A survey on multi-lingual offensive language detection

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

A Survey on Multi-lingual Offensive Language Detection

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

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

Corresponding author:
Khouloud Mnassri

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

ABSTRACT

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

INTRODUCTION

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

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

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

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

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

Motivation and Impact

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

Definition of Hate Speech and Offensive Language

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

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

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

Multilingual vs. Cross-Lingual

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

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

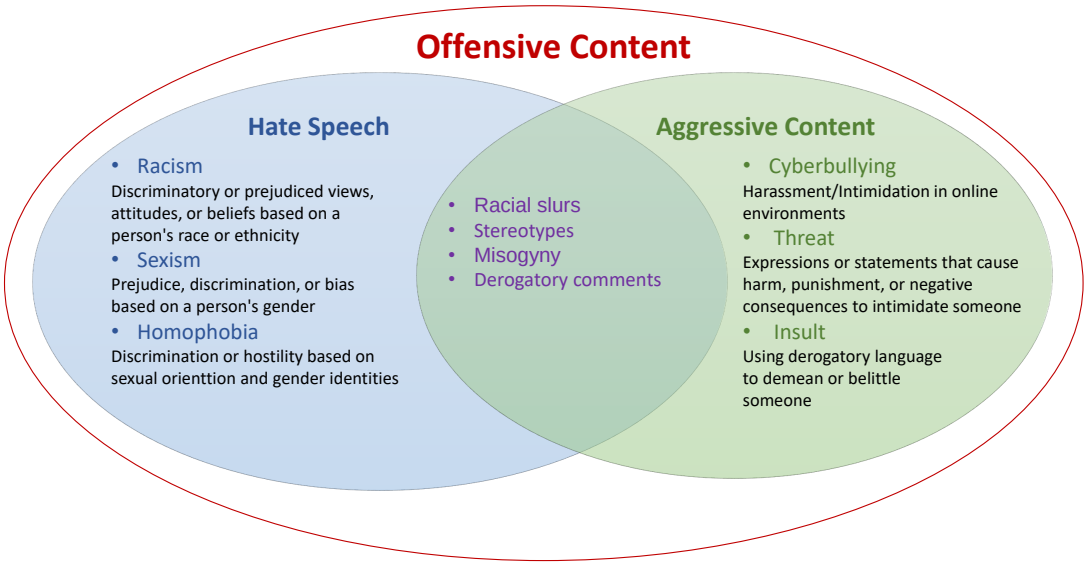

Figure 1. Categories of Offensive Content: Offensive content, hate speech, and aggressive content are distinct but they are often overlapping categories that involve harmful or negative expressions. While there is some overlap between these categories, their distinctions lie in their specific intentions,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

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

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

Structure of this survey

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

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

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

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

BACKGROUND ON MULTILINGUAL HATE SPEECH PHENOMENA

Previous Surveys

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

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

Surveys on Hate Speech - from general perspective:

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

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

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

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

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

Surveys on Multilingual Hate Speech - multilingual perspective:

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

Existing Gaps in the Previous Surveys

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

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

APPROACHES ON MULTILINGUAL HATE SPEECH DETECTION

Existing Approaches

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

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

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

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

Logistic Regression (LR)

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

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

Abbreviation: LR:Logistic Regression, TL:Transfer Learning, MT:Machine Translation, EL:Ensemble Learning, Meta-L:Meta Learning, Multitask-L:Multitask Learning, UL:Unsupervised Learning, GAE:Graph Auto-Encoders. Languages abbreviations are based on ISO 639 language codes list.
♢/♣: ♢ refers to **Multilingual** methods & ♣ refers to **cross-lingual** methods. Feature extraction methods are put between ().

Table 2. Overview of approaches on Multilingual and Cross-Lingual hate speech detection.
(Note: only one representative study of each approach is cited in the table due to space limitation.)

