in arabic content on social networks: A survey. *Procedia Computer Science*,  
1171 189:156–166. AI in Computational Linguistics.
- 1172 Khairy, M., Mahmoud, T. M., and Abd-El-Hafeez, T. (2023). The effect of rebalancing techniques on the  
1173 classification performance in cyberbullying datasets. *Neural Computing and Applications*, pages 1–17.
- 1174 Koshiry, A. M. E., Eliwa, E. H. I., Abd El-Hafeez, T., and Omar, A. (2023). Arabic toxic tweet  
1175 classification: Leveraging the arabert model. *Big Data and Cognitive Computing*, 7(4).
- 1176 Kovács, G., Alonso, P., and Saini, R. (2021). Challenges of hate speech detection in social media: Data  
1177 scarcity, and leveraging external resources. *SN Computer Science*, 2:1–15.
- 1178 Kumar, A., Dikshit, S., and Albuquerque, V. H. C. (2021). Explainable artificial intelligence for sarcasm  
1179 detection in dialogues. *Wireless Communications and Mobile Computing*, 2021:1–13.
- 1180 Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2020). Evaluating aggression identification in  
1181 social media. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages  
1182 1–5.
- 1183 Kumar, R., Ojha, A. K., Zampieri, M., and Malmasi, S., editors (2018a). *Proceedings of the First*  
1184 *Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, Santa Fe, New Mexico, USA.  
1185 Association for Computational Linguistics.
- 1186 Kumar, R., Reganti, A. N., Bhatia, A., and Maheshwari, T. (2018b). Aggression-annotated corpus of

- 1187 hindi-english code-mixed data. *arXiv preprint arXiv:1803.09402*.
- 1188 Leite, J. A., Silva, D. F., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media  
1189 for brazilian portuguese: New dataset and multilingual analysis. *arXiv preprint arXiv:2010.04543*.
- 1190 Ljubešić, N., Erjavec, T., and Fišer, D. (2018). Datasets of slovene and croatian moderated news comments.  
1191 In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pages 124–131.
- 1192 Luu, S. T., Nguyen, K. V., and Nguyen, N. L.-T. (2021). A large-scale dataset for hate speech detection on  
1193 vietnamese social media texts. In *Advances and Trends in Artificial Intelligence. Artificial Intelligence  
1194 Practices: 34th International Conference on Industrial, Engineering and Other Applications of Applied  
1195 Intelligent Systems, IEA/AIE 2021, Kuala Lumpur, Malaysia, July 26–29, 2021, Proceedings, Part I 34*,  
1196 pages 415–426. Springer.
- 1197 Madhu, H., Satapara, S., Modha, S., Mandl, T., and Majumder, P. (2023). Detecting offensive speech in  
1198 conversational code-mixed dialogue on social media: A contextual dataset and benchmark experiments.  
1199 *Expert Systems with Applications*, 215:119342.
- 1200 Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., and Patel, A. (2019). Overview  
1201 of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european  
1202 languages. In *Proceedings of the 11th forum for information retrieval evaluation*, pages 14–17.
- 1203 Mandl, T., Modha, S., Shahi, G. K., Madhu, H., Satapara, S., Majumder, P., Schäfer, J., Ranasinghe, T.,  
1204 Zampieri, M., and Nandini, D. (2021). Overview of the hasoc subtrack at fire 2021: Hate speech and  
1205 offensive content identification in english and indo-aryan languages. *arXiv preprint arXiv:2112.09301*.
- 1206 Mathur, P., Shah, R., Sawhney, R., and Mahata, D. (2018). Detecting offensive tweets in hindi-english  
1207 code-switched language. In *Proceedings of the sixth international workshop on natural language  
1208 processing for social media*, pages 18–26.
- 1209 Mazari, A. C. and Kheddar, H. (2023). Deep learning-based analysis of algerian dialect dataset targeted  
1210 hate speech, offensive language and cyberbullying. *International Journal of Computing and Digital  
1211 Systems*.
- 1212 Mohapatra, S. K., Prasad, S., Bebarta, D. K., Das, T. K., Srinivasan, K., and Hu, Y.-C. (2021). Automatic  
1213 hate speech detection in english-odia code mixed social media data using machine learning techniques.  
