Cross-lingual Offensive Language Detection: A Systematic Review of Dataset, Approach and Challenge

AIQI JIANG, School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom and School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, United Kingdom
ARKAITZ ZUBIAGA, School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom

The growing prevalence and rapid evolution of offensive language in social media amplify the complexities of detection, particularly highlighting the challenges in identifying such content across diverse languages. This survey presents a systematic and comprehensive exploration of Cross-Lingual Transfer Learning (CLTL) techniques in offensive language detection in social media. Our study stands as the first holistic overview to focus exclusively on the cross-lingual scenario in this domain. We analyse 67 relevant papers and categorise these studies across various dimensions, including the characteristics of multilingual datasets used, the cross-lingual resources employed, and the specific CLTL strategies implemented. According to "what to transfer", we also summarise three main CLTL transfer approaches: instance, feature, and parameter transfer. Additionally, we shed light on the current challenges and future research opportunities in this field. Furthermore, we have made our survey resources available online, including two comprehensive tables that provide accessible references to the multilingual datasets and CLTL methods used in the reviewed literature.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; *Machine learning*; • **Human-centered computing** → **Collaborative and social computing;**

Additional Key Words and Phrases: Offensive language detection, hate speech detection, literature review, survey, cross-lingual, multilingual, social media, text classification

ACM Reference Format:
Aiqi Jiang and Arkaitz Zubiaga. 2026. Cross-lingual Offensive Language Detection: A Systematic Review of Dataset, Approach and Challenge. *ACM Comput. Surv.* 58, 10, Article 269 (April 2026), 37 pages. https://doi.org/10.1145/3801736

Disclaimer: *Due to the nature of this work, some examples may contain offensive text and hate speech.*

Aiqi Jiang is funded by China Scholarship Council (CSC Funding, No. 201908510140). Arkaitz Zubiaga acknowledges support from the European Union and UK Research and Innovation under Grant No. 101073351 as part of Marie Skłodowska-Curie Actions (MSCA Hybrid Intelligence to monitor, promote, and analyse transformations in good democracy practices).
Authors' Contact Information: Aiqi Jiang (corresponding author), School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK, School of Mathematical and Computer Sciences and Heriot-Watt University, Edinburgh, UK; e-mail: aiqi.jiang@yahoo.com; Arkaitz Zubiaga, School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK; e-mail: a.zubiaga@qmul.ac.uk.

This work is licensed under a Creative Commons Attribution 4.0 International License.

© 2026 Copyright held by the owner/author(s).
ACM 0360-0300/2026/04-ART269
https://doi.org/10.1145/3801736

1 Introduction

The prevalence of offensive language on social media platforms, such as Facebook and Twitter, has become a pressing concern in recent years. This surge can be attributed to the anonymity provided by these platforms and the lack of strict regulations and effective mechanisms to restrain such behaviour [64]. While social platforms have fostered connections and bridged global distances, they have in turn enabled the spread of **hate speech** (**HS**) and various forms of offensive language [33]. Offensive language broadly refers to expressions that may upset or annoy others, while HS escalates beyond mere offensive behaviour to inciting discrimination, hostility, or violence against individuals or groups based on their race, religion, ethnicity, gender, or other identity factors [62]. HS is considered a human rights violation in many legal systems due to its potential to cause real-world harm and perpetuate discrimination against targeted groups [178]. Hence, there is a growing scholarly focus on developing approaches to mitigate HS in addition to offensive content.

The **Natural Language Processing** (**NLP**) community has proposed a variety of research techniques to detect offensive content based on both traditional machine learning and advanced neural network-based approaches. While early work focused on monolingual settings (i.e., English), research in recent years has increasingly focused on tackling the problem in multilingual settings to broaden the ability to deal with more languages, a key requirement given the presence of diverse languages in social platforms [128]. This however generally hindered by the limited availability of labelled data and the high variability of offensive language across diverse cultures and languages [142].

Cross-Lingual Transfer Learning (**CLTL**) emerges as a promising direction to mitigate challenges associated with data scarcity by leveraging domain knowledge from high- to low-resource languages [139]. Given its adaptability and compatibility with neural networks, CLTL has found successful applications across NLP tasks [130]. Some pioneering efforts have also integrated CLTL for offensive language detection in low-resource languages [87, 168, 197]. However, open challenges remain in building effective and generalisable cross-lingual models and understanding linguistic gaps.

Interest in cross-lingual offensive language detection has increased since the first work in 2018 [111], but this is yet to be covered in a comprehensive survey. Existing surveys concentrate on monolingual detection or specific characteristics of offensive language, while the unique CLTL application in offensive language detection is seldom considered. Relevant surveys exist [33, 128], but their scope goes beyond cross-lingual scenarios, also covering cross-domain and multimodal dimensions, and hence lacking our survey's focus on CLTL.

1.1 Scope of the Survey

Our objective is to provide a systematic review of existing literature on CLTL techniques for offensive language detection, and consequently, this survey discusses **67 papers** in this area. To make this scope explicit, we set out the following key objectives:

- **Map the landscape** of multilingual datasets and cross-lingual resources used in offensive language detection, discussing aspects such as languages, label schemes, and availability.
- **Classify and analyse methods** by introducing a fine-grained taxonomy of CLTL approaches tailored to this task, distinguishing instance-, feature-, and parameter-level transfer.
- **Identify key challenges** that limit progress, with a focus on language diversity, dataset scarcity, cultural variation, and modelling limitations in low-resource scenarios.
- **Recommend future research directions** for developing more generalisable, robust, and ethically sound cross-lingual detection systems.

Furthermore, we make survey resources available online,[1] including two comprehensive tables that summarise multilingual datasets and CLTL methods used in surveyed papers.

1.2 Survey Structure

The rest of the article is organised as follows. Section 2 describes the definitions of offensive language and the framework of cross-lingual detection, and distinguishes 'cross-lingual' from related terminologies such as 'multilingual'. Related surveys are also summarised. Section 3 summarises our survey methodology. Then we summarise and analyse the multilingual datasets leveraged in surveyed papers in Section 4. In Section 5, we describe diverse linguistic resources and tools used in cross-lingual studies. Section 6 discusses the three transfer levels of CLTL in offensive language detection, complemented by transfer strategies based on these layers. Section 7 presents current challenges and future research directions in this area, concluding the survey in Section 8.

2 Background and Related Work

2.1 Cross-Linguality in Offensive Language Detection

Task Definition and Workflow. The task focuses on identifying and categorising offensive or hateful text across languages. The objective of CLTL is to leverage the knowledge acquired from one language (source language with more resources) to enhance HS detection in another language (target language with less resources), especially when the labelled data in the target language is scarce or lacking [128]. The task workflow can be thought of as a four-stage process:

(1) **Data Preparation.** Gather text in both a source language with more labelled examples and a target language with fewer or no labels. The source data provides the foundation for learning, while the target data is where we ultimately want to detect offensive content.
(2) **Cross-lingual Training.** Use the source data alone, or combine it with any available target data, to train a model how to recognise offensive language patterns.
(3) **Model Adaptation.** If any labelled examples exist in the target language, fine-tune the model so it better reflects specific characteristics of the target language.
(4) **Detection on Target Language.** Apply the adapted model to unseen target language text to assign offensive/non-offensive labels.

Formally, we represent the process as follows. Let L_s and L_t be the source and target languages, with datasets $D_s = \{X_s, Y_s\}$ and $D_t = \{X_t, Y_t\}$, where X denotes text and $Y \in \{0, 1\}$ denotes non-offensive (0) or offensive (1) labels. The cross-lingual task seeks a function $f : F \rightarrow Y_t$ mapping a feature space F to the target label space. In the training stage, it can use only D_s or both D_s and D_t:

— **Source-only training:** Train M_s on D_s to learn $f_s : F_s \rightarrow Y$, then adapt f_s to $f_t : F_t \rightarrow Y$.
— **Joint training:** Train M_{st} on $D_s \cup D_t$ to learn $f_{st} : F_{st} \rightarrow Y$ based on joint feature space F_{st}.

If Y_t is available, fine-tune f_t (or f_{st}) on full or few-shot D_t, producing M_t for prediction:

$$\hat{y}_t = f_t(D_t) \quad \text{or} \quad \hat{y}_t = f_{st}(D_t).$$

Cross-lingual and Multilingual and Code-mixing. While "cross-lingual learning" typically denotes the application of CLTL techniques, the term "multilingual learning" is sometimes used interchangeably [139]. Some multilingual data may contain code-mixed content, i.e. a mixture of languages in the same sentence [82]. Although some researchers draw parallels between these terms and the relationship between transfer learning and multi-task learning [184], we adopt a broader

[1] https://github.com/aggiejiang/crosslingual-offensive-language-survey

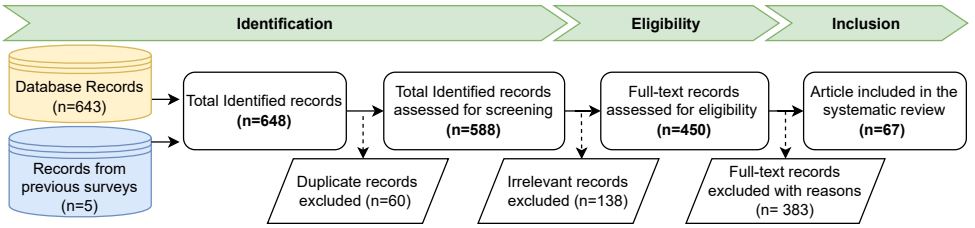

Fig. 1. PRISMA flowchart showing the phases of the selection of research articles in this review.

perspective on cross-lingual learning. We define cross-lingual learning as the process of transferring knowledge between different languages, a concept that inherently subsumes multilingual learning.

2.2 Related Surveys

We found 18 surveys covering offensive language detection from different angles.

General Offensive Language Detection. Surveys on offensive language detection provide general insights into the field [64, 84, 112, 166]. Key works by Schmidt and Wiegand [166] and Fortuna and Nunes [64] stand out as the two most cited reviews in the NLP community, providing definitions of HS and related terms, summarising previous methods, and discussing the existing limitations and challenges. Some of these surveys focus on in-depth discussions of specific aspects when reviewing studies in HS detection, such as data sources [6, 143, 181], methodologies [6, 17, 44], and challenges [17, 69, 93, 192]. Additionally, several surveys have also delved into related areas like cyberbullying [191] and the intersection of HS and fake news [44], and others review ethical considerations in the development of HS detection [93, 181].

Cross-lingual and Multilingual Offensive Language Detection. Several surveys focus on studies in single but non-English languages, such as Arabic [2, 80], Indian [49] and Ethiopian [44] languages. Surveys for HS detection also provide a brief summary of multilingual or cross-lingual contributions [2, 143, 192]. However, systematic reviews focusing explicitly on cross-lingual perspectives remain scarce [33, 128]. Pamungkas et al. [128] provide an overview of datasets and approaches in cross-domain and cross-lingual settings. Chhabra and Vishwakarma [33] conduct a comprehensive review of multimodal and multilingual HS detection, including benchmark datasets, state-of-the-art methods, challenges, and evaluation techniques.

Strength of our survey. In contrast, our work presents:

- *Fine-grained taxonomy of CLTL approaches* for offensive language detection, structured into instance-, feature-, and parameter-level transfer, each with detailed strategy categorisation;
- *Comprehensive dataset library* including rich metadata (e.g., language families, annotation schemes, abuse types, availability, etc.);
- *In-depth analysis of challenges* unique to cross-lingual offensive language detection from aspects of languages, data sources and approaches;
- *Future-oriented perspective* that synthesises promising research directions and provides recommendations for future research.

3 Survey Methodology

We adopt the structured PRISMA framework [125] for our systematic review to ensure both replicability and transparency. A detailed flowchart of our search protocol is provided in Figure 1, which illustrates the progression of publications through different phases of our selection process.

Table 1. Eligibility Criteria for Inclusion/Exclusion Studies

Include	Exclude
(i) Describe cross-lingual approaches to detect general abusive content (toxic, hate, etc.)	(i) Study transfer across dimensions other than languages (e.g. platforms, domains, datasets)
(ii) Describe cross-lingual approaches to detect subtypes of abusive content (e.g. sexism, racism)	(ii) Study low-resource languages without knowledge transfer approaches
(iii) Describe multilingual approach with knowledge transfer between languages	(iii) Relevant studies from competition reports and papers
(iv) Study text classification cases in cross-lingual HS scenarios	(iv) Master or PhD dissertations

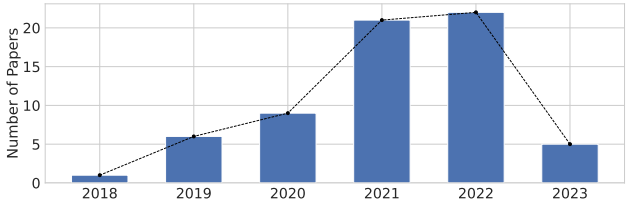

Fig. 2. Publications per year up to July 2023.

Search Databases. We retrieve publications from scientific databases including ACL Anthology,[2] Google Scholar,[3] and DBLP Computer Science Bibliography.[4]

Keyword Selection. We create a list of 99 pairwise combinations of words in (i) and (ii): (i) 9 keywords related to offensive language, such as "hate", "abus*", "offens*", "toxic", "cyberbully", "gender", "sexis*", "misogyn*", "rac*"; (ii) 11 keywords related to multilingual and cross-lingual scenarios, such as "cross-lingual", "multilingual", "low resource", "under resource", "cross language", "between language", "cross linguistic", "transfer", "code switch", "zero shot", "few shot".

Search Process. In addition to studies matching the 99 keyword pairs above, we also identify and include 5 additional studies cited in relevant surveys. This search was conducted on July 29th, 2023.

Eligibility Criteria. Following removal of duplicates and irrelevant papers, the remainder 455 publications were checked against the inclusion / exclusion criteria in Table 1.

Summary. In the end, we obtained 67 relevant papers, with their temporal distribution shown in Figure 2 and showing a growing trend from the first publication in 2018 to 2022.

4 Multilingual Hate Speech Datasets

We investigate 82 datasets utilised in existing surveyed studies on cross-lingual HS detection across different aspects. We summarise these datasets and their characteristics in Table 2.

Topics. The topic distribution among datasets shows that "offensiveness" and "hate speech" are the two most frequently addressed topics, accounting for 34.23% and 32.43% respectively. Some cover general offense, such as abusiveness, toxicity, cyberbullying, harassment and aggressiveness. In addition, some datasets narrow their scope to targeted issues, like sexism, misogyny and racism.

Data Sources. Most studies collect data from social media such as Twitter (now X) and Facebook. Twitter encompasses 47.19% of the datasets, followed by Youtube (11.24%) and Facebook (10.11%).