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

Deep Neural Networks (DNN)

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

Transfer Learning (TL)

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

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

Machine Translation (MT)

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

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

Ensemble Learning (EL)

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

Meta Learning (Meta-L)

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

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

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

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

Multitask Learning (Multitask-L)

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

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

Unsupervised learning (UL)

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

Other Technologies

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

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

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

Analysis of the existing approaches

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

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

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

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

DATASETS ON MULTILINGUAL HATE SPEECH DETECTION

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

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

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

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

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

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

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

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

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

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

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

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

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

RESOURCES FOR MULTILINGUAL HATE SPEECH DETECTION

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

Collaborative International Projects

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

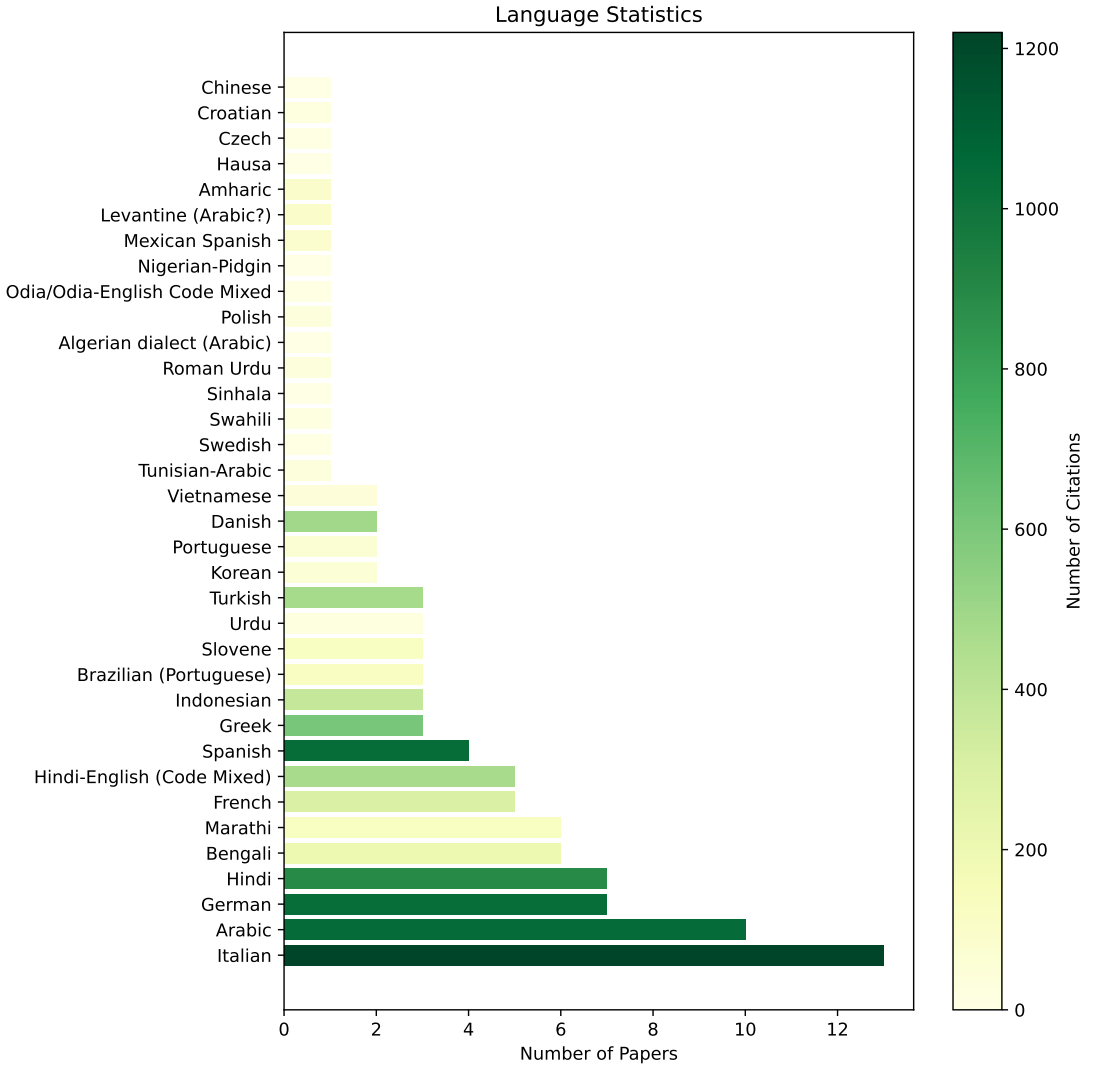

Figure 2. Datasets distribution based on languages. Number of papers studies in this survey and their impact (based on citations @May 2023).
Note: English (47 papers with 11k citations) is excluded from the figure to have a better visualization in the distribution.