1214 *Applied Sciences*, 11(18):8575.
- 1215 Montariol, S., Riabi, A., and Seddah, D. (2022). Multilingual auxiliary tasks training: Bridging the gap  
1216 between languages for zero-shot transfer of hate speech detection models. In *Findings of the Association  
1217 for Computational Linguistics: ACL-IJCNLP 2022*, pages 347–363, Online only. Association for  
1218 Computational Linguistics.
- 1219 Moon, J., Cho, W. I., and Lee, J. (2020). Beep! korean corpus of online news comments for toxic speech  
1220 detection. *arXiv preprint arXiv:2005.12503*.
- 1221 Mossie, Z. and Wang, J.-H. (2020). Vulnerable community identification using hate speech detection on  
1222 social media. *Information Processing & Management*, 57(3):102087.
- 1223 Mozafari, M., Farahbakhsh, R., and Crespi, N. (2022). Cross-lingual few-shot hate speech and offensive  
1224 language detection using meta learning. *IEEE Access*, 10:14880–14896.
- 1225 Mubarak, H., Hassan, S., and Chowdhury, S. A. (2022). Emojis as anchors to detect arabic offensive  
1226 language and hate speech. *arXiv preprint arXiv:2201.06723*.
- 1227 Mulki, H., Haddad, H., Ali, C. B., and Alshabani, H. (2019). L-hsab: A levantine twitter dataset for hate  
1228 speech and abusive language. In *Proceedings of the third workshop on abusive language online*, pages  
1229 111–118.
- 1230 Muti, A., Fericola, F., and Barrón-Cedeño, A. (2023). UniBoe’s at SemEval-2023 task 10: Model-  
1231 agnostic strategies for the improvement of hate-tuned and generative models in the classification  
1232 of sexist posts. In *Proceedings of the The 17th International Workshop on Semantic Evaluation  
1233 (SemEval-2023)*, pages 1138–1147, Toronto, Canada. Association for Computational Linguistics.
- 1234 Nascimento, G., Carvalho, F., Cunha, A. M. d., Viana, C. R., and Guedes, G. P. (2019). Hate speech  
1235 detection using brazilian imageboards. In *Proceedings of the 25th Brazilian Symposium on Multimedia  
1236 and the Web*, pages 325–328.
- 1237 Nozza, D. (2021). Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of  
1238 the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International  
1239 Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online.  
1240 Association for Computational Linguistics.
- 1241 Ocampo, N., Cabrio, E., and Villata, S. (2023). Playing the part of the sharp bully: Generating adver-

- 1242 sarial examples for implicit hate speech detection. In *Findings of the Association for Computational*  
1243 *Linguistics: ACL 2023*, pages 2758–2772, Toronto, Canada. Association for Computational Linguistics.
- 1244 Ollagnier, A., Cabrio, E., Villata, S., and Blaya, C. (2022). Cyberagressionado-v1: a dataset of annotated  
1245 online aggressions in french collected through a role-playing game. In *Language Resources and*  
1246 *Evaluation Conference*.
- 1247 Omar, A. and Abd El-Hafeez, T. (2023). Quantum computing and machine learning for arabic language  
1248 sentiment classification in social media. *Scientific Reports*, 13(1):17305.
- 1249 Omar, A., Mahmoud, T. M., and Abd-El-Hafeez, T. (2020). Comparative performance of machine learning  
1250 and deep learning algorithms for arabic hate speech detection in osns. In Hassanien, A.-E., Azar,  
1251 A. T., Gaber, T., Oliva, D., and Tolba, F. M., editors, *Proceedings of the International Conference on*  
1252 *Artificial Intelligence and Computer Vision (AICV2020)*, pages 247–257, Cham. Springer International  
1253 Publishing.
- 1254 Ombui, E., Muchemi, L., and Wagacha, P. (2019). Hate speech detection in code-switched text mes-  
1255 sages. In *2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies*  
1256 *(ISMSIT)*, pages 1–6. IEEE.