[2] https://aclanthology.org/
[3] https://scholar.google.com/
[4] https://dblp.org/

Table 2. Summary of Datasets used in Cross-lingual Offensive Language Detection

Ref	Year	Topic	Source	Language	Size	Label	#Cit	CM?	Avail?
[132]	2023	cyberbullying	Twitter	en, hi	22k	1	<50	✓	✗
[85]	2022	offense	NAVER-news, Youtube	ko	40,429	1,2,3	<10	✗	✓
[86]	2022	Sexism	Weibo	zh	8,969	1,2,3	<50	✗	✓
[162]	2022	offense	-	en, fr, de, es	37,221	1	<10	✓	✓
[45]	2022	offense	Weibo, Zhihu	zh	37,480	1,2	<50	✗	✓
[73]	2022	abuse	ShareChat	hi, kn, ml, ta, te	92,881	1	<10	✓	✓
[30]	2022	offense	Youtube	ta	60k	1,3	<50	✓	✓
[13]	2022	offense	Twitter	es	9,834	2	<10	✗	✓
[20]	2022	aggressiveness, misogyny	Youtube	bn	25k	1	<100	✗	✓
[95]	2022	cyberbullying	Facebook, Twitter	en, hi	6,500	1	<50	✓	✗
[133]	2021	toxicity	Civil Comments	en	10,629	1,2	<100	✗	✓
[29]	2021	toxicity	Ask.fm	nl	10,189	2	<10	✗	✗
[78]	2021	offense	Youtube	kn, ml, ta	71,691	1,3	<20	✓	✓
[109]	2021	hate speech	Gab, Twitter	en	20,148	1,2	<500	✗	✓
[156]	2021	hate speech	Facebook, Youtube	bn	30k	1	<50	✗	✗
[113]	2021	hate speech, offense	Twitter	en, hi, mr	13,755	1,4	<100	✗	✓
[9]	2021	abuse	Twitter	ur	8,400	1	<50	✗	✓
[182]	2021	hate speech	open-source platform	en	41,255	1,2	<100	✗	✓
[67]	2021	offense	Twitter	mr	2499	1,2,3	<50	✗	✓
[92]	2021	hate speech, offense	Twitter	roman ur	5k	1	<50	✗	✗
[118]	2021	offense	Twitter	ar	10k	1,2	<250	✗	✓
[140]	2020	offense	Twitter	el	4,779	1	<250	✗	✓
[31]	2020	hate speech	Twitter	en, fr, de, el, es	264,035	1	<50	✗	✓
[70]	2020	hate speech, abuse	Facebook, Fox news, Twitter, Wikipedia	sq, hr, en, de, ru, tr	109,955	1	<50	✗	✓
[167]	2020	offense	24sata, Eesti Ekspress, Večernji List	hr, et	62.6m	-	<10	✗	✗
[1]	2020	offense, abuse	Youtube	Roman ur, ur	12,171	1	<100	✗	✓
[152]	2020	hate speech, offense	Twitter	Roman ur	10,012	1,2	<50	✗	✓
[106]	2020	offense	Twitter, Youtube	en, de, hi, ml, ta	15,047	1,2	<250	✓	✓
[163]	2020	hate speech	Twitter	it	12,081	1	<100	✗	✓
[91]	2020	hate speech	Gab	en	27,655	1	<250	✗	✓
[36]	2020	offense	Twitter	tr	36,232	1,3	<250	✗	✓
[96]	2020	aggressiveness	Youtube	bn, en, hi	15k	1	<250	✗	✓
[169]	2020	abuse, offense	Facebook, Reddit, Twitter	da	3,600	1,3	<250	✗	✓
[77]	2020	offense	Youtube	kn	7,671	1,3	<100	✓	✗
[195]	2020	offense	Twitter	ar, da, en, el, tr	9m	1,3	<500	✗	✓
[104]	2019	offense	Facebook	en, sl	22,877	1,2	<50	✓	✗
[14]	2019	cyberbullying	Twitter	ja	4,096	1	<10	✗	✗
[10]	2019	hate speech	Twitter	en	14,949	1	<250	✗	✓
[83]	2019	abuse, hate speech	Twitter	id	5,561	1,2,3	<250	✗	✓
[119]	2019	abuse, hate speech, toxicity	Twitter	at	5,846	1	<250	✗	✗

(Continued)

Table 2. Continued

Ref	Year	Topic	Source	Language	Size	Label	#Cit	CM?	Avail?
[136]	2019	hate speech	Twitter	es	6k	1	<100	✗	✓
[65]	2019	hate speech, offense	Twitter	pt	5,668	1	<100	✗	✓
[143]	2019	cyberbullying, toxicity	Twitter	pl	11,041	1,2	<50	✗	✓
[175]	2019	offense	Twitter	de	7,025	1,2	<250	✗	✓
[19]	2019	aggressiveness	Twitter	en	19,600	1,3	<750	✗	✓
[194]	2019	offense	Twitter	en, es	14k	1,3	<700	✗	✓
[107]	2019	hate speech, offense	Twitter, Facebook	en, de, hi	17,657	1,2,3	<100	✗	✓
[35]	2019	hate speech	Twitter	en, fr, it	4,078	1,2	<250	✗	✓
[123]	2019	hate speech	Twitter	ar, en, fr	13,014	1,2,3	<250	✗	✓
[34]	2019	sexism, misogyny	Twitter	fr	3,085	1	<50	✗	✗
[25]	2018	hate speech	Facebook, Twitter	en, it	4k	1	<250	✗	✓
[41]	2018	hate speech	Stormfront	en	10,568	1	<500	✗	✓
[144]	2018	cyberbullying	Formspring.me	en	12,772	1	<50	✗	✗
[3]	2018	abuse, offense	Youtube	ar	167,549	1	<100	✗	✓
[24]	2018	hate speech	Twitter	en, hi	4,575	1	<250	✓	✓
[53]	2018	hate speech	Twitter	en	27,330	1,2	<250	✗	✓
[57]	2018	misogyny	Twitter	en, it	20k	1,2,3	<250	✗	✓
[110]	2018	abuse, hate speech, offense	Twitter	en, hi	17,698	1,2	<100	✗	✓
[97]	2018	aggressiveness	Facebook, Twitter	en, hi	39k	1,2,3	<250	✓	✓
[111]	2018	offense	Twitter	en, hi	3,679	1,2	<250	✓	✗
[164]	2018	aggressiveness, hate speech, offense	Twitter	it	6k	2,4	<250	✗	✓
[66]	2018	abuse, hate speech	Twitter	en	80k	1,2	<550	✗	✓
[188]	2018	hate speech, offense	Twitter	de	8,541	1,2	<500	✗	✓
[58]	2018	misogyny	Twitter	en, es	8,115	1,2,3	<250	✗	✓
[71]	2017	harassment, offense, racism	BlockTogether, Twitter	en	35k	1	<100	✗	✓
[48]	2017	offense	Twitter	ru	493	1,2	<50	✗	✗
[68]	2017	hate speech	Fox news	en	1,528	1	<250	✗	✓
[189]	2017	hate speech	Wikipedia	en	115,737	1	<750	✗	✓
[43]	2017	hate speech, offense	g1.globo.com	pt	1,250	1	<100	✗	✓
[157]	2017	hate speech, offense	Twitter	de	541	1,2,4	<500	✗	✓
[4]	2017	hate speech	Twitter	id	713	1	<250	✗	✓
[63]	2017	hate speech	Twitter	pt	5,668	1,2	<50	✗	✓
[27]	2017	hate speech, offense	Facebook	de	5,836	1,3	<100	✗	✓
[40]	2017	offense	Twitter	en	24,802	1,2	<2200	✗	✓
[116]	2017	abuse, hate speech, offense	Twitter	ar	32k	1	<500	✗	✓
[186]	2016	hate speech	Twitter	en	16,914	1	<1550	✗	✓
[179]	2015	cyberbullying	Ask.fm	nl	85,462	2,4	<250	✗	✗
[51]	2010	cyberbullying	informal websites of Japanese secondary schools	ja	2,999	1	<100	✗	✗

"#Cit" = "number of citations", "CM?" = "whether or not the dataset is code-mixed", "Avail?" = "dataset availability", and "Label" denotes: (1) binary labels, (2) fine-grained offense category, (3) attack target, and (4) intensity score.

While these mainstream platforms dominate, news websites like Fox News and NAVER news, and open forums like Reddit and 2ch, also contribute to multilingual datasets. Sources like Weibo (Chinese), 2ch (Russian), g1.globo (Brazilian Portuguese) and Eesti Ekspress (Estonian) represent non-English speaking regions as more culturally specific platforms. Some datasets originate from

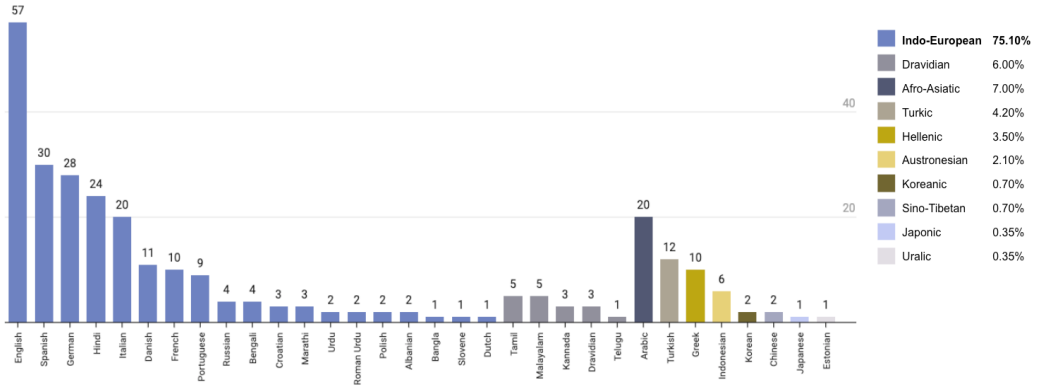

Fig. 3. Distribution of languages and language families covered in the datasets.

platforms known for controversial content or specific user bases, such as Stormfront and alt-right websites.

Languages and Families. Datasets cover 32 languages across 10 language families, where one dataset may contain more than one language or family. We show the distribution of languages and families in Figure 3. It highlights a representative language of English, accounting for 25.95% of the data, followed by Hindi at 9.16%, and German at 6.87%. Notably, it exhibits a relatively balanced representation across several languages, such as Spanish, Arabic, Italian, French, Portuguese and Turkish, each ranging from around 2% to 5%. It also reveals a strong emphasis on Indo-European languages in HS studies, complemented by a notable presence of Afro-Asiatic languages (represented primarily by Arabic), with other language families being less represented. Some datasets (14.6%) are also dedicated to code-mixed content.

Dataset Size. Over half of the datasets, accounting for 51.20%, contain 10^4-10^5 instances, closely followed by 35.40% that have instances between 10^3 and 10^4. Besides, large datasets are rare, with only one exceeding 10^7, and two between 10^6 and 10^7. The majority of datasets fall within the smaller size ranges, which indicates the challenges in collecting and manually annotating large-scale datasets, as well as model generalisability for cross-lingual HS.

Data Labelling and Distributions. According to diverse annotation schemes, four different types of labels are often employed: (i) binary labels, (ii) fine-grained categories of offensive content, (iii) attack targets, and (iv) intensity scores. A majority of datasets, 76 in total, utilise the simpler binary labelling, of which 36 solely provide binary labels. There are also datasets combining binary labels with either fine-grained categories (20), attack targets (9), or intensity scores (1). A smaller number of datasets (10) combine binary labels, fine-grained categories, and attack targets. Interestingly, only four datasets consider intensity scores, suggesting that quantifying the severity of offensive content is less common in current research.

Dataset Availability. We find 66 out of 82 datasets (80%) are publicly accessible, with the other 16 unavailable, in some cases with the possibility of contacting the authors to request the data.

Competitions and Shared Tasks. The rising concern surrounding online abuse has led to the establishment of numerous competitions and shared tasks within both national (e.g., GermEval, Evalita, IberLEF) and international (e.g., SemEval) evaluation campaigns. These open scientific competitions release benchmark datasets and invite participants to submit detection results and detailed system reports [143]. Table 3 lists 24 such competitions and shared tasks from 2018 to 2023, each addressing different facets of offensive content across multiple languages.

Table 3. Summary of Competitions and Shared Tasks in Automated Identification of Cross-Lingual HS

Name	Description	Year	Topic	Language	#Teams
GermEval [188]	Identification of Offensive Language	2018	offense	de	20
AMI@IberEval [58]	Automatic Misogyny Identification	2018	misogyny	en, es	11
AMI@Evalita [57]	Automatic Misogyny Identification	2018	misogyny	en, it	16
HaSpeeDe@Evalita [25]	Hate Speech Detection	2018	hate speech	it	9
TRAC-1 [97]	Aggression Identification of Hindi-English Code-mixed Data	2018	aggressive-ness	en, hi	30
HatEval@SemEval [19]	Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter	2019	hate speech	en, es	74
HASOC@FIRE [107]	Offensive Content Identification in Indo-European Languages	2019	hate speech, offense	en, hi, de	37
Task2@GermEval [175]	Identification of Offensive Language	2019	offense	de	13
Task6@PolEval [145]	Automatic Cyberbullying Detection in Polish Twitter	2019	cyberbullying	pl	9
OffensEval@Se-mEval [195]	Multilingual Offensive Language Identification in Social Media	2020	offense	en, ar, da, el, tr	145
HASOC@FIRE [106]	Offensive Language Identification in Tamil, Malayalam, Hindi, English and German	2020	hate speech, offense	en, hi, de, ta, ml	40+
HaSpeeDe@Evalita [163]	Hate Speech Detection	2020	hate speech	it	14
TRAC-2 [96]	Aggression and Gendered Aggression Identification	2020	aggressive-ness	en, hi, bn	19
Jigsaw Toxic@Kaggle	Jigsaw Multilingual Toxic Comment Classification	2020	toxicity	en, es, tr, pt	1621
OSACT4 [117]	Arabic Offensive Language Detection	2020	offense	ar	27
HASOC@FIRE [113]	Offensive Content Identification in English and Indo-Aryan Languages and Conversational Hate Speech	2021	hate speech, offense	en, hi, mr	65
EXIST@IberLEF [154]	Sexism Identification in Social Networks	2021	sexism	en, es	31
UrduThreat@FIRE [9]	Abusive and Threatening Language Detection in Urdu	2021	abuse, threat	ur	19
IIIT-D@Kaggle	Moj Multilingual Abusive Comment Identification across Indic Languages	2021	abuse	10+ Indic	54
HASOC@FIRE [165]	Offensive Content Identification in English and Indo-Aryan Languages	2022	hate speech, offense	en, hi, de, mr	12
EXIST@IberLEF [155]	Sexism Identification in Social Networks	2022	sexism	en, es	19
HaSpeeDe@Evalita [98]	Political and Religious Hate Speech Detection	2023	hate speech	it	6
EXIST@IberLEF [141]	Sexism Identification in Social Networks	2023	sexism	en, es	28
HASOC@FIRE	Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages	2023	hate speech, offense	en, hi, de, Indo-Aryan	-

Relevant monolingual tasks in non-English languages are also included. "#Teams" = "number of teams participated in the competition and submitted runs". All task links are added to the "Name" column.

Prominent topics include offense and HS. Specifically, GermEval competitions in 2018 [188] and 2019 [175] centre around identifying offensive content in German tweets. OffensEval, initially in 2019 focused only on English content [194], expanded to include multiple languages in the subsequent edition in 2020 [195]. The task in OSACT4 [117], the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, focuses on offensive language detection in Arabic, especially its dialectal forms. HaSpeeDe has consistently addressed HS with various general and particular topics [25, 98, 163]. For instance, HaSpeeDe [25] and HaSpeeDe 2 [163] identify HS against immigrants and Muslims, while HaSpeeDe 3 [98] explores HS in strongly polarised debates, particularly political and religious topics. Similarly, the HASOC series, recurring annually from 2019 to 2023, identify HS and offensive content across various languages [106, 107, 113, 165]. TRAC workshops, held in 2018 [97] and 2020 [96], tackled trolling, aggression, and cyberbullying. Other competitions, like PolEval [145] and UrduThreat [9], focus on specific languages (such as Polish and Urdu) or topics (e.g., cyberbullying and threatening languages). The Jigsaw Multilingual Toxic Comment Classification competition on Kaggle attracted over 1,600 participants to build multilingual models using English-only training data. Another Kaggle competition, IIIT-D Multilingual Abusive Comment Identification, focuses on abusive comments in Indic languages. Additionally, some tasks focus on more specific topics, such as sexism and misogyny. AMI is organised by IberEval [58] and Evalita [57] in 2018, specifically addressing Spanish and Italian misogynous content respectively. EXIST is held for three consecutive years from 2021 to 2023 and aims to identify sexism in social networks [141, 154, 155]. Most of these benchmark datasets originate from social media platforms, and often conduct annotations that go beyond simple binary labels to finer-grained ones.