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

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

Table 5. Collaborative International Projects on Hate speech.

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

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

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

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

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

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

Available Products

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

CHALLENGES AND LIMITATIONS

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

Challenges

Technical Challenges

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

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

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

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

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

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

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

Non-Technical Challenges

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

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

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

Limitations

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

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

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

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

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

Future Directions

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

on DataSet

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

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

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

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

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

on Modeling

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

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

on Low Resource Languages

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

CONCLUSION

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

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

REFERENCES

Abercrombie, G., Hovy, D., and Prabhakaran, V. (2023). Temporal and second language influence on intra-annotator agreement and stability in hate speech labelling. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 96–103, Toronto, Canada. Association for Computational Linguistics.

Aharoni, R., Johnson, M., and Firat, O. (2019). Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Ahn, H., Sun, J., Park, C. Y., and Seo, J. (2020a). NLPDove at SemEval-2020 task 12: Improving offensive language detection with cross-lingual transfer. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1576–1586, Barcelona (online). International Committee for Computational Linguistics.

Ahn, H., Sun, J., Park, C. Y., and Seo, J. (2020b). Nlpdove at semeval-2020 task 12: Improving offensive language detection with cross-lingual transfer. *CoRR*, abs/2008.01354.

Ahuja, K., Diddee, H., Hada, R., Ochieng, M., Ramesh, K., Jain, P., Nambi, A., Ganu, T., Segal, S., Axmed, M., Bali, K., and Sitaram, S. (2023). Mega: Multilingual evaluation of generative ai.

Akhtar, S., Basile, V., and Patti, V. (2021). Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *arXiv preprint arXiv:2106.15896*.

Akram, M. H., Shahzad, K., and Bashir, M. (2023). Ise-hate: A benchmark corpus for inter-faith, sectarian, and ethnic hatred detection on social media in urdu. *Information Processing & Management*, 60(3):103270.

Al-Hassan, A. and Al-Dossari, H. (2019). Detection of hate speech in social networks: a survey on multilingual corpus. In *6th international conference on computer science and information technology*, volume 10, pages 10–5121.

Aldjanabi, W., Dahou, A., Al-qaness, M. A. A., Elaziz, M. A., Helmi, A. M., and Damaševičius, R. (2021). Arabic offensive and hate speech detection using a cross-corpora multi-task learning model. *Informatics*, 8(4).

Alfina, I., Mulia, R., Fanany, M. I., and Ekanata, Y. (2017). Hate speech detection in the indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 233–238.

Ali, R., Farooq, U., Arshad, U., Shahzad, W., and Beg, M. O. (2022a). Hate speech detection on twitter using transfer learning. *Computer Speech & Language*, 74:101365.

Ali, R., Farooq, U., Arshad, U., Shahzad, W., and Beg, M. O. (2022b). Hate speech detection on twitter using transfer learning. *Computer Speech Language*, 74:101365.

Aliyu, S. M., Wajiga, G. M., Murtala, M., Muhammad, S. H., Abdulmumin, I., and Ahmad, I. S. (2022). Herdphobia: A dataset for hate speech against fulani in nigeria. *arXiv preprint arXiv:2211.15262*.

Aluru, S. S., Mathew, B., Saha, P., and Mukherjee, A. (2020). Deep learning models for multilingual hate speech detection. *CoRR*, abs/2004.06465.

Aluru, S. S., Mathew, B., Saha, P., and Mukherjee, A. (2021). A deep dive into multilingual hate speech classification. In Dong, Y., Ifrim, G., Mladenić, D., Saunders, C., and Van Hoecke, S., editors, *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track*, pages 423–439, Cham. Springer International Publishing.