- 1257 Otter, D. W., Medina, J. R., and Kalita, J. K. (2018). A survey of the usages of deep learning in natural  
1258 language processing. *CoRR*, abs/1807.10854.
- 1259 Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., and Yeung, D.-Y. (2019). Multilingual and multi-aspect  
1260 hate speech analysis. *arXiv preprint arXiv:1908.11049*.
- 1261 Pamungkas, E. W., Basile, V., and Patti, V. (2021a). A joint learning approach with knowledge injection  
1262 for zero-shot cross-lingual hate speech detection. *Inf. Process. Manage.*, 58(4).
- 1263 Pamungkas, E. W., Basile, V., and Patti, V. (2021b). Towards multidomain and multilingual abusive  
1264 language detection: A survey. *Personal Ubiquitous Comput.*, 27(1):17–43.
- 1265 Pamungkas, E. W. and Patti, V. (2019). Cross-domain and cross-lingual abusive language detection: A  
1266 hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual*  
1267 *Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370,  
1268 Florence, Italy. Association for Computational Linguistics.
- 1269 Pavlopoulos, J., Malakasiotis, P., and Androutsopoulos, I. (2017). Deep learning for user comment  
1270 moderation. *arXiv preprint arXiv:1705.09993*.
- 1271 Pelicon, A., Shekhar, R., Martinc, M., Škrlj, B., Purver, M., and Pollak, S. (2021a). Zero-shot cross-lingual  
1272 content filtering: Offensive language and hate speech detection. In *Proceedings of the EACL Hackshop*  
1273 *on News Media Content Analysis and Automated Report Generation*, pages 30–34, Online. Association  
1274 for Computational Linguistics.
- 1275 Pelicon, A., Shekhar, R., Škrlj, B., Purver, M., and Pollak, S. (2021b). Investigating cross-lingual training  
1276 for offensive language detection. *PeerJ Computer Science*, 7:e559.
- 1277 Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., and Camacho-Collados, M. (2019). Detecting  
1278 and monitoring hate speech in twitter. *Sensors*, 19(21):4654.
- 1279 Pfeiffer, J., Goyal, N., Lin, X., Li, X., Cross, J., Riedel, S., and Artetxe, M. (2022). Lifting the curse  
1280 of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of*  
1281 *the North American Chapter of the Association for Computational Linguistics: Human Language*  
1282 *Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.
- 1283 Pham, N.-Q., Niehues, J., Ha, T.-L., and Waibel, A. (2019). Improving zero-shot translation with  
1284 language-independent constraints. In *Proceedings of the Fourth Conference on Machine Translation*  
1285 *(Volume 1: Research Papers)*, pages 13–23, Florence, Italy. Association for Computational Linguistics.
- 1286 Pikuliak, M., Šimko, M., and Bieliková, M. (2021). Cross-lingual learning for text processing: A survey.  
1287 *Expert Systems with Applications*, 165:113765.
- 1288 Pitenis, Z., Zampieri, M., and Ranasinghe, T. (2020). Offensive language identification in greek. *arXiv*  
1289 *preprint arXiv:2003.07459*.
- 1290 Plaza-Del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A., and Martín-Valdivia, M. T. (2021).  
1291 A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access*,  
1292 9:112478–112489.
- 1293 Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2021). Resources and benchmark corpora  
1294 for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.
- 1295 Ptaszynski, M., Pieciukiewicz, A., and Dybała, P. (2019). Results of the poleval 2019 shared task 6: First  
1296 dataset and open shared task for automatic cyberbullying detection in polish twitter.

- 1297 Putra, I. F. and Purwarianti, A. (2020). Improving indonesian text classification using multilingual  
1298 language model. *CoRR*, abs/2009.05713.
- 1299 Rahman, M. M., Balakrishnan, D., Murthy, D., Kutlu, M., and Lease, M. (2021). An information retrieval  
1300 approach to building datasets for hate speech detection. *arXiv preprint arXiv:2106.09775*.