5 Cross-Lingual Resources

This section reviews the cross-lingual resources that enable offensive language detection, including datasets, lexicons, models, and auxiliary tools that facilitate transfer across languages.

Multilingual linguistic Resources. There are some basic but essential linguistic resources that facilitate the task of cross-lingual HS detection across two or more languages. Among these, multilingual lexicons and parallel corpora stand out as foundational resources, frequently utilised by researchers to bridge linguistic gaps and enhance model performance in CLTL.

- **Multilingual Lexicons (word-aligned):** contain words, phrases, or terms in two or more languages. These lexicons often provide direct translations or equivalents of terms across the languages they cover, such as HurtLex.[5]
- **Parallel Corpora (sentence-aligned):** refer to datasets that consist of texts in two or more languages, where each text in one language has a direct translation in the other languages.

Due to the lack of sufficient labelled datasets for HS, multilingual lexicons can help bridge this gap by providing connections between resource-rich and resource-poor languages [87, 127], while parallel corpora provide translations of labelled instances from resource-rich to resource-poor languages, to facilitate CLTL [23, 151, 173]. Cross-lingual word embeddings are also created by utilising these multilingual resources. These resources are most often created for specific domains, such as HS and abusive language. While HS patterns can be universal, they can also be specific to certain cultures or languages. They provide insights into how HS manifests differently across languages and cultures, furthering the understanding of cultural nuances [121].

Machine Translation and Transliteration Tools. Translation refers to converting content from one language into another while preserving meaning of the original text, whereas transliteration is the process of converting text from one script into another by preserving phonetic form rather than

[5] https://github.com/valeriobasile/hurtlex

semantic content [78, 111]. Machine translation tools can translate datasets from one language to another, such as Google Translate,[6] Microsoft Translator,[7] DeepL[8]) and machine translation models (e.g., mBART [103] and mT5 [190]). Besides, machine transliteration tools are essential for code-mixed datasets [78, 95, 111], such as Google Transliteration,[9] Microsoft Transliteration,[10] AI4Bharat transliteration application,[11] and transliteration python packages (e.g., indic-transliteration[12]). However, in offensive language detection, some translation tools may lead to translation errors or cultural nuances that are inconsistent with the original meaning [127]. For instance, derogatory terms or slurs without direct equivalents in other languages, or their severity might differ across cultures. Additionally, idiomatic expressions that convey abuse in one language might lose their offensive meaning when translated literally. Therefore, translation quality is crucial to maintain the effectiveness and robustness of cross-lingual models.

Multilingual Representations. Multilingual representations, as language independent representations, are often used in CLTL. With these representations we can directly address the difference between source and target languages by projecting the text into a shared feature space. Then, it is easier to project the behaviour of the model. In general, we can distinguish between using word-level representations and sentence-level representations.

Multilingual distributional representations encode words from multiple languages in a single distributional word vector space. Cross-lingual transfer of word embeddings aims to establish the semantic mappings among words in different languages by learning the transformation functions over the corresponding word embedding spaces. In this space, semantically similar words (e.g. 'cat' in English and 'katze' in German) are close together independently of the language. Some prominent distributional embeddings are Multilingual GloVe [56], Multilingual FastText [72], Babylon [170], and **Multilingual Unsupervised or Supervised word Embeddings** (**MUSE**) [99, 100]. During the training, multilingual distributional representations usually require additional cross-lingual resources, e.g., bilingual dictionaries or parallel corpora.

Multilingual contextualised representations on sentence level work on a similar principle, but they use sentences instead of words for alignments. They are usually based on an auto-encoder architecture, in which the model is pushed to create similar and context-aware representations for parallel sentences, such as **Language Agnostic Sentence Representations** (**LASER**) [15] and **Language Agnostic BERT Sentence Embeddings** (**LabSE**) [55].

Monolingual and Multilingual Pre-trained Language Models. Pre-trained language models (**PLMs**) are a SOTA technique in NLP, as a language model trained on large amounts of data. PLMs can be initialised and further trained or fine-tuned with target data for CLTL. The best-known examples of PLMs are **Embeddings from Language Models** (**ELMo**) [137] and **Bidirectional Encoder Representations from Transformers** (**BERT**) [47]. **Multilingual PLMs** are an extension of monolingual PLMs, trained on multiple languages. They learn the connections between languages by developing a shared linguistic representation. There is no need to provide multilingual PLMs with any additional information about interlanguage relations, and cross-lingual transfer is possible between any languages [139]. Prominent models include **Multilingual BERT** (**mBERT**) [47] and XLM-RoBERTa (XLM-R) [37], demonstrating remarkable capabilities in understanding and processing multiple languages simultaneously.

[6] https://translate.google.com
[7] https://translator.microsoft.com/
[8] https://www.deepl.com/en/translator
[9] https://www.google.co.in/inputtools/try/
[10] https://www.microsoft.com/en-us/translator/business/translator-api/
[11] https://github.com/AI4Bharat/IndianNLP-Transliteration
[12] https://github.com/indic-transliteration/indic_transliteration_py

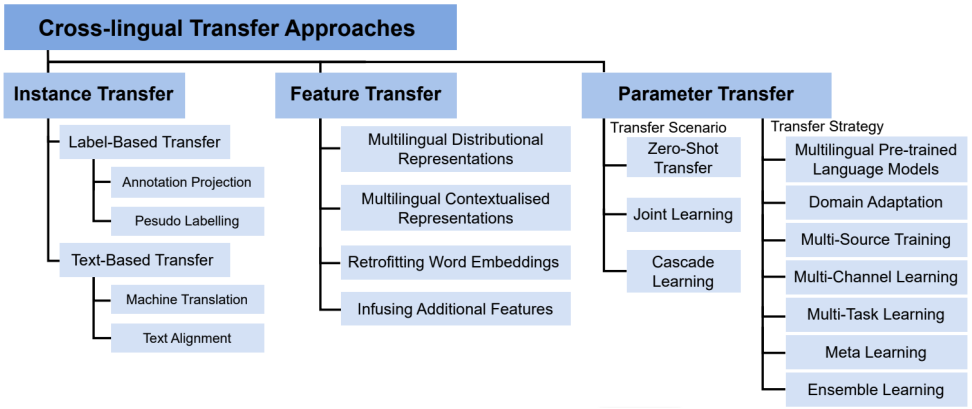

Fig. 4. Hierarchy of cross-lingual transfer approaches.

Table 4. Transfer Levels Applied in Studied Papers

Transfer Level	Reference
Instance	[7, 8, 12, 26, 34, 39, 46, 52, 78, 82, 87, 95, 111, 126, 127, 129, 131, 161, 162, 168, 172, 180, 193, 197, 198]
Feature	[7, 11, 18, 21–23, 38, 42, 74, 88, 132, 134, 153, 168, 171, 183]
Parameter	[7, 8, 12, 16, 21–23, 26, 29, 34, 39, 42, 46, 52, 54, 60, 67, 70, 73, 76, 78, 81, 87, 90, 95, 108, 111, 114, 115, 120, 121, 123, 126, 127, 129, 131, 134, 135, 138, 142, 146–149, 153, 158, 159, 161, 162, 168, 171, 172, 174, 177, 180, 193, 197, 198]

Language-Agnostic Resources Language-agnostic resources (e.g., emojis, capitals, and punctuations) can be seen as common traits among different languages, sharing similar knowledge to enhance the semantic connection across languages. For instance, emojis are generally associated with emotional expressions, which in turn are associated with various forms of online hate [38].

6 Cross-Lingual Transfer Approaches

In this section, we systematically describe and compare different cross-lingual techniques for offensive language. Referring to the way of identifying different transfer learning techniques in Pan and Yang [130] and Pikuliak et al. [139], we analyse cross-lingual approaches in all reviewed papers and ultimately decide to categorise them according to *"what to transfer"* – the level at which knowledge transfer or sharing occurs. The overall hierarchy of various cross-lingual transfer approaches is depicted in Figure 4. Cross-lingual transfer paradigms consist of three main categories:

- **Instance Transfer:** Instances are transferred on the data level between source and target languages, including texts or labels transfer.
- **Feature Transfer:** Linguistic knowledge is shared or transferred on the feature level between source and target languages.
- **Parameter Transfer:** Parameter values are transferred between language models. This effectively transfers the behaviour of the model from source to target languages.

We summarise studies based on transfer levels in Table 4 and model types in Table 5. Additionally, a comprehensive Table is provided online[13] to summarise all techniques utilised in these studies.

[13]https://github.com/aggiejiang/crosslingual-offensive-language-survey

Table 5. Types of Models Used in Studied Papers

Model Type	Model	Reference
Machine Learning	NB	[12, 29, 82, 153]
	LR	[7, 8, 13, 46, 132, 180]
	SVM	[21, 82, 87, 126, 129, 132, 153]
	other	[74]
Deep Learning	MLP	[87, 95, 132, 134]
	CNN	[21–23, 52, 87, 88, 108, 111, 132]
	LSTM	[12, 13, 21–23, 29, 34, 87, 90, 95, 108, 123, 126, 127, 129, 132, 138, 168, 174, 183]
	other	[13, 18, 42, 46, 87, 95, 123, 180]
Transformers	BERT variants	[7, 11, 13, 26, 76, 81, 88, 108, 120, 126, 131, 132, 142, 171, 180, 193, 197, 198]
	mBERT	[7, 8, 16, 22, 23, 26, 29, 39, 42, 46, 54, 70, 78, 87, 114, 120, 121, 127, 134, 135, 149, 153, 171, 172, 177, 193, 197]
	XLM variants	[38, 76, 114, 158, 159, 174, 197]
	XLM-R	[13, 16, 42, 54, 60, 67, 70, 73, 76, 78, 87, 114, 115, 121, 131, 142, 146–149, 161, 162, 172, 177, 198]
	other	[39, 73, 76, 78, 81, 88, 114, 131, 135, 142, 153]
LLMs	FLAN-T5, mT0	[142]

Multilingual PLMs are kept in Transformers and others are grouped into *variants.

6.1 Instance Transfer

Instances in the cross-lingual HS task are comprised of two elements (texts and corresponding labels) in both source and target languages. Although the single source language data is not directly applicable in cross-lingual transfer, there are certain parts of the instance that can still be repurposed together with target language data via instance transfer. Instance transfer (or instance projection) revolves around the transfer of specific data elements (either text or label information) between source and target languages, which is a key technique in the realm of cross-lingual HS detection.

To implement instance transfer, a correspondence between instances in the source and target languages must be created on the data level. **Correspondence** refers to the pair of instances with the same meaning of texts or identical labels [139], e.g., parallel data with the same label matching, and translated text retains the same label of its original counterpart. Based on the established correspondence, texts or labels can be transitioned from one language to the other. Subsequently, these projected texts or labels can be used for further training and fine-tuning stages.

Pairs of instances in source and target languages that share identical labels are referred to as corresponding instances. All data-based transfer techniques need corresponding texts/labels or a mechanism to generate them. Techniques for instance transfer are introduced distinctly at the text and label levels. For label-based transfer, labels from a source language are projected onto target-language data (e.g., via annotation projection or pseudo labelling), without transferring the source text. In contrast, text-based transfer is to transfer or generate text across languages (e.g., via machine translation or text alignment), preserving the original labels and creating new training instances in the target language. These two strategies differ in which component of an instance is projected across languages, but both establish correspondence between source and target data.

6.1.1 *Label-Based Transfer.*

Annotation Projection. It leverages parallel corpora to transfer annotations from a source to a target language. Parallel corpora consist of sentence-aligned source-to-target data, ensuring higher quality than machine-generated translations. Since data is paired between source and target languages, labels in the source language can be directly projected onto the target language. For example, if a sentence in the source language is labelled as hateful, its corresponding text in the target language can be inferred to the same label. It can effectively create labelled data for the target language without manual annotation, especially when building datasets for languages where annotating from scratch might be challenging due to linguistic complexities or a lack of annotators.

Pseudo Labelling. It is a semi-supervised learning technique, where a model trained on labelled source data is used to make predictions on unlabeled target data [102]. High-confidence predictions, namely predicted labels with high probability scores, are treated as "pseudo-labels" for the unlabelled target data. By using high-confidence predictions from a model trained in the source language, it is able to generate labelled target data in another language. This augmented target data can then be used to train or fine-tune models for offensive language detection in the target language. To generate pseudo labels, some studies use an ensemble-based approach to generate labels based on majority voting for unlabelled target datasets, and use bootstrapped target dadasets to further fine-tune the trained model with source training samples [21–23]. Hande et al. [78] directly use a pre-trained multilingual language model (i.e., XLM-R [37]) to predict the pseudo labels. Zia et al. [198] firstly fine-tune a pre-trained XLM-R on gold-labeled source language data, and then use it to create a new psuedo-labelled dataset in the target language.

6.1.2 *Text-Based Transfer.*

Machine Translation. In the absence of sufficient annotated parallel data resources, machine translation provides an efficient alternative strategy to achieve text-level transmission. In general, it leverages translation tools [126, 127] or models [70] to translate labelled datasets between source and target languages, thereby augmenting the training data. Labels for translated data are the same as those in the original labelled dataset. Translation can be either directed or undirected: (i) target to source [7, 8, 12, 26, 82]; (ii) source to target [70, 162, 172]; (iii) translate both source and target to each other [39, 87, 129, 171, 193]. In addition, back-translation is commonly used as a data augmentation technique in cross-lingual scenarios by using various transformations on the original source data to create new samples for further model training [124]. Back-translation refers to translating a source sentence to a target language and then reverting it back to the original language, aiming to generate paraphrased instances of the original texts.

Text Alignment. Similar to machine translation, existing mapping techniques between languages are employed to generate labelled data for the low-resource language. Shi et al. [168] begin by using a shared space between two language vectors to identify the most similar sentence in the target language to the labelled data in the source language. The identified sentences are then treated as labelled data for the target language, and assigned the same labels of the source data. Ryzhova et al. [161] augment training datasets by generating attacked source datasets. They apply an adversarial attack algorithm [89] to the source sentence by only replacing words that the model considers important, ensuring that the new sentence is semantically similar to the old one.

6.2 Feature Transfer

Feature transfer transforms linguistic features to aid in HS detection. Rather than translating source texts or projecting source labels, they focus on extracting salient linguistic features from source and target languages, and aligning them into a shared feature space, retaining the essence of the

text consistent across languages. **Cross-lingual Word Embeddings (CLWEs)**, also known as **Multilingual Word Embeddings (MWEs)**, are commonly used for feature transfer. They train monolingual word embeddings (WEs), such as Word2Vec or FastText, on multiple languages, yielding vectors that capture semantic similarities relying on shared representations between languages and thus enabling representations that are common across languages. In cross-lingual offensive language detection, various approaches utilise existing CLWEs or retrofit CLWEs. We discuss three main techniques: multilingual distributional representations, multilingual contextualised representations, and retrofitting WEs.

6.2.1 Multilingual Distributional Representations. Pre-trained distributional WEs are most commonly used as embedding layers in neural networks for different language data inputs. Among these, MUSE embeddings have emerged as a predominant choice. They are extensively utilised to extract multilingual features and are often integrated with both traditional machine learning models, such as Gradient Boosted Decision Trees [12], and advanced deep learning architectures (i.e. LSTM [11, 29, 126, 127, 129], BiLSTM [21–23, 87], CNN-GRU [7, 8, 46], Capsule Networks [87], Sluice networks [123]), and BERT [126]). Other notable multilingual distributional vectors, such as Multilingual GloVe and FastText, have also been employed in architectures like BiLSTM [34] and Capsule Networks [87]. In addition, Babylon multilingual embeddings have been paired with Sluice networks for multilingual HS detection [123]. Bansal et al. [18] integrate bilingual switching features into the **Hierarchical Attention Network** (**HAN**) architecture, enhancing the transfer knowledge for cross-lingual detection.