Álvarez-Carmona, M. Á., Guzmán-Falcón, E., Montes-y Gómez, M., Escalante, H. J., Villasenor-Pineda, L., Reyes-Meza, V., and Rico-Sulayes, A. (2018). Overview of mex-a3t at ibereval 2018: Authorship

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Chiril, P., Benamara Zitoune, F., Moriceau, V., Coulomb-Gully, M., and Kumar, A. (2019). Multilingual and multitarget hate speech detection in tweets. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume II : Articles courts*, pages 351–360, Toulouse, France. ATALA.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Chung, Y.-L., Kuzmenko, E., Tekiroglu, S. S., and Guerini, M. (2019). Conan–counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. *arXiv preprint arXiv:1910.03270*.

Cohen, S., Presil, D., Katz, O., Arbili, O., Messica, S., and Rokach, L. (2023). Enhancing social network hate detection using back translation and gpt-3 augmentations during training and test-time. *Information Fusion*, 99:101887.

Çöltekin, Ç. (2020). A corpus of turkish offensive language on social media. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6174–6184.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Dabre, R., Chu, C., and Kunchukuttan, A. (2020). A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

Dadu, T. and Pant, K. (2020). Team rouges at SemEval-2020 task 12: Cross-lingual inductive transfer to detect offensive language. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2183–2189, Barcelona (online). International Committee for Computational Linguistics.

Das, A. K., Al Asif, A., Paul, A., and Hossain, M. N. (2021). Bangla hate speech detection on social media using attention-based recurrent neural network. *Journal of Intelligent Systems*, 30(1):578–591.

Das, M., Pandey, S. K., and Mukherjee, A. (2023). Evaluating chatgpt's performance for multilingual and emoji-based hate speech detection.

Davidson, T., Bhattacharya, D., and Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.

De la Peña Sarracén, G. L. and Rosso, P. (2022). Unsupervised embeddings with graph auto-encoders for multi-domain and multilingual hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2196–2204, Marseille, France. European Language Resources Association.

de Pelle, R. P. and Moreira, V. P. (2017). Offensive comments in the brazilian web: a dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*. SBC.

Del Vigna12, F., Cimino23, A., Dell'Orletta, F., Petrocchi, M., and Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the first Italian conference on cybersecurity (ITASEC17)*, pages 86–95.

Deshpande, N., Farris, N., and Kumar, V. (2022a). Highly generalizable models for multilingual hate speech detection. *ArXiv*, abs/2201.11294.

Deshpande, N., Farris, N., and Kumar, V. (2022b). Highly generalizable models for multilingual hate speech detection. *CoRR*, abs/2201.11294.

d'Sa, A. G., Illina, I., and Fohr, D. (2020). Bert and fasttext embeddings for automatic detection of toxic speech. In *2020 International Multi-Conference on:"Organization of Knowledge and Advanced Technologies"(OCTA)*, pages 1–5. IEEE.

Elouali, A., Elberrichi, Z., and Elouali, N. (2020). Hate speech detection on multilingual twitter using convolutional neural networks. *Rev. d'Intelligence Artif.*, 34(1):81–88.

Eronen, J., Ptaszynski, M., Masui, F., Arata, M., Leliwa, G., and Wroczynski, M. (2022). Transfer language selection for zero-shot cross-lingual abusive language detection. *Information Processing Management*, 59(4):102981.

Evkoski, B., Pelicon, A., Mozetič, I., Ljubešić, N., and Kralj Novak, P. (2022). Retweet communities reveal the main sources of hate speech. *Plos one*, 17(3):e0265602.

Fang, M., Li, Y., and Cohn, T. (2017). Learning how to active learn: A deep reinforcement learning ap-

proach. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Copenhagen, Denmark. Association for Computational Linguistics.

Fanton, M., Bonaldi, H., Tekiroglu, S. S., and Guerini, M. (2021). Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. *arXiv preprint arXiv:2107.08720*.

Fernquist, J., Lindholm, O., Kaati, L., and Akrami, N. (2019). A study on the feasibility to detect hate speech in swedish. In *2019 IEEE international conference on big data (Big Data)*, pages 4724–4729. IEEE.