- 1301 Rajamanickam, S., Mishra, P., Yannakoudakis, H., and Shutova, E. (2020). Joint modelling of emotion  
1302 and abusive language detection. In *Proceedings of the 58th Annual Meeting of the Association for*  
1303 *Computational Linguistics*, pages 4270–4279, Online. Association for Computational Linguistics.
- 1304 Ranasinghe, T., Anuradha, I., Premasiri, D., Silva, K., Hettiarachchi, H., Uyangodage, L., and Zampieri,  
1305 M. (2022). Sold: Sinhala offensive language dataset. *arXiv preprint arXiv:2212.00851*.
- 1306 Ranasinghe, T. and Zampieri, M. (2020). Multilingual offensive language identification with cross-lingual  
1307 embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*  
1308 *Processing (EMNLP)*, pages 5838–5844, Online. Association for Computational Linguistics.
- 1309 Ranasinghe, T. and Zampieri, M. (2021a). An evaluation of multilingual offensive language identification  
1310 methods for the languages of india. *Information*, 12(8).
- 1311 Ranasinghe, T. and Zampieri, M. (2021b). Multilingual offensive language identification for low-resource  
1312 languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(1).
- 1313 Ranasinghe, T. and Zampieri, M. (2023). Teacher and student models of offensive language in social  
1314 media. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3910–3922,  
1315 Toronto, Canada. Association for Computational Linguistics.
- 1316 Riabi, A., Montariol, S., and Seddah, D. (2022). Tâches auxiliaires multilingues pour le transfert de  
1317 modèles de détection de discours haineux (multilingual auxiliary tasks for zero-shot cross-lingual  
1318 transfer of hate speech detection). In *Actes de la 29e Conférence sur le Traitement Automatique des*  
1319 *Langues Naturelles. Volume 1 : conférence principale*, pages 413–423, Avignon, France. ATALA.
- 1320 Rizwan, H., Shakeel, M. H., and Karim, A. (2020). Hate-speech and offensive language detection in roman  
1321 urdu. In *Proceedings of the 2020 conference on empirical methods in natural language processing*  
1322 *(EMNLP)*, pages 2512–2522.
- 1323 Röttger, P., Nozza, D., Bianchi, F., and Hovy, D. (2022a). Data-efficient strategies for expanding hate  
1324 speech detection into under-resourced languages. In *Proceedings of the 2022 Conference on Empirical*  
1325 *Methods in Natural Language Processing*, pages 5674–5691, Abu Dhabi, United Arab Emirates.  
1326 Association for Computational Linguistics.
- 1327 Röttger, P., Seelawi, H., Nozza, D., Talat, Z., and Vidgen, B. (2022b). Multilingual HateCheck: Functional  
1328 tests for multilingual hate speech detection models. In Narang, K., Mostafazadeh Davani, A., Mathias,  
1329 L., Vidgen, B., and Talat, Z., editors, *Proceedings of the Sixth Workshop on Online Abuse and Harms*  
1330 *(WOAH)*, pages 154–169, Seattle, Washington (Hybrid). Association for Computational Linguistics.
- 1331 Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., and Pierrehumbert, J. (2021). HateCheck:  
1332 Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the*  
1333 *Association for Computational Linguistics and the 11th International Joint Conference on Natural*  
1334 *Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational  
1335 Linguistics.
- 1336 Roy, S. G., Narayan, U., Raha, T., Abid, Z., and Varma, V. (2021a). Leveraging multilingual transformers  
1337 for hate speech detection. In *Fire*.
- 1338 Roy, S. G., Narayan, U., Raha, T., Abid, Z., and Varma, V. (2021b). Leveraging multilingual transformers  
1339 for hate speech detection. *CoRR*, abs/2101.03207.