Furthermore, some studies have ventured into cross-lingual projection using the MUSE mapping method, aligning monolingual FastText embeddings in English and German to produce their own CLWEs [21–23]. Paul et al. [132] align English and Hindi FastText embeddings by performing **Canonical Correlation Analysis** (**CCA**),[14] and project them into a shared vector space where they are maximally correlated. Kapoor et al. [90] use multilingual datasets to train embeddings in order to capture specific distributional representations of tweets. De la Peña Sarracén and Rosso [42] propose a graph auto-encoder framework to learn embeddings of a set of texts in an unsupervised way, adding language-specific knowledge via **Universal Sentence Encoder** (**USE**).

6.2.2 Multilingual Contextualised Representations. Multilingual contextualised representations, typically derived at the sentence level from PLMs (such as mBERT [47] and XLM-R [37]), can capture deeper semantic and contextual relationships between words and phrases. Some researchers have leveraged hidden features with richer semantic information from the first embedding layer of mBERT [11, 127, 153], DistilmBERT [78, 183], XLM [174], XLM-R [78] as feature extraction layers. These embeddings are either extracted as standalone embedding layers [11, 127, 153, 183] or utilised as frozen embedding layers [78, 174] in deep neural networks, such as BiLSTM, yielding enhanced performance over traditional distributional embeddings. Multilingual sentence embeddings LASER have also been used across various models, including Support Vector Machine [153], Logistic Regression [7, 8, 46, 127], Random Forest [11, 153], XGBoost [11], Multi-Layer Perceptron [134] and LSTM [183]. Rodríguez et al. [153] propose to use LabSE and mBERT representations with SVM-based and tree-based classifiers, achieving improved performance.

6.2.3 Retrofitting Word Embeddings. Retrofitting pre-trained WEs by infusing multilingual domain knowledge stands as a promising strategy to amplify the semantic relationships across languages, thereby enhancing the efficacy and scalability of models in cross-lingual HS detection. Pant and Dadu [131] employ a supervised FastText model trained on the sarcasm detection corpus [176] to improve the identification of HS content in English and Hindi languages. Arango et al.

[14] https://www.mathworks.com/help/stats/canoncorr.html

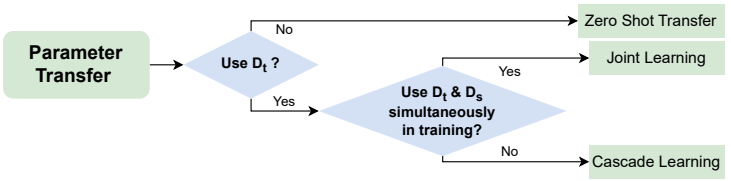

Fig. 5. Different scenarios in parameter transfer for automated detection of cross-lingual HS.

[11] construct HS-specific WEs by aligning monolingual Word2Vec embeddings using hate-specific bilingual dictionaries (such as HurtLex), which are able to capture non-traditional translations of words between languages. Additionally, Hahn et al. [74] learn semantic subspace-based representations to model profane languages on both word and sentence levels. Jiang and Zubiaga [88] propose a cross-lingual domain-aware semantic specialisation system to construct sexism-aware WEs. They retrofit pre-trained FastText word vectors by integrating in-domain and out-of-domain linguistic knowledge (such as lexico-semantic relations) into the specialised feature space.

6.2.4 Infusing Additional Features. Additional external features are commonly used in identifying HS across languages, such as domain-specific [74, 129], language-specific [11, 197], and typographic features [95, 108]. Specifically, domain-specific features are essential to assist the hate-related knowledge transfer. The multilingual lexicon, HurtLex, has been employed in multiple studies [87, 126, 127, 129]. Some researchers also utilise hateful word pairs to construct domain-aware or semantic-based representations across languages [11, 74, 88], due to the lack of hate-specific patterns in general-purpose multilingual embeddings. In addition, infusing language-specific features can help integrate linguistic and cultural knowledge for geographically sensitive tasks. It can be in different forms, such as language switching pattern matrix [18], bilingual pairs [88], cross-cultural similarities [11], and social dynamics among users [121]. This feature can provide a richer understanding of cultural dimensions in hateful content [197]. Furthermore, typographic features provide a language-agnostic and domain-agnostic perspective to help identify online hate due to their shared meanings across languages, which includes capitals, punctuations, and emojis [38, 95, 108]. Stylometric features also remain persistent with respect to language variations, as the writing style of toxic content can be correlated to emotional profile of social media users [108].

6.3 Parameter Transfer

The performance of NLP models is dependent on model parameters. Parameter transfer assumes that prior distributions of parameters are transferred between different languages within one model or between individual models. Most parameter transfer approaches rely on inductive transfer learning in cross-lingual offensive language detection [130]. Inspired by Pikuliak et al. [139], we outline three distinct scenarios for surveyed studies, depending on how the source and target data are utilised during the process of parameter transfer, namely zero-shot transfer, joint learning, and cascade learning (see Figure 5). Only source data is used for training in zero-shot transfer, while both source and target data are applied during the training process in joint learning and cascade learning. Joint learning utilises both simultaneously to train the model, but cascade learning leverages them in the different stages of training.

Parameter-level transfer strategies can be sophisticated and diverse in terms of their model architectures and training procedures, incorporating one or more of the transfer scenarios above. Table 6 has summarised the possible correlations between these transfer strategies and their application scenarios, and will discuss more details in the following subsections regarding scenarios and strategies utilised on parameter-level transfer.

Table 6. Correlations between Scenarios and Strategies in Parameter Transfer for Automated Detection of Cross-Lingual HS

Transfer Strategy	Zero-Shot Transfer	Joint Learning	Cascade Learning
Multilingual PLMs	✓✓	✓	✓✓
Domain Adaptation		✓	✓✓
Multi-source Training	✓✓	✓✓	✓
Multi-channel Learning	✓	✓✓	
Multi-task Learning	✓	✓✓	✓
Meta Learning		✓	✓
Ensemble Learning	✓	✓✓	✓

✓✓means higher frequency than ✓.

6.3.1 Transfer Scenarios. We discuss three different scenarios of parameter transfer.

Scenario 1: Zero-Shot Transfer. It is a promising scenario when labelled data in the target language is scarce or lacking, by enabling detection in a language not seen during training. The model is trained on labelled source data from one or more languages not including the target language, and is then tested on the unseen target language. This is sometimes called direct or model transfer. We found 25 papers under this transfer paradigm, including subtypes:

— *Single Source to Single Target:* Considered a type of domain adaptation (see Section 6.3.2), models are trained on labelled data from a single source language, then used on an unseen target language [39, 54, 60, 67, 70, 73, 114, 120, 121, 134, 142, 147, 148, 153, 177, 197]. It is the most frequently observed (62.5%) of zero-shot transfer strategies in our survey.

— *Multiple Sources to Single Target:* Models are trained on labelled data from multiple source languages (excluding the target language) and then evaluated on a single target language [7, 114, 161]. It is a type of multi-source training strategy where the combination of training data excludes the target (see Section 6.3.2).

— *Pseudo-Target Augmented Training:* The source data is translated into the parallel target data via machine translation, and then models are trained on both the original source data and the translated target data together [174] or in parallel [87, 126, 127, 129]. It falls under the multi-channel learning strategy when using both source and translated target data in parallel[15] and fusing them before the output layer of the model. This allows the model to learn nuances specific to the target language from the pseudo content.

— *Parameter Frozen:* A model is trained on source data first, and then a subset of its parameters (e.g., weights from specific layers) is saved to initialise another model for target language evaluation [67, 138, 146, 149]. It is a type of domain adaptation,[16] and captures language knowledge from certain layers of the source model to provide a head start for the target model, potentially leading to faster convergence and better generalisation.

It is worth highlighting that these methods rely on multilingual PLMs, especially often used for Single Source to Single Target and Multiple Sources to Single Target scenarios. mBERT and XLM-R are two of the most prominent multilingual PLMs in surveyed papers, because of their popularity, widespread adoption and SOTA performance across multilingual NLP tasks. They together account

[15]Pseudo-Target Augmented Training has overlaps with the joint learning scenario. It distinguishes itself by using pseudo target texts for training instead of real target texts.

[16]Parameter frozen in zero-shot transfer scenario shares similarities with the cascade learning scenario. However, while cascade learning continues to fine-tune the trained model with target data, zero-shot transfer directly applies the trained model to test the target data without further fine-tuning.

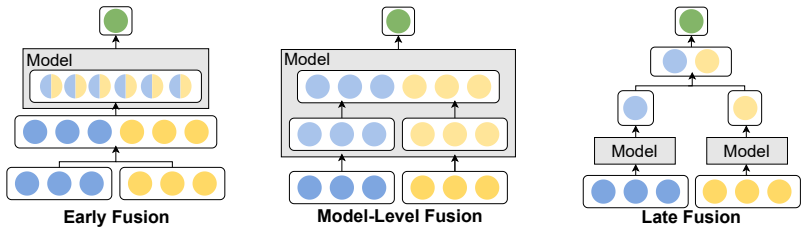

Fig. 6. Different fusion stages in the joint learning scenario.

for 80.8% of papers in the zero-shot transfer scenario. Additionally, some papers release domain- and language-specific PLMs in a multilingual setting [73], choose other multilingual PLMs like XLM [174] and language-specific MuRIL [39, 73], or combine MWEs with monolingual neural models like LSTM [87, 129] and BERT [126].

Scenario 2: Joint Learning. It typically involves training a model between languages based on multiple channels, tasks or objectives simultaneously, or jointly assimilating multiple types of data or modalities. The model might have some parameters that are shared across languages as well as some that are specific to each, so it is also called *parameter sharing* [139]. The transfer of knowledge happens only on shared parameters. Changes to the shared parameters can affect each other between languages. They are interrelated and can provide complementary information.

In this scenario, both source and target data are used at the same time during the training stage. Parameters can be shared between source and target channels within one model or between individual source and target models. We found 23 papers in this group. The distinction among these strategies mainly lies in the stage of knowledge transfer across languages and the extent of parameter sharing (some or all parameters shared) [95]. Based on these criteria, we categorise them into three primary fusion stages (see Figure 6):

— *Early fusion:* A combination of datasets or embeddings from both source and target languages, with the entire model sharing parameters. It is commonly adopted in few-shot learning [39, 60, 67, 73, 174, 197] and multi-source training [34, 39, 46, 52, 73, 108, 120, 147, 162, 180], where datasets from the target language are incorporated during training. In few-shot learning, the model is trained by utilising a very small amount of target samples combined with a more extensive source dataset, while multi-source training typically takes a combined dataset of source and target languages or integrates multiple datasets including multiple non-target datasets together with the target dataset for training.
— *Model-level fusion:* A concatenation of high-level hidden features within a model, and only specific layers or components of the model share parameters, facilitating a more nuanced transfer of knowledge. It commonly occurs in parallel model architectures, such as multi-channel learning [95, 171], multi-task learning [76, 123] and meta-learning [16, 115].
— *Late fusion:* A combination of different channels or models before the output layer, and only the parameters of the output layer for predictions are shared, ensuring a final consolidation of knowledge from various languages, often used in multi-channel architectures [82].

Early fusion is the most prevalent strategy due to its straightforward integration of source and target data for better generalisation, while late fusion provides a modular approach, allowing for the independent training of individual models that can later be merged. However, unlike early fusion, model-level fusion effectively overcomes the curse of high dimensionality and the synchronisation demands among diverse features. Moreover, it ensures that interactions between different languages are not isolated as in late fusion [95], showing its robustness and resilience.

Scenario 3: Cascade Learning. In the cascade learning scenario, both source and target data are used at different times during the training. First, an existing PLM is directly used or a new model is trained with source data, and then the trained model is fine-tuned with target data, where subsets of parameters are shared between different stages of training and fine-tuning. Different strategies for training and fine-tuning have been used in 19 of our surveyed papers.

Many use existing SOTA pre-trained multilingual models serving as the foundation,[17] while some studies train a model with one or multiple source datasets. Subsequently, the fine-tuning stage in cascade learning is applied across an entire [8, 60, 81, 90, 135] or only a limited subset of target data (i.e. few-shot learning) [135, 147], optimising the model adaptability. Röttger et al. [158] indicate that an initial training phase on source data (English) could increase model performance when there is little fine-tuning data in the target language, where source data can partly substitute target data in few-shot settings. Some researchers delve into consistent fine-tuning [16, 115, 161] or iterative fine-tuning (i.e. meta learning) [16, 115], performing multiple rounds of fine-tuning on the PLMs to further enhance its ability to transfer knowledge.

Additionally, given the limited availability of target data, data augmentation techniques have emerged. They aim at enriching the target dataset to enable fine-tuning, by enhancing label diversity or expanding the text corpus. Some studies have attempted to generate pseudo labels for unlabelled target datasets [21–23, 198] (see Section 6.1.1), while others augment and refine pseudo texts by exploring advanced techniques, such as word alignment to project samples [168], adversarial algorithm to generate attack samples [161], and domain-specific target data filtering [70].

6.3.2 Hybrid Transfer Strategies. According to the utilisation of source and target data, these strategies can be broadly covered in one or more parameter transfer scenarios mentioned in Section 6.3.1. That is, these strategies are not mutually exclusive. A single scenario can seamlessly integrate multiple transfer strategies (see Table 6).

Multilingual Pre-Trained Language Models. These models are regarded as a cornerstone of parameter-level transfer techniques for low-resource scenarios, especially in cross-lingual offensive language detection. They are trained on large-scale multilingual corpora, learning commonalities and differences between languages. Therefore, a single PLM can share the same underlying sub-word vocabulary and semantic representations across languages [47]. Multilingual PLMs can be initialised as the base model with pre-trained parameters, then fine-tuned on available HS data in all three scenarios of parameter transfer (zero-shot transfer, joint learning and cascade learning) [60, 120]. Beyond this, they can also employed in the other two transfer levels (data and feature). They are able to augment target datasets by predicting pseudo labels for unlabelled data and generating synthetic target texts. And they can be used as multilingual representations, where subsequent models are constructed based on these source representations [153].

Domain Adaptation. It involves taking a model trained (or pre-trained) in the source language and adapting its parameters to initialise a new model for the low-resource language, aiming to leverage source data to improve performance on target data [70, 101]. Its advantage is that it can capture offensive language knowledge and linguistic similarities between the languages within the model's parameters. In the zero-shot transfer scenario, parameters of the trained model are entirely or partially used on the target data for prediction [67, 149]. In addition, cascade learning adopts a more subtle approach. That is, the entire or specific subsets of the model's parameters (like specific layers) are frozen to retain their original state [147], while others are fine-tuned to better align with the target data, often using a smaller target dataset (as few-shot learning) [135, 158].

[17] Cascade learning fine-tunes PLMs using target data, while zero-shot transfer fine-tunes PLMs using non-target data.

Multi-Source Training. This approach often entails the amalgamation of labelled data from multiple source languages and potentially the target language in diverse ways. By synthesising information across multiple languages, this strategy is able to capture the diversities and commonalities of linguistic patterns, demonstrating significant efficacy in the cross-lingual setting. It is extensively utilised across all three transfer scenarios, serving as an augmented cross-lingual learning approach during different stages of training or fine-tuning. In multi-source training, the most prevalent way is to train or fine-tune the model on multiple source languages. This can be executed in the training phase of joint learning [34, 39, 46, 52, 76, 108, 120, 159, 162, 180, 193] or the fine-tuning phase of cascade learning [73, 147] in conjunction with the target language, or in isolation from the target language for zero-shot transfer [114, 161]. Deshpande et al. [46] offer a more focused perspective by narrowing the linguistic scope to specific language families. By training models on languages within one language family, they then assess the model's performance on each member of that family, revealing the interplay of linguistic knowledge within related languages. It is worth noting that merging training examples from different languages can be detrimental to cross-lingual performance in cases where offensive language domains are too distant [70].