Fersini, E., Rosso, P., and Anzovino, M. (2018). Overview of the task on automatic misogyny identification at ibereval 2018. *Ibereval@ sepln*, 2150:214–228.

Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR.

Fišer, D., Erjavec, T., and Ljubešić, N. (2017). Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in slovene. In *Proceedings of the first workshop on abusive language online*, pages 46–51.

Fortuna, P., da Silva, J. R., Wanner, L., and Nunes, S. (2019). A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the third workshop on abusive language online*, pages 94–104.

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Fortuna, P., Soler, J., and Wanner, L. (2020). Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th language resources and evaluation conference*, pages 6786–6794.

Gaikwad, S. S., Ranasinghe, T., Zampieri, M., and Homan, C. (2021). Cross-lingual offensive language identification for low resource languages: The case of Marathi. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 437–443, Held Online. INCOMA Ltd.

Ghadery, E. and Moens, M.-F. (2020). LIIR at SemEval-2020 task 12: A cross-lingual augmentation approach for multilingual offensive language identification. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2073–2079, Barcelona (online). International Committee for Computational Linguistics.

Ghosal, S. and Jain, A. (2023). Hatecircle and unsupervised hate speech detection incorporating emotion and contextual semantics. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(4).

Glavaš, G., Karan, M., and Vulić, I. (2020). XHate-999: Analyzing and detecting abusive language across domains and languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6350–6365, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Gokhale, O., Kane, A., Patankar, S., Chavan, T., and Joshi, R. (2022). Spread love not hate: Undermining the importance of hateful pre-training for hate speech detection.

Goldzycher, J., Preisig, M., Amrhein, C., and Schneider, G. (2023). Evaluating the effectiveness of natural language inference for hate speech detection in languages with limited labeled data.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Guerreiro, N. M., Alves, D., Waldendorf, J., Haddow, B., Birch, A., Colombo, P., and Martins, A. F. T. (2023). Hallucinations in large multilingual translation models.

Guest, E., Vidgen, B., Mittos, A., Sastry, N., Tyson, G., and Margetts, H. (2021). An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350.

Haddad, H., Mulki, H., and Oueslati, A. (2019). T-hsab: A tunisian hate speech and abusive dataset. In *Arabic Language Processing: From Theory to Practice: 7th International Conference, ICALP 2019, Nancy, France, October 16–17, 2019, Proceedings 7*, pages 251–263. Springer.

Hajmohammadi, M. S., Ibrahim, R., Selamat, A., and Fujita, H. (2015). Combination of active learning and self-training for cross-lingual sentiment classification with density analysis of unlabelled samples. *Information Sciences*, 317:67–77.

Hammersley, M. (2001). On 'systematic' reviews of research literatures: A 'narrative' response to evans & benefield. *British Educational Research Journal*, 27(5):543–554.

Hanu, L. and Unitary team (2020). Detoxify. Github. https://github.com/unitaryai/detoxify.

Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., and Kamar, E. (2022). ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Huang, X., Xing, L., Dernoncourt, F., and Paul, M. J. (2020). Multilingual Twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1440–1448, Marseille, France. European Language Resources Association.

Ibrohim, M. O. and Budi, I. (2018). A dataset and preliminaries study for abusive language detection in indonesian social media. *Procedia Computer Science*, 135:222–229.

Ibrohim, M. O. and Budi, I. (2019a). Multi-label hate speech and abusive language detection in indonesian twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57.

Ibrohim, M. O. and Budi, I. (2019b). Translated vs non-translated method for multilingual hate speech identification in twitter. *Int. J. Adv. Sci. Eng. Inf. Technol*, 9(4):1116–1123.

Ishmam, A. M. and Sharmin, S. (2019). Hateful speech detection in public facebook pages for the bengali language. In *2019 18th IEEE international conference on machine learning and applications (ICMLA)*, pages 555–560. IEEE.

Jacobs, C., Rakotonirina, N. C., Chimoto, E. A., Bassett, B. A., and Kamper, H. (2023). Towards hate speech detection in low-resource languages: Comparing asr to acoustic word embeddings on wolof and swahili.