- 1340 Roychowdhury, S. and Gupta, V. (2023). Data-efficient methods for improving hate speech detection. In  
1341 *Findings of the Association for Computational Linguistics: EACL 2023*, pages 125–132, Dubrovnik,  
1342 Croatia. Association for Computational Linguistics.
- 1343 Safi Samghabadi, N., Hatami, A., Shafaei, M., Kar, S., and Solorio, T. (2020). Attending the emotions to  
1344 detect online abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*,  
1345 pages 79–88, Online. Association for Computational Linguistics.
- 1346 Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., and Stranisci, M. (2018). An italian twitter corpus of  
1347 hate speech against immigrants. In *Proceedings of the eleventh international conference on language*  
1348 *resources and evaluation (LREC 2018)*.
- 1349 Satapara, S., Modha, S., Mandl, T., Madhu, H., and Majumder, P. (2021). Overview of the hasoc subtrack  
1350 at fire 2021: Conversational hate speech detection in code-mixed language. *Working Notes of FIRE*.
- 1351 Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing.

- 1352 In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*,  
1353 pages 1–10, Valencia, Spain. Association for Computational Linguistics.
- 1354 Sharif, O., Hossain, E., and Hoque, M. M. (2021). Nlp-cuet@dravidianlangtech-eacl2021: Offensive  
1355 language detection from multilingual code-mixed text using transformers. *CoRR*, abs/2103.00455.
- 1356 Shi, X., Liu, X., Xu, C., Huang, Y., Chen, F., and Zhu, S. (2022). Cross-lingual offensive speech  
1357 identification with transfer learning for low-resource languages. *Computers and Electrical Engineering*,  
1358 101:108005.
- 1359 Sigurbergsson, G. I. and Derczynski, L. (2019). Offensive language and hate speech detection for danish.  
1360 *arXiv preprint arXiv:1908.04531*.
- 1361 Snell, J., Swersky, K., and Zemel, R. (2017). Prototypical networks for few-shot learning. *Advances in*  
1362 *neural information processing systems*, 30.
- 1363 Stappen, L., Brunn, F., and Schuller, B. W. (2020). Cross-lingual zero- and few-shot hate speech detection  
1364 utilising frozen transformer language models and AXEL. *CoRR*, abs/2004.13850.
- 1365 Steinberger, J., Brychcín, T., Hercig, T., and Krejzl, P. (2017). Cross-lingual flames detection in news  
1366 discussions. In *RANLP*, pages 694–700.
- 1367 Takawane, G., Phaltankar, A., Patwardhan, V., Patil, A., Joshi, R., and Takalikar, M. S. (2023). Leveraging  
1368 language identification to enhance code-mixed text classification.
- 1369 Tita, T. and Zubiaga, A. (2021). Cross-lingual hate speech detection using transformer models. *CoRR*,  
1370 abs/2111.00981.
- 1371 Vadakkekara Suresh, G., Chakravarthi, B. R., and McCrae, J. P. (2022). Meta-learning for offensive  
1372 language detection in code-mixed texts. In *Proceedings of the 13th Annual Meeting of the Forum*  
1373 *for Information Retrieval Evaluation, FIRE '21*, page 58–66, New York, NY, USA. Association for  
1374 Computing Machinery.
- 1375 Vashistha, N. and Zubiaga, A. (2021). Online multilingual hate speech detection: Experimenting with  
1376 hindi and english social media. *Information*, 12(1).
- 1377 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin,  
1378 I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus,  
1379 R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*,  
1380 volume 30. Curran Associates, Inc.
- 1381 Vidgen, B. and Derczynski, L. (2020). Directions in abusive language training data, a systematic review:  
1382 Garbage in, garbage out. *Plos one*, 15(12):e0243300.
- 1383 Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., and Margetts, H. (2019). Challenges and  
1384 frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language*  
1385 *Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.
- 1386 Vinyals, O., Blundell, C., Lillicrap, T., and Wierstra, D. (2016). Matching networks for one shot learning.  
1387 *Advances in neural information processing systems*, 29.
- 1388 Vitiugin, F., Senarath, Y., and Purohit, H. (2021). Efficient detection of multilingual hate speech by using  
1389 interactive attention network with minimal human feedback. In *13th ACM Web Science Conference*  
1390 *2021, WebSci '21*, page 130–138, New York, NY, USA. Association for Computing Machinery.