Multi-Channel Learning. This strategy typically refers to the use of multiple types of input representations or sources concurrently for a single task, where multiple embeddings or pre-processed versions of the text are often channelled as parallel input streams. In this architecture, each channel processes the input independently through parallel layers. These independent processing channels are eventually combined based on their outputs for subsequent learning or classification stages. Such a design ensures that diverse input representations are holistically integrated, capturing a broader spectrum of linguistic knowledge across languages. Multi-channel learning finds its applications mainly in the zero-shot transfer and joint learning scenarios. When applied to cross-lingual offensive language detection, this approach often treats datasets from different languages as distinct input channels. Each channel adopts different neural networks or PLMs based on diverse input languages, and these channels are then merged into the final classification stage [95, 171, 174]. A significant improvement to this strategy is the incorporation of machine translation. By translating labelled source data to pseudo target data, low-resource target data is enriched and a bridge is constructed between source and target languages, enhancing knowledge relationships and a more seamless transfer of languages [82, 87, 126, 127, 129].

Multi-Task Learning. A joint learning environment training a model on several interrelated tasks simultaneously, where the proficiency learned from one task can boost performance on another. A key point of this strategy is the shared representation of hidden features across specific layers of the model. Such shared layers enable different tasks to mutually benefit from common representations, while also preserving their distinctiveness through task-specific output layers. The training process is holistic, aiming to optimise the model's performance across all tasks. This is typically achieved by incorporating the loss functions from each task. Multi-task learning emerges as a powerful method for cross-lingual offensive language detection within the parameter-level transfer techniques across three transfer scenarios. Models can be trained on offensive language detection and a wide range of related auxiliary tasks, such as Sentiment Analysis, **Named Entity Recognition** (**NER**), Dependency Parsing, and **Part-Of-Speech** (**POS**) Tagging [76, 114, 123]. Additionally, they can be fine-tuned on data in various offensive categories, such as aggressive content [146, 149]. This multi-task training strategy enables the model to identify and process offensive content with higher accuracy and enhances the model's ability to grasp linguistic nuances across languages.

Meta Learning. This strategy, often referred to as "learning to learn", has become a popular few-shot learning technique in NLP. It involves training models to learn the optimal initialisation of

parameters, allowing them to be fine-tuned with a small amount of data from the target language. This technique finds its application within the joint learning and cascade learning transfer scenarios. It enables rapid adaptation to new unseen languages, making it especially useful in scenarios where labelled data is limited. The main meta learning methods include optimisation-based [59] and metric-based techniques [94]. Two papers in our review employ optimisation-based meta learning framework. Montariol et al. [114] study meta learning for the problem of few-shot HS detection in low-resource languages. They propose a cross-lingual meta learning-based approach based on optimisation-based **Model-Agnostic Meta-Learning** (**MAML**) and Proto-MAML models, fine-tuning the base learner XLM-R with parallel few-shot datasets in different target languages. Awal et al. [16] propose HateMAML, a model-agnostic meta-learning-based framework that uses a semi-supervised self-refinement strategy to fine-tune a better PLM for unseen data in target language, showing effective performance for HS detection in low-resource languages.

Ensemble Learning. It leverages multiple models or "learners" to collaboratively make decisions. Instead of relying on the output of a single model, ensemble learning combines all predictions from multiple models to produce a final prediction based on majority voting [108, 132]. It takes advantage of the different strengths of each model, mitigating the weaknesses and biases of any single model and avoiding overfitting, improving robustness and reliability over a single classifier. This is particularly beneficial in cross-lingual settings, where language biases can pose challenges to individual models. In cross-lingual offensive language detection, ensemble learning is mainly applied in the zero-shot, joint learning, and cascade learning transfer scenarios. Deep ensemble models have emerged as the predominant choice in recent studies [21–23, 132], while the integration of PLMs within ensemble learning frameworks is also common [29, 108, 161]. Each model in the ensemble can be trained on different languages or linguistic features, allowing the ensemble model to capture diverse linguistic patterns and reduce test errors for target languages [161].

6.4 Summary of Cross-Lingual Approaches

A review of the surveyed work reveals several key developments in cross-lingual offensive language detection. First, there has been a **rise of multilingual PLMs**, beginning with general-purpose PLMs such as mBERT and XLM-R, which quickly replaced feature- or translation-based methods as the dominant approach [37, 47]. Meanwhile, there is a move **from general-purpose PLMs to more language- and culture-specific PLMs**, such as MuRIL for Indic languages [39], AraBERT for Arabic dialects [81], and XLM-T trained on social media [114]. More recently, a new trend has emerged with the rise of **large language models (LLMs)**, capable of zero- or few-shot multilingual tasks via prompting [28, 122, 142]. See future challenges and directions in PLMs and LLMs in Section 7.4. Second, we observe a **shift from feature-based to parameter-based transfer**. Earlier approaches relied on CLWEs, lexicons, or projection methods [11, 129]. With the rise of multilingual PLMs, parameter-based fine-tuning has become the core transfer mechanism, while handcrafted features now play a complementary role. Moreover, there is a **growing focus on low-resource languages and dialects**. While early research centred on English and a few European languages, more recent work increasingly covers Indic [39], Arabic [81], and Asian languages [85, 86]. However, many languages still lack dedicated datasets and evaluation benchmarks.

Model Distributions. After reviewing diverse models and their frequency of use in recent studies, we can see traditional machine learning-based models have been employed 24 times, while deep learning-based models have seen a slightly higher usage (38 times) due to their complex structure and learning capacity from a large amount of data. Transformer-based models have been overwhelmingly preferred (99 times), highlighting their growing dominance in NLP. Large Language Models have also recently been investigated in this area due to their excellent emergent ability.

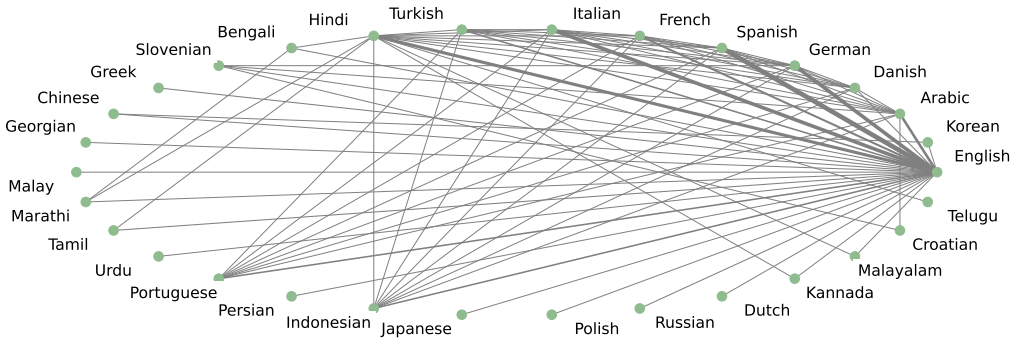

Fig. 7. Synergies between languages. A link indicates that both languages have been used in a model, as source or target language, or to learn multilingual feature spaces. Thicker lines for higher frequencies.

Evaluation Metrics. Macro F1 is the most used metric given class imbalance in datasets, while accuracy is also popular. Some studies also consider precision, recall, and weighted F1 score. Eronen et al. [54] propose a new linguistic similarity metric based on the **World Atlas of Language Structures** (**WALS**) [50] to select optimal transfer languages for automatic HS detection.

Source Code Availability. Out of 67 papers reviewed, 26 (38%) made their source code available, whereas the remainder 41 (62%) do not, hindering transparency and reproducibility.

Language Pairs. We present the frequency of language pairs in Figure 7. English prevails as the source language, while the most frequent target languages are Spanish, German, and Italian. These languages share typological similarities with English, which likely facilitates their frequent pairing in cross-lingual tasks. This emphasises the impact of data availability and language similarity on the training and application of cross-lingual methods, and also reveals the limitations resulting from data scarcity. Many studies have also explored cross-lingual learning by transferring English to Arabic, Hindi and French, as these languages have a relatively large amount of datasets available (see Figure 3). Additionally, there are studies working on the transfer between non-English languages, among which Danish, Indonesian, Portuguese, and Turkish are frequent.

Comparative Strengths and Tradeoffs. While each transfer level offers viable solutions, their effectiveness can vary depending on the task setting and language pair characteristics. For example, **instance transfer** is often most effective when high-quality parallel corpora or reliable translation tools are available, enabling direct label projection or text translation. It tends to work well for typologically similar languages or when cultural and topical overlap is high. **Feature transfer** is particularly useful when parallel data is scarce but high-quality multilingual representations exist; it can capture shared semantic spaces while retaining language-specific nuances. This approach is generally more robust to moderate typological distance, but may struggle when the target language is under-represented in pre-training corpora. **Parameter transfer** offers the best flexibility across settings, but its success depends heavily on the pre-trained model coverage of the target language and the domain similarity between training and testing data. Parameter sharing in multilingual PLMs often generalises best within the same language family, while cross-family transfer may require additional adaptation (e.g., domain-specific fine-tuning or language-specific modules). In practice, selecting an appropriate transfer strategy requires balancing data availability, typological similarity between languages, and computational resources. Hybrid approaches that combine strategies can often yield the most reliable performance across diverse language families, such as using instance transfer for initial alignment and parameter transfer for fine-tuning.

7 Challenges and Future Directions

In this section, we examine key challenges encountered in cross-lingual offensive language detection in three aspects (language, dataset and approach), and then outline potential future directions.

7.1 Language-Related Challenges

7.1.1 Low-Resource and Underrepresented Languages. A central challenge is the limited availability of resources for many non-mainstream languages. As illustrated in Figure 7, the majority of studies focus on English as the source language and transfer primarily to typologically similar European languages (e.g., Spanish, German, French). In contrast, languages from other families (particularly African, Indigenous, and smaller Asian languages) remain underrepresented in both datasets and modelling efforts. This imbalance reflects broader trends in multilingual PLMs, where training corpora disproportionately favour high-resource languages [121, 158]. As a result, CLTL systems often exhibit strong performance for well-represented languages but much weaker transferability for low-resource or distant languages [46, 54]. The scarcity of labelled data, the uneven distribution of languages in PLMs, and the difficulty of transferring knowledge across typologically distant families represent major obstacles.

7.1.2 Diverse Linguistic Structures. Language diversity encompasses various aspects from basic grammatical rules to complex nuances unique to each language. *Linguistic distance* captures how typologically similar or different two languages are. Transfer tends to be more successful across languages within a certain family that shares vocabulary, morphology, or syntactic patterns (e.g., English–German, Spanish–Italian), while performance often degrades for distant language pairs (e.g., English–Swahili, English–Chinese) [46, 54]. Figure 7 also demonstrates that cross-lingual experiments are concentrated within Indo-European languages, while distant language pairs are far less represented. This indicates that challenges for CLTL are not only the scarcity of data in low-resource languages, but also the uneven ability of models to generalise across varying degrees of linguistic distance. Furthermore, the presence of *dialectal forms* within a single language amplifies its complexity. For example, the Arabic language is spoken by a wide range of countries across Asia, with complexities arising due to its multiple dialects [81].

7.1.3 Code Mixing. Non-English social media content is often characterised by code-mixing, where users blend languages, use transliterations, or incorporate multiple scripts within a single sentence or conversation, such as English-Hindi [78, 82] and English-Chinese [88]. The English-Hindi (or Hinglish) combination appears in many cases, because the unique policies on regulating such speech in these regions add layers of complexity and make detection more challenging [90].

7.1.4 Cultural Variations. Culture is multifaceted and complex. The perception of offensive language can differ significantly across languages and cultures [121]. What is considered non-offensive in one culture or language might be misinterpreted as signals of offense in another, leading to detection discrepancies. Even within a single language, there can be cultural diversity. For example, categorising English as a representative of "western cultural background" might ignore the differences between American and British cultures [197]. Such cultural variations can hinder the effective transferability of language models [159]. Moreover, hateful expressions often employ figurative language, rhetorical figures and idioms, which are also language-dependent. This requires models to interpret underlying intent rather than relying solely on literal meanings [127].

7.1.5 Explicit vs. Implicit Offensive Language. Offensive language can be expressed explicitly or implicitly. Explicit abuse, such as direct insults or slurs, often transfers more easily across languages because it relies on surface lexical markers. In contrast, implicit abuse (expressed through sarcasm, stereotypes, or code-mixed language) is far more context-dependent and culturally situated. As

a result, implicit forms are much harder to transfer from high-resource to low-resource settings, where contextual cues may differ significantly [79].

7.2 Dataset-Related Challenges

7.2.1 Inconsistent Definition of Offensive Language. The diversity of HS encompassing from misogyny and racism to other forms of discrimination often leads to discrepancies in research [121]. This has a bigger impact on low-resource languages [21–23], and introduces ambiguities that challenge the task, especially in multilingual scenarios [183]. The lack of definition consistency across datasets in turn limits cross-lingual research, because many datasets in either source or target languages may be incompatible [22].

7.2.2 Data Scarcity and Quality. **Limited Labelled Datasets.** Low-resource languages often do not have the attention that major languages receive, leading to limited data availability [29] and limited studies (see Figures 3 and 7). Additionally, creating a high-quality labelled dataset requires extensive manual effort, which can potentially lead to high annotation and time costs as well as be harmful to annotators [158]. Annotating offensive language is even more challenging due to the need for cultural and linguistic knowledge. Ethical and privacy concerns also arise when dealing with real-world data from social media platforms or other online sources [86], because there are potential risks of exposing sensitive information or inadvertently promoting offensive language when studying it or trying to combat it.

Dataset Imbalance. Given that the majority of social media content is non-hateful, labels are typically skewed towards the non-hate label [22, 23, 172]. This skewness can lead to training issues, especially when working with small training corpora. In addition, when considering multilingual or merged datasets, inter-language imbalance is another challenge [46]. Such datasets often display significant differences in the number of examples available for each language. This imbalance can bias models towards the semantic tendencies of languages that are overrepresented.

Dataset Bias. Biases such as topic, authorship, political affiliation, social media platform, and collection period can be one of the major issues and lead to potential generalisation challenges for cross-lingual models. Among them, topic bias has been taken seriously in many works [12, 158, 159, 183]. While some datasets include general offensive language or HS, others might concentrate on specific issues such as immigrants, misogyny, politics, or religion. Such topic-specific biases across datasets can harm model performance in cross-lingual classification [127, 174]. Besides, the temporal aspect of data collection, affected by events in different periods, can further introduce bias, leading to datasets with varied topical foci [174].

Data Source Obsolescence. As online language rapidly transforms, especially influenced by moderation and real-world events, datasets can quickly become obsolete [5, 183]. This obsolescence is further exacerbated by the emergence of new slang, metaphors, and colloquialisms that vary across languages and regions. For instance, terms that normally have neutral meanings (such as donkey) may be used offensively in sexist text [88]. The constrained and informal style of tweets, which are more like oral expressions than written language, complicates the preprocessing steps and leads to erroneous predictions [8, 126]. Emojis, associated with various forms of online harassment, further pose unique challenges [38]. Moreover, the domain and target of HS can shift significantly over time [61]. Real-world events, from local incidents to global crises like the COVID-19 pandemic, can give rise to new domains and terms of HS [114]. Then, it becomes increasingly challenging to create and continually update datasets customised for every possible language and domain, leading to frequent low-resource issues.

7.2.3 Annotation Bias. Offensive language datasets can exhibit systematic gaps and biases in their annotations [159]. First, different labelled datasets often utilise distinct annotation frameworks

or strategies [64]. Second, the subjective nature of offensive language leads to tentatively disputed annotations among human annotators, where hateful content could be interpreted differently depending on annotators' understanding [39]. The annotation of these contents is easily influenced by individual sociodemographic and cultural backgrounds, often resulting in low inter-annotator agreement [108]. Additionally, there is an "annotation dilemma" [8], whereby ambiguous instances might be annotated incorrectly by annotators but predicted accurately by the model.