Jahan, M. S. and Oussalah, M. (2021). A systematic review of hate speech automatic detection using natural language processing. *CoRR*, abs/2106.00742.

Jayanthi, S. M. and Gupta, A. (2021). SJ_AJ@DravidianLangTech-EACL2021: Task-adaptive pre-training of multilingual BERT models for offensive language identification. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 307–312, Kyiv. Association for Computational Linguistics.

Jiang, A. and Zubiaga, A. (2021). Cross-lingual capsule network for hate speech detection in social media. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, HT '21, page 217–223, New York, NY, USA. Association for Computing Machinery.

Kar, P. and Debbarma, S. (2023). Multilingual hate speech detection sentimental analysis on social media platforms using optimal feature extraction and hybrid diagonal gated recurrent neural network. *The Journal of Supercomputing*, pages 1–32.

Keung, P., Lu, Y., and Bhardwaj, V. (2019). Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and NER. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1355–1360, Hong Kong, China. Association for Computational Linguistics.

Khairy, M., Mahmoud, T. M., and Abd-El-Hafeez, T. (2021). Automatic detection of cyberbullying and abusive language in arabic content on social networks: A survey. *Procedia Computer Science*, 189:156–166. AI in Computational Linguistics.

Khairy, M., Mahmoud, T. M., and Abd-El-Hafeez, T. (2023). The effect of rebalancing techniques on the classification performance in cyberbullying datasets. *Neural Computing and Applications*, pages 1–17.

Koshiry, A. M. E., Eliwa, E. H. I., Abd El-Hafeez, T., and Omar, A. (2023). Arabic toxic tweet classification: Leveraging the arabert model. *Big Data and Cognitive Computing*, 7(4).

Kovács, G., Alonso, P., and Saini, R. (2021). Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources. *SN Computer Science*, 2:1–15.

Kumar, A., Dikshit, S., and Albuquerque, V. H. C. (2021). Explainable artificial intelligence for sarcasm detection in dialogues. *Wireless Communications and Mobile Computing*, 2021:1–13.

Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2020). Evaluating aggression identification in social media. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 1–5.

Kumar, R., Ojha, A. K., Zampieri, M., and Malmasi, S., editors (2018a). *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kumar, R., Reganti, A. N., Bhatia, A., and Maheshwari, T. (2018b). Aggression-annotated corpus of

hindi-english code-mixed data. *arXiv preprint arXiv:1803.09402*.

Leite, J. A., Silva, D. F., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. *arXiv preprint arXiv:2010.04543*.

Ljubešić, N., Erjavec, T., and Fišer, D. (2018). Datasets of slovene and croatian moderated news comments. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pages 124–131.

Luu, S. T., Nguyen, K. V., and Nguyen, N. L.-T. (2021). A large-scale dataset for hate speech detection on vietnamese social media texts. In *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices: 34th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2021, Kuala Lumpur, Malaysia, July 26–29, 2021, Proceedings, Part I 34*, pages 415–426. Springer.

Madhu, H., Satapara, S., Modha, S., Mandl, T., and Majumder, P. (2023). Detecting offensive speech in conversational code-mixed dialogue on social media: A contextual dataset and benchmark experiments. *Expert Systems with Applications*, 215:119342.

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., and Patel, A. (2019). Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th forum for information retrieval evaluation*, pages 14–17.

Mandl, T., Modha, S., Shahi, G. K., Madhu, H., Satapara, S., Majumder, P., Schäfer, J., Ranasinghe, T., Zampieri, M., and Nandini, D. (2021). Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages. *arXiv preprint arXiv:2112.09301*.

Mathur, P., Shah, R., Sawhney, R., and Mahata, D. (2018). Detecting offensive tweets in hindi-english code-switched language. In *Proceedings of the sixth international workshop on natural language processing for social media*, pages 18–26.

Mazari, A. C. and Kheddar, H. (2023). Deep learning-based analysis of algerian dialect dataset targeted hate speech, offensive language and cyberbullying. *International Journal of Computing and Digital Systems*.