- 1391 Vu, X.-S., Vu, T., Tran, M.-V., Le-Cong, T., and Nguyen, H. (2020). Hsd shared task in vlsp campaign  
1392 2019: Hate speech detection for social good. *arXiv preprint arXiv:2007.06493*.
- 1393 Wang, C.-C., Day, M.-Y., and Wu, C.-L. (2022). Political hate speech detection and lexicon building: A  
1394 study in taiwan. *IEEE Access*, 10:44337–44346.
- 1395 Wang, R., Tan, X., Luo, R., Qin, T., and Liu, T.-Y. (2021). A survey on low-resource neural machine  
1396 translation. In Zhou, Z.-H., editor, *Proceedings of the Thirtieth International Joint Conference*  
1397 *on Artificial Intelligence, IJCAI-21*, pages 4636–4643. International Joint Conferences on Artificial  
1398 Intelligence Organization. Survey Track.
- 1399 Wang, S., Liu, J., Ouyang, X., and Sun, Y. (2020). Galileo at SemEval-2020 task 12: Multi-lingual  
1400 learning for offensive language identification using pre-trained language models. In *Proceedings of*  
1401 *the Fourteenth Workshop on Semantic Evaluation*, pages 1448–1455, Barcelona (online). International  
1402 Committee for Computational Linguistics.
- 1403 Weller, O., Marone, M., Braverman, V., Lawrie, D., and Van Durme, B. (2022). Pretrained models for  
1404 multilingual federated learning. *arXiv preprint arXiv:2206.02291*.
- 1405 Wigand, C. and Voin, M. (2017). Speech by commissioner jourová—10 years of the eu fundamental  
1406 rights agency: A call to action in defence of fundamental rights, democracy and the rule of law.

- 1407 Yadav, A., Chandel, S., Chatufale, S., and Bandhakavi, A. (2023). Lahm: Large annotated dataset for  
1408 multi-domain and multilingual hate speech identification. *arXiv preprint arXiv:2304.00913*.
- 1409 Yang, K., Jang, W., and Cho, W. I. (2022). Apeach: Attacking pejorative expressions with analysis on  
1410 crowd-generated hate speech evaluation datasets. *arXiv preprint arXiv:2202.12459*.
- 1411 Yin, W. and Zubiaga, A. (2021). Towards generalisable hate speech detection: a review on obstacles and  
1412 solutions. *PeerJ Computer Science*, 7:e598.
- 1413 zahra El-Alami, F., Ouatik El Alaoui, S., and En Nahnahi, N. (2022). A multilingual offensive language  
1414 detection method based on transfer learning from transformer fine-tuning model. *Journal of King Saud*  
1415 *University - Computer and Information Sciences*, 34(8, Part B):6048–6056.
- 1416 Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis,  
1417 Z., and Çöltekin, Ç. (2020). Semeval-2020 task 12: Multilingual offensive language identification in  
1418 social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*.
- 1419 Zampieri, M., Ranasinghe, T., Chaudhari, M., Gaikwad, S., Krishna, P., Nene, M., and Paygude, S. (2022).  
1420 Predicting the type and target of offensive social media posts in marathi. *Social Network Analysis and*  
1421 *Mining*, 12(1):77.
- 1422 Zhang, B., Williams, P., Titov, I., and Sennrich, R. (2020). Improving massively multilingual neural  
1423 machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the*  
1424 *Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational  
1425 Linguistics.
- 1426 Zia, H. B., Castro, I., Zubiaga, A., and Tyson, G. (2022). Improving zero-shot cross-lingual hate  
1427 speech detection with pseudo-label fine-tuning of transformer language models. *Proceedings of the*  
1428 *International AAI Conference on Web and Social Media*, 16(1):1435–1439.
- 1429 Zoph, B. and Le, Q. V. (2016). Neural architecture search with reinforcement learning. *arXiv preprint*  
1430 *arXiv:1611.01578*.