7.3 Approach-Related Challenges

7.3.1 Limited Ability of Multilingual Pre-Trained Language Models. Cross-lingual models, especially multilingual PLMs, often lack generalisability and exhibit model bias [74, 127, 135]. The generalisability of transformer-based multilingual PLMs can be inconsistent, especially for typologically diverse languages, because they might be pre-trained with a highly focused set of languages or some languages with insufficient training samples [46]. This can lead to bias and instability in model performance, depending on the model architectures, topical focus, linguistic distance, or cultural context in datasets [115, 127]. For example, PLMs may exhibit systematic bias when handling dialects, minority languages with various linguistic distances, and content towards marginalised communities. Since such varieties could be under-represented in pre-training corpora, PLMs tend towards standardised forms of dominant languages. This can result in degraded model performance, erasure of dialect-specific expressions, or misinterpretation of community-specific discourse. Model overfitting across target languages is another important issue. Although multilingual models are less prone to overfitting on dataset-specific features than monolingual models, they will achieve poorer performance than monolingual ones in higher-resource settings, and they may require very different calibrations and adaptations across languages [158, 159].

7.3.2 Cross-lingual Transfer Performance. CLTL has emerged as a promising solution and has shown effectiveness addressing data scarcity [22, 171]. However, offensive language is deeply rooted in the specificity and diversity of language and culture, and its complexity poses significant obstacles to cross-lingual transfer, especially in zero-shot settings [114, 121]. Zero-shot transfer to multilingual training often suffers from performance deficiencies when compared with models trained on actual target language data [135]. It relies heavily on surface-level lexical and structural features, making them less effective when confronted with cultural contexts or implicit forms of abuse. This limitation has been consistently observed in empirical studies: zero-shot transfer yields significantly weaker results than target-language fine-tuning, even when only small amounts of labelled target data are available [135]. Moreover, cultural and linguistic variations across communities can further erode transferability, leading to biased or unreliable detection outcomes [121].

7.3.3 Limitations of Machine Translation. Machine translation can help augment datasets and alleviate data scarcity issues. A prevalent practice is to utilise Google Translate API.[18] [8, 87, 127, 193] However, the effectiveness of cross-lingual detection models is directly linked to the quality and accuracy of the machine translation. True semantics of the target text may change during the translation process. Machine translation may diminish the toxicity degree and thereby reduce the perception of offensive content, especially in context-sensitive situations [39]. Therefore, models that rely on such translations can experience performance degradation, especially when predicting instances with semantic shifts. However, despite translation errors and uncertainties, they can still be valuable supplementary inputs for text classification [171].

7.3.4 Poor Model Interpretability. While advanced deep learning-based models have shown good performance in detecting HS, they often operate as "black boxes" with a lack of transparency in the

[18] https://translate.google.com

decision-making process [8]. This opacity makes it hard for humans to understand the underlying reasons for the model performance in a particular language or scenario, and analyse the model errors [46, 183]. To bridge this interpretability gap, some tools are utilised to explain what models are doing, such as Local Interpretable Model-agnostic Explanations (LIME) [150] and SHapley Additive exPlanations (SHAP) [105], and theoretical approaches from cognitive linguistics are also applied, e.g. frame semantics [183]. A human-in-the-loop paradigm can also help by integrating human expertise into the model's learning process, but it can be difficult and uncertain for individuals to provide effective feedback to enhance the model performance [183].

7.4 Future Directions

This section outlines promising research directions and synthesises lessons from prior work.

7.4.1 Dataset Creation. A recurring finding is that **collecting even small amounts of target-language data** works well: fine-tuning with just a few hundred annotated examples consistently outperforms pure zero-shot transfer [158]. Conversely, relying solely on zero-shot transfer does not work well, as it leads to significant drops in performance for typologically distant or culturally distinct languages [135]. Datasets should therefore be designed to cover **multiple languages, dialects, and cultural contexts** [123], while avoiding narrow topical bias from single platforms [158]. Complementary benchmark resources such as XHATE-999 [70] and Multilingual HateCheck (MHC) [159] provide more reliable cross-lingual evaluation.

7.4.2 Data Annotation. Research suggests that **incremental and selective annotation** works well: iteratively expanding small subsets increases efficiency while avoiding redundant labelling [158]. In contrast, large-scale undirected annotation does not work well, as benefits quickly diminish relative to cost. To advance annotation quality, researchers should **recruit diverse annotators and provide culturally contextualised guidelines**, which mitigates annotation bias. Moreover, **active learning and semi-supervised strategies** have been shown to optimise annotator effort by focusing on ambiguous or uncertain cases [158]. Expert feedback is particularly valuable for subtle and implicit abuse [183], but annotator well-being and ethical safeguards must also be prioritised, providing appropriate training and support to avoid potential harm.

7.4.3 Integration of Additional Features. Studies consistently show that integration of additional features can also be effective. Adding **language-agnostic signals such as typographic features** (capitalisation, punctuation, emojis) works well [38, 95, 108]. Emojis, in particular, often carry stable meanings across languages and provide additional emotional cues that complement textual signals. **Stylometric features** such as function word usage, sentence length, or character-level distributions can also strengthen detection, since they capture writing style cues of toxic content that can be correlated with specific social media users [108]. Similarly, **domain-specific resources** such as HurtLex and tailored embeddings work well in capturing hate-related semantics across languages [88, 126]. For example, multilingual lexicons of offensive terms have been shown to boost recall by reducing false negatives, while domain-aware embeddings such as sexism-specific representations (SexWEs) provide finer-grained distinctions for subtle or implicit abuse. Relying only on general-purpose multilingual embeddings does not perform well, which often misses culture-specific or implicit abuse patterns [11]. They may capture general semantic similarity but struggle with figurative expressions, reclaimed slurs, or locally grounded derogatory terms. Over-reliance on generic embeddings risks reinforcing biases by ignoring context-specific linguistic cues. Furthermore, infusing **cultural features** underscores the significance of language-specific knowledge for geographically sensitive tasks. The creation of culturally-aware models, informed by multidisciplinary studies involving anthropologists and sociologists, can provide a richer understanding of cultural dimensions in hateful content [197]. Such work can produce different forms of

cultural features, such as switching pattern matrix [18], bilingual language pairs [88], cross-cultural similarities [11], and social dynamics among users [121].

7.4.4 Multilingual Pre-Trained Language Models. Multilingual PLMs can be applied to low-resource target languages without further training, thus bridging the data availability gap between high- and low-resource languages. However, their large number of parameters carries significant computational costs during training and fine-tuning [139]. Future research is geared towards optimising these multilingual PLMs for enhanced efficiency and interpretability. Additionally, two strategies are emerging to improve the generalisability and scalability of multilingual PLMs so that they can scale to handle large datasets or multiple languages simultaneously [12]. The first emphasises pre-training models using **data from relevant sources or language communities** (such as social platforms and hateful domains), as demonstrated by XLM-T [114] and AbuseXLMR [73]. This can mitigate model bias across language communities. The second strategy focuses on models based on **specific low-resource languages, dialects or those from similar language families**, as seen with MuRIL for Indic languages [39] and AraBERT for dialects [81], which requires inclusion of dialectal and under-represented language data in both training and evaluation. Meanwhile, it is important to **evaluate transfer effectiveness across both similar and distant language pairs**, rather than relying solely on high-resource or typologically similar examples. Otherwise, multilingual PLMs risk reinforcing existing performance gaps, performing well for Indo-European languages while neglecting typologically distant or underrepresented languages.

7.4.5 Cross-Lingual Training Strategies. Various cross-lingual training strategies have been proven effective. **Adaptive training techniques** can dynamically adjust to the specificities of languages and dialects, as well as bridge the gap between offensive language detection and PLMs. By leveraging meta-learning or few-shot learning, models can be trained to quickly adapt to new low-resource languages with limited labelled data [115]. Furthermore, collaborative **multi-task learning** can make models more generalisable in cultural variations and resistant to overfitting by training them and sharing representations on multiple auxiliary tasks simultaneously across languages or similar NLP tasks [114]. The use of **adversarial training**, where models are trained against adversarial examples, can also enhance their robustness and generalisation. This is especially important for offensive language, as attackers often use subtle language tricks to obfuscate and bypass detection [89, 161, 168]. Moreover, **cascade learning** can greatly bolster model performance and enhance out-of-domain generalisability by initially training models on source data or more diverse datasets and subsequently fine-tuning the results by simultaneously or progressively incorporating target language data [135, 158]. Besides, Vitiugin et al. [183] highlight that **human feedback** can be of great value to help detect subtleties of abuse and phrases during model training, Ranasinghe and Zampieri [149] explore **language-specific preprocessing** like segmentation in morphologically rich languages (e.g., Arabic, Turkish), and **multi-source training**, with augmented datasets, is also able to capture the diversities and commonalities of different linguistic patterns [34, 76, 108]. Future research should focus on systematically comparing these strategies and identifying which are most effective under varying resource and language-family conditions.

7.4.6 Application of Large Language Models. With the emergence of LLMs, the landscape of offensive language detection has rapidly shifted. Models such as GPT-3 [28], GPT-4 [122], **Fine-tuned LAnguage Net** (**FLAN**) [187] have demonstrated impressive multilingual capabilities, and recent surveys have highlighted their growing importance [196]. **Instruction-tuned LLMs with prompting** can already perform competitive detection across monolingual and multilingual settings without the need for task-specific fine-tuning [142]. Prompt design strategies, such as incorporating **task descriptions or input-label demonstrations**, further enhance performance [32, 75, 160]. LLMs also show promise in addressing data scarcity by **generating synthetic examples** for

low-resource languages [79], and in **providing explanations** that improve interpretability [185]. Despite their advantages, LLMs are not a full replacement for prior methods. They often underperform specialised multilingual PLMs on fine-grained tasks when little prompting expertise is available, and they remain costly to deploy at scale. Traditional approaches, such as leveraging domain-specific lexicons (e.g., HurtLex) or designing language-family-specific PLMs (e.g., MuRIL, AraBERT), still *work well* for targeted low-resource or domain-specific scenarios, offering more efficiency and control.

Overall, while early work on LLMs in this area was scarce at the time of our survey design, the rapid pace of development since then underscores their growing importance. Future studies should explicitly and systematically compare LLM-based approaches with prior CLTL techniques to establish where LLMs offer clear advantages and where CLTL methods remain more effective. Another important research direction is to explore the complementarity between LLMs and CLTL approaches. For example, lexicon features, linguistic distance, and culturally-aware representations can be integrated as prompts or external knowledge sources to guide LLMs, while smaller multilingual PLMs can be lightweight alternatives in resource-constrained environments.

8 Conclusion

This survey provides a detailed exploration of CLTL techniques in detecting offensive language online, Our survey is the first to specifically focus on cross-lingual detection, analysing a collection of 67 papers. We summarise the characteristics of multilingual offensive language datasets and cross-lingual resources. We further emphasise the diversity and complexity of CLTL techniques, highlighting the varying levels of transfer—instance, feature, and parameter– and popular strategies to improve performance. The survey also reveals several ongoing challenges in cross-lingual offensive language detection, such as linguistic diversity, dataset scarcity, and model generalisability. We also point out potential research opportunities, emphasising the need for new datasets across low-resource languages, as well as more robust, generalisable, and ethically sound CLTL approaches. Our survey highlights the importance of advancing CLTL approaches, which can serve as both a key reference for current practices and a guide for future research in this evolving field.

Limitations

While this survey provides a comprehensive synthesis of CLTL-based offensive language detection, several limitations remain.

- **Focused scope:** We focus exclusively on cross-lingual text classification for offensive language detection, excluding related modalities (e.g., multimodal detection) and other transfer scenarios such as cross-domain or cross-platform learning.
- **Coverage of literature:** Our selection presents a comprehensive review of the literature which is updated up to the cut-off point of July 2023.
- **Benchmark comparison:** Although we summarise performance trends, we do not conduct benchmarking experiments with reported results, which may vary due to evaluation setups.

Ethical Considerations

CLTL-based offensive language detection systems can raise several ethical considerations with the datasets, methods, and challenges we have discussed.

- **Bias in multilingual datasets:** Multilingual offensive language datasets often suffer from topic bias, imbalanced language representation, and inconsistent annotation guidelines, which can propagate through CLTL models across languages and communities. Balanced

and culturally diverse data collection, transparent reporting of dataset composition, and fairness-aware evaluation are needed.

— **Cultural sensitivity:** Offensive expressions are often context-dependent, and models trained on data from one cultural context may misclassify content in another, e.g., sarcastic expressions such as "well done" or idioms like calling someone a "clever fox" may be benign in one language but misinterpreted as neutral or offensive in another cultural context when translated. Culturally-aware annotation and evaluation with native speakers from target communities are necessary.

— **Model transparency and interpretability:** CLTL systems, particularly those based on PLMs, often act as"black boxes". This lack of interpretability can hinder error analysis. Techniques such as attention or post-hoc visualisations (e.g., LIME, SHAP), and clear model training and evaluation setups can improve transparency.

— **Responsible use and risk mitigation:** While CLTL-based systems can support harm reduction, they also carry risks of misuse for censorship, political suppression, or disproportionate moderation of certain language groups. Deployment should have clear governance frameworks, stakeholder engagement, and human-in-the-loop review mechanisms.

References

[1] Muhammad Pervez Akhter, Zheng Jiangbin, Irfan Raza Naqvi, Mohammed Abdelmajeed, and Muhammad Tariq Sadiq. 2020. Automatic detection of offensive language for urdu and roman urdu. *IEEE Access* 8 (2020), 91213–91226.

[2] Areej Al-Hassan and Hmood Al-Dossari. 2019. Detection of hate speech in social networks: a survey on multilingual corpus. In *Proc. of COMIT*, Vol. 10. 10–5121.

[3] Azalden Alakrot, Liam Murray, and Nikola S Nikolov. 2018. Dataset construction for the detection of anti-social behaviour in online communication in Arabic. *Procedia Comp. Sci.* 142 (2018), 174–181.

[4] Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017. Hate speech detection in the Indonesian language: A dataset and preliminary study. In *Proc. of ICACSIS*. 233–238.

[5] Rabab Alkhalifa, Elena Kochkina, and Arkaitz Zubiaga. 2023. Building for tomorrow: Assessing the temporal persistence of text classifiers. *Information Processing & Management* 60, 2 (2023), 103200.

[6] Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information* 13, 6 (2022), 273.

[7] Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465* (2020).

[8] Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2021. A deep dive into multilingual hate speech classification. In *Proc. of ECML PKDD*. Springer, Cham, 423–439.

[9] Maaz Amjad, Alisa Zhila, Grigori Sidorov, Andrey Labunets, Sabur Butt, Hamza Imam Amjad, Oxana Vitman, and Alexander Gelbukh. 2022. UrduThreat@FIRE2021: Shared track on abusive threat identification in Urdu. In *Proc. of FIRE* (Virtual Event, India). ACM, New York, NY, USA, 9–11.

[10] Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proc. of ACM SIGIR* (Paris, France). ACM, New York, NY, USA, 45–54.

[11] Aymé Arango, Jorge Pérez, and Barbara Poblete. 2021. Cross-lingual hate speech detection based on multilingual domain-specific word embeddings. *arXiv preprint arXiv:2104.14728* (2021).

[12] Aymé Arango, Jorge Pérez, and Barbara Poblete. 2022. Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems* 105, C (2022), 101584.

[13] Ayme Arango Monnar, Jorge Perez, Barbara Poblete, Magdalena Saldaña, and Valentina Proust. 2022. Resources for multilingual hate speech detection. In *Proc. of the WOAH Workshop*. ACL, Seattle, Washington (Hybrid), 122–130.