Mohapatra, S. K., Prasad, S., Bebarta, D. K., Das, T. K., Srinivasan, K., and Hu, Y.-C. (2021). Automatic hate speech detection in english-odia code mixed social media data using machine learning techniques. *Applied Sciences*, 11(18):8575.

Montariol, S., Riabi, A., and Seddah, D. (2022). Multilingual auxiliary tasks training: Bridging the gap between languages for zero-shot transfer of hate speech detection models. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 347–363, Online only. Association for Computational Linguistics.

Moon, J., Cho, W. I., and Lee, J. (2020). Beep! korean corpus of online news comments for toxic speech detection. *arXiv preprint arXiv:2005.12503*.

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

Mozafari, M., Farahbakhsh, R., and Crespi, N. (2022). Cross-lingual few-shot hate speech and offensive language detection using meta learning. *IEEE Access*, 10:14880–14896.

Mubarak, H., Hassan, S., and Chowdhury, S. A. (2022). Emojis as anchors to detect arabic offensive language and hate speech. *arXiv preprint arXiv:2201.06723*.

Mulki, H., Haddad, H., Ali, C. B., and Alshabani, H. (2019). L-hsab: A levantine twitter dataset for hate speech and abusive language. In *Proceedings of the third workshop on abusive language online*, pages 111–118.

Muti, A., Fernicola, F., and Barrón-Cedeño, A. (2023). UniBoe's at SemEval-2023 task 10: Model-agnostic strategies for the improvement of hate-tuned and generative models in the classification of sexist posts. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1138–1147, Toronto, Canada. Association for Computational Linguistics.

Nascimento, G., Carvalho, F., Cunha, A. M. d., Viana, C. R., and Guedes, G. P. (2019). Hate speech detection using brazilian imageboards. In *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web*, pages 325–328.

Nozza, D. (2021). Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

Ocampo, N., Cabrio, E., and Villata, S. (2023). Playing the part of the sharp bully: Generating adver-

sarial examples for implicit hate speech detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2758–2772, Toronto, Canada. Association for Computational Linguistics.

Ollagnier, A., Cabrio, E., Villata, S., and Blaya, C. (2022). Cyberagressionado-v1: a dataset of annotated online aggressions in french collected through a role-playing game. In *Language Resources and Evaluation Conference*.

Omar, A. and Abd El-Hafeez, T. (2023). Quantum computing and machine learning for arabic language sentiment classification in social media. *Scientific Reports*, 13(1):17305.

Omar, A., Mahmoud, T. M., and Abd-El-Hafeez, T. (2020). Comparative performance of machine learning and deep learning algorithms for arabic hate speech detection in osns. In Hassanien, A.-E., Azar, A. T., Gaber, T., Oliva, D., and Tolba, F. M., editors, *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020)*, pages 247–257, Cham. Springer International Publishing.

Ombui, E., Muchemi, L., and Wagacha, P. (2019). Hate speech detection in code-switched text messages. In *2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 1–6. IEEE.

Otter, D. W., Medina, J. R., and Kalita, J. K. (2018). A survey of the usages of deep learning in natural language processing. *CoRR*, abs/1807.10854.

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., and Yeung, D.-Y. (2019). Multilingual and multi-aspect hate speech analysis. *arXiv preprint arXiv:1908.11049*.

Pamungkas, E. W., Basile, V., and Patti, V. (2021a). A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. *Inf. Process. Manage.*, 58(4).

Pamungkas, E. W., Basile, V., and Patti, V. (2021b). Towards multidomain and multilingual abusive language detection: A survey. *Personal Ubiquitous Comput.*, 27(1):17–43.

Pamungkas, E. W. and Patti, V. (2019). Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370, Florence, Italy. Association for Computational Linguistics.

Pavlopoulos, J., Malakasiotis, P., and Androutsopoulos, I. (2017). Deep learning for user comment moderation. *arXiv preprint arXiv:1705.09993*.

Pelicon, A., Shekhar, R., Martinc, M., Škrlj, B., Purver, M., and Pollak, S. (2021a). Zero-shot cross-lingual content filtering: Offensive language and hate speech detection. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 30–34, Online. Association for Computational Linguistics.