[14] Masaki Arata. 2019. *Study on Change of Detection Accuracy Over Time in Cyberbullying Detection*. Ph. D. Dissertation. Master's thesis, Kitami Institute of Technology, Department of Computer Science.

[15] Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the ACL* 7 (2019), 597–610.

[16] Md Rabiul Awal, Roy Ka-Wei Lee, Eshaan Tanwar, Tanmay Garg, and Tanmoy Chakraborty. 2023. Model-agnostic meta-learning for multilingual hate speech detection. *IEEE Trans. on Computational Social Systems* 11, 1 (2023), 1086–1095.

[17] Femi Emmanuel Ayo, Olusegun Folorunso, Friday Thomas Ibharalu, and Idowu Ademola Osinuga. 2020. Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review* 38 (2020), 100311.

[18] Srijan Bansal, Vishal Garimella, Ayush Suhane, Jasabanta Patro, and Animesh Mukherjee. 2020. Code-switching patterns can be an effective route to improve performance of downstream NLP applications: A case study of humour, sarcasm and hate speech detection. In *Proc. of ACL*. ACL, Online, 1018–1023.

[19] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proc. of the SemEval Workshop*. ACL, Minneapolis, Minnesota, USA, 54–63.

[20] Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr Ojha. 2020. Developing a multilingual annotated corpus of misogyny and aggression. *arXiv preprint arXiv:2003.07428* (2020).

[21] Irina Bigoulaeva, Viktor Hangya, and Alexander Fraser. 2021. Cross-lingual transfer learning for hate speech detection. In *Proc. of the LT-EDI Workshop*. ACL, Kyiv, 15–25.

[22] Irina Bigoulaeva, Viktor Hangya, Iryna Gurevych, and Alexander Fraser. 2022. Addressing the challenges of cross-lingual hate speech detection. *arXiv preprint arXiv:2201.05922* (2022).

[23] Irina Bigoulaeva, Viktor Hangya, Iryna Gurevych, and Alexander Fraser. 2023. Label modification and bootstrapping for zero-shot cross-lingual hate speech detection. *Language Resources and Evaluation* 57, 4 (2023), 1515–1546.

[24] Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proc. of PEOPLES Workshop*. ACL, New Orleans, Louisiana, USA, 36–41.

[25] Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, Tesconi Maurizio, et al. [n. d.]. Overview of the evalita 2018 hate speech detection task. In *Proc. of CEUR Workshop*, Vol. 2263. CEUR, 1–9.

[26] Zakaria Boulouard, Mariya Ouaissa, Mariyam Ouaissa, Moez Krichen, Mutiq Almutiq, and Karim Gasmi. 2022. Detecting hateful and offensive speech in Arabic social media using transfer learning. *Applied Sciences* 12, 24 (2022), 12823.

[27] Uwe Bretschneider and Ralf Peters. 2017. Detecting offensive statements towards foreigners in social media. Hawaii. *International Conference on System Sciences*. (2017).

[28] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Proc. of NeurIPS* 33 (2020), 1877–1901.

[29] Ben Burtenshaw and Mike Kestemont. 2021. A Dutch dataset for cross-lingual multilabel toxicity detection. In *Proc. of the BUCC Workshop*. INCOMA Ltd., Online (Virtual Mode), 75–79.

[30] Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2022. Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. *Language Resources and Evaluation* 56, 3 (2022), 765–806.

[31] Polychronis Charitidis, Stavros Doropoulos, Stavros Vologiannidis, Ioannis Papastergiou, and Sophia Karakeva. 2020. Towards countering hate speech against journalists on social media. *Online Social Networks and Media* 17 (2020), 100071.

[32] Yulong Chen, Yang Liu, Li Dong, Shuohang Wang, Chenguang Zhu, Michael Zeng, and Yue Zhang. 2022. AdaPrompt: Adaptive model training for prompt-based NLP. In *Findings of EMNLP*. ACL, Abu Dhabi, UAE, 6057–6068.

[33] Anusha Chhabra and Dinesh Kumar Vishwakarma. 2023. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems* 29, 3 (2023), 1203–1230.

[34] Patricia Chiril, Farah Benamara Zitoune, Véronique Moriceau, Marlène Coulomb-Gully, and Abhishek Kumar. 2019. Multilingual and multitarget hate speech detection in Tweets. In *Proc. of TALN-RECITAL*. ATALA, Toulouse, France, 351–360.

[35] Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through Nichesourcing: A Multilingual Dataset of Responses to Fight Online Hate Speech. In *Proc. of ACL*. ACL, Florence, Italy, 2819–2829.

[36] Çağrı Çöltekin. 2020. A corpus of Turkish offensive language on social media. In *Proc. of LREC*. 6174–6184.

[37] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proc. of ACL*. ACL, Online, 8440–8451.

[38] Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. Hybrid emoji-based masked language models for zero-shot abusive language detection. In *Findings of EMNLP*. ACL, Online, 943–949.

[39] Mithun Das, Somnath Banerjee, and Animesh Mukherjee. 2022. Data bootstrapping approaches to improve low resource abusive language detection for indic languages. In *Proc. of ACM HT* (Barcelona, Spain). ACM, New York, NY, USA, 32–42.

[40] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proc. of ICWSM* (Montreal, Canada). 512–515.
[41] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proc. of the ALW Workshop*. ACL, Brussels, Belgium, 11–20.
[42] Gretel Liz De la Peña Sarracén and Paolo Rosso. 2022. Unsupervised embeddings with graph auto-encoders for multi-domain and multilingual hate speech detection. In *Proc. of LREC*. ELRA, Marseille, France, 2196–2204.
[43] Rogers de Pelle and Viviane Moreira. 2017. Offensive comments in the Brazilian web: A dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining* (São Paulo). SBC, Porto Alegre, RS, Brasil.
[44] Wubetu Barud Demilie and Ayodeji Olalekan Salau. 2022. Detection of fake news and hate speech for Ethiopian languages: A systematic review of the approaches. *Journal of Big Data* 9, 1 (2022), 66.
[45] Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. COLD: A benchmark for Chinese offensive language detection. In *Proc. of EMNLP*. ACL, Abu Dhabi, UAE, 11580–11599.
[46] Neha Deshpande, Nicholas Farris, and Vidhur Kumar. 2022. Highly generalizable models for multilingual hate speech detection. *arXiv preprint arXiv:2201.11294* (2022).
[47] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT*. ACL, Minneapolis, Minnesota, 4171–4186.
[48] Dmitry Devyatkin, Ivan Smirnov, Margarita Ananyeva, Maria Kobozeva, Andrey Chepovskiy, and Fyodor Solovyev. 2017. Exploring linguistic features for extremist texts detection (on the material of Russian-speaking illegal texts). In *Proc. of IEEE ISI*. 188–190.
[49] Suman Dowlagar and Radhika Mamidi. 2021. A survey of recent neural network models on code-mixed indian hate speech data. In *Proc. of FIRE*. 67–74.
[50] Matthew S. Dryer and Martin Haspelmath (Eds.). 2013. *WALS Online (v2020.3)*. Zenodo. DOI: https://doi.org/10.5281/zenodo.7385533
[51] Michal Ptaszynski1 Pawel Dybala, Tatsuaki Matsuba2 Fumito Masui, Rafal Rzepka, and Kenji Araki. 2010. Machine learning and affect analysis against cyber-bullying. In *Proc. of the Linguistic and Cognitive Approaches to Dialog Agents Symposium*.
[52] Aya Elouali, Zakaria Elberrichi, and Nadia Elouali. 2020. Hate speech detection on multilingual twitter using convolutional neural networks. *Revue d'Intelligence Artificielle* 34, 1 (2020), 81–88.
[53] Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018. Peer to peer hate: Hate speech instigators and their targets. In *Proc. of ICWSM*, Vol. 12.
[54] Juuso Eronen, Michal Ptaszynski, Fumito Masui, Masaki Arata, Gniewosz Leliwa, and Michal Wroczynski. 2022. Transfer language selection for zero-shot cross-lingual abusive language detection. *Information Processing & Management* 59, 4 (2022), 102981.
[55] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proc. of ACL*. ACL, Dublin, Ireland, 878–891.
[56] Daniel C. Ferreira, André F. T. Martins, and Mariana S. C. Almeida. 2016. Jointly learning to embed and predict with multiple languages. In *Proc. of ACL)*. ACL, Berlin, Germany, 2019–2028.
[57] Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the evalita 2018 task on automatic misogyny identification (AMI). In *EVALITA@CLiC-it*.
[58] Elisabetta Fersini, Paolo Rosso, and Mary E. Anzovino. 2018. Overview of the task on automatic misogyny identification at IberEval 2018. In *IberEval@SEPLN*.
[59] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proc. of ICML*. PMLR, 1126–1135.
[60] Anderson Almeida Firmino, Cláudio Souza de Baptista, and Anselmo Cardoso de Paiva. 2021. Using cross lingual learning for detecting hate speech in Portuguese. In *Database and Expert Systems Applications*. Springer, Cham, 170–175.
[61] Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences* 10, 12 (2020), 4180.
[62] Rights for Peace. [n. d.]. What is Hate Speech? Retrieved from https://www.rightsforpeace.org/hate-speech. Accessed: 2023-10-24.
[63] Paula Fortuna. 2017. Automatic detection of hate speech in text: An overview of the topic and dataset annotation with hierarchical classes. https://hdl.handle.net/10216/106028
[64] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.* 51, 4, Article 85 (July 2019), 1–30.
[65] Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A hierarchically-labeled Portuguese hate speech dataset. In *Proc. of the ALW Workshop*. ACL, Florence, Italy, 94–104.

[66] Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proc. of ICWSM*, Vol. 12.

[67] Saurabh Sampatrao Gaikwad, Tharindu Ranasinghe, Marcos Zampieri, and Christopher Homan. 2021. Cross-lingual Offensive Language Identification for Low Resource Languages: The Case of Marathi. In *Proc. of RANLP*. INCOMA Ltd., Held Online, 437–443.

[68] Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. *arXiv preprint arXiv:1710.07395* (2017).

[69] Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling bias in toxic speech detection: A survey. 55, 13s, Article 264 (jul 2023), 32 pages.

[70] Goran Glavaš, Mladen Karan, and Ivan Vulić. 2020. XHate-999: Analyzing and detecting abusive language across domains and languages. In *Proc. of COLING*. ICCL, Barcelona, Spain (Online), 6350–6365.

[71] Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, et al. 2017. A large labeled corpus for online harassment research. In *Proc. of ACM WebSci* (Troy, New York, USA). ACM, New York, NY, USA, 229–233.

[72] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proc. of LREC*.

[73] Vikram Gupta, Sumegh Roychowdhury, Mithun Das, Somnath Banerjee, Punyajoy Saha, Binny Mathew, Animesh Mukherjee, et al. 2022. Multilingual abusive comment detection at scale for indic languages. *Proc. of NeurIPS* 35 (2022), 26176–26191.

[74] Vanessa Hahn, Dana Ruiter, Thomas Kleinbauer, and Dietrich Klakow. 2021. Modeling profanity and hate speech in social media with semantic subspaces. In *Proc. of the WOAH Workshop*. ACL, Online, 6–16.

[75] Lawrence Han and Hao Tang. 2022. Designing of prompts for hate speech recognition with in-context learning. In *Proc. of CSCI*. IEEE, 319–320.

[76] Adeep Hande, Siddhanth U. Hegde, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. Benchmarking multi-task learning for sentiment analysis and offensive language identification in under-resourced dravidian languages. *arXiv preprint arXiv:2108.03867* (2021).

[77] Adeep Hande, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2020. KanCMD: Kannada CodeMixed dataset for sentiment analysis and offensive language detection. In *Proc. of the PEOPLES Workshop*. ACL, Barcelona, Spain (Online), 54–63.

[78] Adeep Hande, Karthik Puranik, Konthala Yasaswini, Ruba Priyadharshini, Sajeetha Thavareesan, Anbukkarasi Sampath, Kogilavani Shanmugavadivel, Durairaj Thenmozhi, and Bharathi Raja Chakravarthi. 2021. Offensive language identification in low-resourced code-mixed dravidian languages using pseudo-labeling. *arXiv preprint arXiv:2108.12177* (2021).

[79] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proc. of ACL*. ACL, Dublin, Ireland, 3309–3326.

[80] Fatemah Husain and Ozlem Uzuner. 2021. A survey of offensive language detection for the Arabic language. *ACM TALLIP* 20, 1 (2021), 1–44.

[81] Fatemah Husain and Ozlem Uzuner. 2022. Transfer learning across Arabic dialects for offensive language detection. In *Proc. of IALP*. 196–205.

[82] Muhammad Okky Ibrohim and Indra Budi. 2019. Translated vs non-translated method for multilingual hate speech identification in Twitter. *International Journal on Advanced Science, Engineering and Information Technology* 9, 4 (1 Jan. 2019), 1116–1123.

[83] Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proc. of the ALW Workshop*. ACL, Florence, Italy, 46–57.

[84] Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of Hate Speech automatic detection using Natural Language Processing. *Neurocomputing* 546 (2023), 126232.

[85] Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. KOLD: Korean offensive language dataset. In *Proc. of EMNLP*. ACL, Abu Dhabi, UAE, 10818–10833.

[86] Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. SWSR: A Chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media* 27 (2022), 100182.

[87] Aiqi Jiang and Arkaitz Zubiaga. 2021. Cross-lingual capsule network for hate speech detection in social media. In *Proc. of ACM HT* (Virtual Event, USA). ACM, New York, NY, USA, 217–223.

[88] Aiqi Jiang and Arkaitz Zubiaga. 2023. SexWEs: Domain-aware word embeddings via cross-lingual semantic specialisation for Chinese sexism detection in social media. *Proc. of ICWSM* 17, 1 (Jun. 2023), 447–458.

[89] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proc. of AAAI*, Vol. 34. 8018–8025.

[90] Raghav Kapoor, Yaman Kumar, Kshitij Rajput, Rajiv Ratn Shah, Ponnurangam Kumaraguru, and Roger Zimmermann. 2019. Mind your language: Abuse and offense detection for code-switched languages. *Proc. of AAAI* 33, 01 (Jul. 2019), 9951–9952.

[91] Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. *arXiv preprint arXiv:2005.02439* (2020).

[92] Muhammad Moin Khan, Khurram Shahzad, and Muhammad Kamran Malik. 2021. Hate speech detection in Roman Urdu. *ACM Transactions on Asian and Low-Resource Language Information Processing* 20, 1, Article 9 (Mar. 2021).

[93] Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C. Fraser. 2021. Confronting abusive language online: A survey from the ethical and human rights perspective. *Artificial Intelligence Research* 71 (2021), 431–478.

[94] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, Vol. 2. Lille.

[95] Akshi Kumar and Nitin Sachdeva. 2022. Multi-input integrative learning using deep neural networks and transfer learning for cyberbullying detection in real-time code-mix data. *Multimedia systems* 28, 6 (2022), 2027–2041.

[96] Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2020. Evaluating aggression identification in social media. In *Proc. of the TRAC Workshop*. 1–5.

[97] Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-annotated Corpus of Hindi-English code-mixed data. In *Proc. of LREC*. ELRA, Miyazaki, Japan.

[98] Mirko Lai, Fabio Celli, Alan Ramponi, Sara Tonelli, Cristina Bosco, and Viviana Patti. 2023. Haspeede3 at evalita 2023: Overview of the political and religious hate speech detection task. In *EVALITA*.

[99] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *Proc. of ICLR*.

[100] Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proc. of ICLR*.

[101] Chanhee Lee, Kisu Yang, Taesun Whang, Chanjun Park, Andrew Matteson, and Heuiseok Lim. 2021. Exploring the data efficiency of cross-lingual post-training in pretrained language models. *Applied Sciences* 11, 5 (2021), 1974.

[102] Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on challenges in representation learning*, Vol. 3. Atlanta, 896.

[103] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the ACL* 8 (2020), 726–742.

[104] Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2019. The FRENK datasets of socially unacceptable discourse in Slovene and English. In *Proc. of TSD*. Springer, 103–114.

[105] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proc. of NeurIPS 30*. Curran Associates, Inc., 4765–4774.

[106] Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2021. Overview of the HASOC track at FIRE 2020: Hate speech and offensive language identification in Tamil, Malayalam, Hindi, English and German. In *Proc. of FIRE* (Hyderabad, India). ACM, New York, NY, USA, 29–32.

[107] Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in Indo-European Languages. In *Proc. of FIRE* (Kolkata, India). ACM, New York, NY, USA, 14–17.

[108] Ilia Markov, Nikola Ljubešić, Darja Fišer, and Walter Daelemans. 2021. Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In *Proc. of WASSA*. ACL, Online, 149–159.

[109] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proc. of AAAI*, Vol. 35. 14867–14875.

[110] Puneet Mathur, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. 2018. Did you offend me? Classification of offensive tweets in Hinglish language. In *Proc. of the ALW Workshop*. ACL, Brussels, Belgium, 138–148.

[111] Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting offensive tweets in Hindi-English code-switched language. In *Proc. of the SocialNLP Workshop*. ACL, Melbourne, Australia, 18–26.

[112] Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Tackling online abuse: A survey of automated abuse detection methods. *arXiv preprint arXiv:1908.06024* (2019).

[113] Sandip Modha, Thomas Mandl, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Tharindu Ranasinghe, and Marcos Zampieri. 2021. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages and conversational hate speech. In *Proce. of FIRE*. 1–3.

[114] Syrielle Montariol, Arij Riabi, and Djamé Seddah. 2022. Multilingual auxiliary tasks training: Bridging the gap between languages for zero-shot transfer of hate speech detection models. In *Findings of AACL-IJCNLP*. ACL, Online only, 347–363.

[115] Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2022. Cross-lingual few-shot hate speech and offensive language detection using meta learning. *IEEE Access* 10 (2022), 14880–14896.
[116] Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on Arabic social media. In *Proc. of the ALW Workshop*. ACL, Vancouver, BC, Canada, 52–56.
[117] Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa. 2020. Overview of OSACT4 Arabic offensive language detection shared task. In *Proc. of the OSACT Workshop*. ELRA, Marseille, France, 48–52.
[118] Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2021. Arabic offensive language on Twitter: Analysis and experiments. In *Proc. of WANLP*. ACL, Kyiv, Ukraine (Virtual), 126–135.
[119] Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-hsab: A levantine twitter dataset for hate speech and abusive language. In *Proc. of the ALW Workshop*. 111–118.
[120] Arianna Muti and Alberto Barrón-Cedeño. 2022. A checkpoint on multilingual misogyny identification. In *Proc. of ACL SRW*. ACL, Dublin, Ireland, 454–460.
[121] Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proc. of ACL-IJCNLP*. ACL, Online, 907–914.
[122] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774. (2023).
[123] Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proc. of EMNLP-IJCNLP*. ACL, Hong Kong, China, 4675–4684.
[124] Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In *Proc. of EMNLP*. ACL, Online and Punta Cana, Dominican Republic, 27–38.
[125] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, et al. 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *PLoS Med* 18, 3 (2021), e1003593.
[126] Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny detection in twitter: A multilingual and cross-domain study. *Information Processing & Management* 57, 6 (2020), 102360.
[127] Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2021. A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. *Information Processing & Management* 58, 4 (2021), 102544.
[128] Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2021. Towards multidomain and multilingual abusive language detection: A survey. *Personal and Ubiquitous Computing* 27, 1 (2021), 17–43.
[129] Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proc. of ACL SRW*. ACL, Florence, Italy, 363–370.
[130] Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Trans. on Knowledge and Data Engineering* 22, 10 (2010), 1345–1359.
[131] Kartikey Pant and Tanvi Dadu. 2020. Towards code-switched classification exploiting constituent language resources. In *Proc. of AACL-IJCNLP SRW*. ACL, Suzhou, China, 37–43.
[132] Sayanta Paul, Sriparna Saha, and Jyoti Prakash Singh. 2023. COVID-19 and cyberbullying: Deep ensemble model to identify cyberbullying from code-switched languages during the pandemic. *Multimedia Tools and Applications* 82, 6 (2023), 8773–8789.
[133] John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 Task 5: Toxic Spans Detection. In *Proc. of the SemEval Workshop*. ACL, Online, 59–69.
[134] Andraž Pelicon, Ravi Shekhar, Matej Martinc, Blaž Škrlj, Matthew Purver, and Senja Pollak. 2021. Zero-shot cross-lingual content filtering: Offensive language and hate speech detection. In *Proc.of the EACL Hackashop*. ACL, Online, 30–34.
[135] Andraž Pelicon, Ravi Shekhar, Blaž Škrlj, Matthew Purver, and Senja Pollak. 2021. Investigating cross-lingual training for offensive language detection. *PeerJ Computer Science* 7 (2021), e559.
[136] Juan Carlos Pereira-Kohatsu, Lara Quijano-Sánchez, Federico Liberatore, and Miguel Camacho-Collados. 2019. Detecting and monitoring hate speech in twitter. *Sensors* 19, 21 (2019), 4654.
[137] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL-HLT*. ACL, New Orleans, Louisiana, 2227–2237.
[138] Tung Minh Phung and Jan Cloos. 2022. An exploratory experiment on Hindi, Bengali hate-speech detection and transfer learning using neural networks. *arXiv preprint arXiv:2201.01997* (2022).
[139] Matúš Pikuliak, Marián Šimko, and Mária Bieliková. 2021. Cross-lingual learning for text processing: A survey. *Expert Systems with Applications* 165 (2021), 113765.
[140] Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. In *Proc. of LREC*. ELRA, Marseille, France, 5113–5119.
[141] Laura Plaza, Jorge Carrillo-de Albornoz, Roser Morante, Enrique Amigó, Julio Gonzalo, Damiano Spina, and Paolo Rosso. 2023. Overview of EXIST 2023: sEXism identification in social NeTworks. In *Proc. of ECIR*. Springer, 593–599.

[142] Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *Proc. of the WOAH workshop*. ACL, Toronto, Canada, 60–68.
[143] Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation* 55 (2021), 477–523.
[144] Michał Ptaszyński, Gniewosz Leliwa, Mateusz Piech, and Aleksander Smywiński-Pohl. 2018. Cyberbullying detection–technical report 2/2018, department of computer science AGH, university of science and technology. *arXiv preprint arXiv:1808.00926* (2018).
[145] Michal Ptaszynski, Agata Pieciukiewicz, and Paweł Dybała. 2019. Results of the poleval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in polish Twitter. In *Proc. of the PolEval 2019 Workshop*. 89–110.
[146] Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. In *Proc. of EMNLP*. ACL, Online, 5838–5844.
[147] Tharindu Ranasinghe and Marcos Zampieri. 2021. An evaluation of multilingual offensive language identification methods for the languages of India. *Information* 12, 8 (2021), 306.
[148] Tharindu Ranasinghe and Marcos Zampieri. 2021. MUDES: Multilingual detection of offensive spans. In *Proc. of NAACL-HLT: Demonstrations*. ACL, Online, 144–152.
[149] Tharindu Ranasinghe and Marcos Zampieri. 2021. Multilingual offensive language identification for low-resource languages. *ACM Transactions on Asian and Low-Resource Language Information Processing* 21, 1 (2021), 13.
[150] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?": Explaining the predictions of any classifier. In *Proc. of NAACL: Demonstrations*. ACL, San Diego, California, 97–101.
[151] Bilingual Word Embeddings Rivaling. 2021. Llod-driven bilingual word embeddings rivaling cross-lingual transformers in quality of life concept detection from french online health communities. In *Proc. of SEMANTiCS*, Vol. 53. 89.
[152] Hammad Rizwan, Muhammad Haroon Shakeel, and Asim Karim. 2020. Hate-speech and offensive language detection in roman Urdu. In *Proc. of EMNLP*. 2512–2522.
[153] Sebastián E. Rodríguez, Héctor Allende-Cid, and Héctor Allende. 2021. Detecting hate speech in cross-lingual and multi-lingual settings using language agnostic representations. In *Proc. of CIARP*. Springer, Cham, 77–87.
[154] Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. 2021. Overview of exist 2021: Sexism identification in social networks. *Procesamiento del Lenguaje Natural* 67 (2021), 195–207.
[155] Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Adrián Mendieta-Aragón, Guillermo Marco-Remón, Maryna Makeienko, María Plaza, Julio Gonzalo, Damiano Spina, and Paolo Rosso. 2022. Overview of exist 2022: Sexism identification in social networks. *Procesamiento del Lenguaje Natural* 69 (2022), 229–240.
[156] Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam. 2021. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In *Proc. of IJCACI*. Springer, 457–468.
[157] Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118* (2017).
[158] Paul Röttger, Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Data-efficient strategies for expanding hate speech detection into under-resourced languages. In *Proc. of EMNLP*. ACL, Abu Dhabi, UAE, 5674–5691.
[159] Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual HateCheck: Functional tests for multilingual hate speech detection models. In *Proc. of the WOAH Workshop*. ACL, Seattle, Washington (Hybrid), 154–169.
[160] Sarthak Roy, Ashish Harshavardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing LLMs for hate speech detection: strengths and vulnerabilities. *arXiv preprint arXiv:2310.12860* (2023).
[161] Anastasia Ryzhova, Dmitry Devyatkin, Sergey Volkov, and Vladimir Budzko. 2022. Training multilingual and adversarial attack-robust models for hate detection on social media. *Procedia Comp. Sci.* 213 (2022), 196–202.
[162] Cesa Salaam, Franck Dernoncourt, Trung Bui, Danda Rawat, and Seunghyun Yoon. 2022. Offensive content detection via synthetic code-switched text. In *Proc. of COLING*. ICCL, Gyeongju, Republic of Korea, 6617–6624.
[163] Manuela Sanguinetti, Gloria Comandini, Elisa Di Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, and Irene Russo. 2020. Haspeede 2@ evalita2020: Overview of the evalita 2020 hate speech detection task. *EVALITA Evaluation of NLP and Speech Tools for Italian*. 93–101.
[164] Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter corpus of hate speech against immigrants. In *Proc. of LREC*. ELRA, Miyazaki, Japan.
[165] Shrey Satapara, Prasenjit Majumder, Thomas Mandl, Sandip Modha, Hiren Madhu, Tharindu Ranasinghe, Marcos Zampieri, Kai North, and Damith Premasiri. 2023. Overview of the HASOC subtrack at FIRE 2022: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages. In *Proc. of FIRE* (Kolkata, India). ACM, New York, NY, USA, 4–7.

[166] Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proc. of the SocialNLP Workshop*. 1–10.
[167] Ravi Shekhar, Marko Pranjic, Senja Pollak, Andraz Pelicon, and Matthew Purver. 2020. Automating news comment moderation with limited resources: Benchmarking in croatian and estonian. *Journal for Language Technology and Computational Linguistics* 34, 1 (2020), 49–79.
[168] Xiayang Shi, Xinyi Liu, Chun Xu, Yuanyuan Huang, Fang Chen, and Shaolin Zhu. 2022. Cross-lingual offensive speech identification with transfer learning for low-resource languages. *Computers and Electrical Engineering* 101 (2022), 108005.
[169] Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive language and hate speech detection for Danish. In *Proc. of LREC*. ELRA, Marseille, France, 3498–3508.
[170] Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proc. of ICLR*.
[171] Hajung Sohn and Hyunju Lee. 2019. MC-BERT4HATE: Hate speech detection using multi-channel BERT for different languages and translations. In *2019 International Conference on Data Mining Workshops (ICDMW)*. 551–559.
[172] Guizhe Song, Degen Huang, and Zhifeng Xiao. 2021. A study of multilingual toxic text detection approaches under imbalanced sample distribution. *Information* 12, 5 (2021), 205.
[173] Vivian Stamou, Iakovi Alexiou, Antigone Klimi, Eleftheria Molou, Alexandra Saivanidou, and Stella Markantonatou. 2022. Cleansing & expanding the HURTLEX(el) with a multidimensional categorization of offensive words. In *Proc. of WOAH*. ACL, 102–108.
[174] Lukas Stappen, Fabian Brunn, and Bj"orn Schuller. 2020. Cross-lingual zero-and few-shot hate speech detection utilising frozen transformer language models and AXEL. *arXiv preprint arXiv:2004.13850* (2020).
[175] Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of GermEval Task 2, 2019 shared task on the identification of offensive language. In *Proc. of KONVENS*. München, 352 – 363.
[176] Sahil Swami, Ankush Khandelwal, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A corpus of english-hindi code-mixed tweets for sarcasm detection. *arXiv preprint arXiv:1805.11869* (2018).
[177] Teodor Tita and Arkaitz Zubiaga. 2021. Cross-lingual hate speech detection using transformer models. *arXiv preprint arXiv:2111.00981* (2021).
[178] UNSPAHS. 2019. UN Strategy and Plan of Action on Hate Speech. Retrieved from https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml. Accessed: 2023-10-24.
[179] Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2015. Automatic detection and prevention of cyberbullying. In *Proc. of HUSO*. IARIA, 13–18.
[180] Neeraj Vashistha and Arkaitz Zubiaga. 2021. Online multilingual hate speech detection: Experimenting with Hindi and English social media. *Information* 12, 1 (2021), 5.
[181] Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos One* 15, 12 (2020), e0243300.
[182] Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2020. Learning from the worst: Dynamically generated datasets to improve online hate detection. *arXiv preprint arXiv:2012.15761* (2020).
[183] Fedor Vitiugin, Yasas Senarath, and Hemant Purohit. 2021. Efficient detection of multilingual hate speech by using interactive attention network with minimal human feedback. In *Proc. of ACM WebSci* (Virtual Event, United Kingdom). ACM, New York, NY, USA, 130–138.
[184] Dong Wang and Thomas Fang Zheng. 2015. Transfer learning for speech and language processing. In *Proc. of APSIPA*. IEEE, 1225–1237.
[185] Han Wang, Ming Shan Hee, Md Rabiul Awal, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2023. Evaluating GPT-3 generated explanations for hateful content moderation. In *Proc. of IJCAI*. 6255–6263.
[186] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proc. of NAACL SRW*. ACL, San Diego, California, 88–93.
[187] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652* (2021).
[188] Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. In *Proceedings of the GermEval 2018, 14th Conference on Natural Language*. Vienna, Austria.
[189] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proc. of WWW*. 1391–1399.
[190] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proc. of NAACL-HLT*. ACL, Online, 483–498.

[191] Peiling Yi and Arkaitz Zubiaga. 2023. Session-based cyberbullying detection in social media: A survey. *Online Social Networks and Media* 36 (2023), 100250.

[192] Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: A review on obstacles and solutions. *PeerJ Computer Science* 7 (2021), e598.

[193] Fatima zahra El-Alami, Said Ouatik El Alaoui, and Noureddine En Nahnahi. 2022. A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model. *Computer and Information Sciences* 34, 8, Part B (2022), 6048–6056.

[194] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 Task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proc. of the SemEval Workshop*. ACL, Minneapolis, Minnesota, USA, 75–86.

[195] Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proc. of the SemEval Workshop*. ICCL, Barcelona (online), 1425–1447.

[196] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

[197] Li Zhou, Laura Cabello, Yong Cao, and Daniel Hershcovich. 2023. Cross-cultural transfer learning for Chinese offensive language detection. In *Proc. of the C3NLP Workshop)*. ACL, Dubrovnik, Croatia, 8–15.

[198] Haris Bin Zia, Ignacio Castro, Arkaitz Zubiaga, and Gareth Tyson. 2022. Improving zero-shot cross-lingual hate speech detection with pseudo-label fine-tuning of transformer language models. In *Proc. of International AAAI Conference on Web and Social Media* 16, 1 (2022), 1435–1439.

Received 17 January 2024; revised 12 February 2026; accepted 1 March 2026