Pelicon, A., Shekhar, R., Škrlj, B., Purver, M., and Pollak, S. (2021b). Investigating cross-lingual training for offensive language detection. *PeerJ Computer Science*, 7:e559.

Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., and Camacho-Collados, M. (2019). Detecting and monitoring hate speech in twitter. *Sensors*, 19(21):4654.

Pfeiffer, J., Goyal, N., Lin, X., Li, X., Cross, J., Riedel, S., and Artetxe, M. (2022). Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

Pham, N.-Q., Niehues, J., Ha, T.-L., and Waibel, A. (2019). Improving zero-shot translation with language-independent constraints. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 13–23, Florence, Italy. Association for Computational Linguistics.

Pikuliak, M., Šimko, M., and Bieliková, M. (2021). Cross-lingual learning for text processing: A survey. *Expert Systems with Applications*, 165:113765.

Pitenis, Z., Zampieri, M., and Ranasinghe, T. (2020). Offensive language identification in greek. *arXiv preprint arXiv:2003.07459*.

Plaza-Del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A., and Martín-Valdivia, M. T. (2021). A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access*, 9:112478–112489.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

Ptaszynski, M., Pieciukiewicz, A., and Dybała, P. (2019). Results of the poleval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in polish twitter.

Putra, I. F. and Purwarianti, A. (2020). Improving indonesian text classification using multilingual language model. *CoRR*, abs/2009.05713.

Rahman, M. M., Balakrishnan, D., Murthy, D., Kutlu, M., and Lease, M. (2021). An information retrieval approach to building datasets for hate speech detection. *arXiv preprint arXiv:2106.09775*.

Rajamanickam, S., Mishra, P., Yannakoudakis, H., and Shutova, E. (2020). Joint modelling of emotion and abusive language detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4270–4279, Online. Association for Computational Linguistics.

Ranasinghe, T., Anuradha, I., Premasiri, D., Silva, K., Hettiarachchi, H., Uyangodage, L., and Zampieri, M. (2022). Sold: Sinhala offensive language dataset. *arXiv preprint arXiv:2212.00851*.

Ranasinghe, T. and Zampieri, M. (2020). Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online. Association for Computational Linguistics.

Ranasinghe, T. and Zampieri, M. (2021a). An evaluation of multilingual offensive language identification methods for the languages of india. *Information*, 12(8).

Ranasinghe, T. and Zampieri, M. (2021b). Multilingual offensive language identification for low-resource languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(1).

Ranasinghe, T. and Zampieri, M. (2023). Teacher and student models of offensive language in social media. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3910–3922, Toronto, Canada. Association for Computational Linguistics.

Riabi, A., Montariol, S., and Seddah, D. (2022). Tâches auxiliaires multilingues pour le transfert de modèles de détection de discours haineux (multilingual auxiliary tasks for zero-shot cross-lingual transfer of hate speech detection). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 413–423, Avignon, France. ATALA.

Rizwan, H., Shakeel, M. H., and Karim, A. (2020). Hate-speech and offensive language detection in roman urdu. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 2512–2522.

Röttger, P., Nozza, D., Bianchi, F., and Hovy, D. (2022a). Data-efficient strategies for expanding hate speech detection into under-resourced languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5674–5691, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Röttger, P., Seelawi, H., Nozza, D., Talat, Z., and Vidgen, B. (2022b). Multilingual HateCheck: Functional tests for multilingual hate speech detection models. In Narang, K., Mostafazadeh Davani, A., Mathias, L., Vidgen, B., and Talat, Z., editors, *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., and Pierrehumbert, J. (2021). HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Roy, S. G., Narayan, U., Raha, T., Abid, Z., and Varma, V. (2021a). Leveraging multilingual transformers for hate speech detection. In *Fire*.

Roy, S. G., Narayan, U., Raha, T., Abid, Z., and Varma, V. (2021b). Leveraging multilingual transformers for hate speech detection. *CoRR*, abs/2101.03207.

Roychowdhury, S. and Gupta, V. (2023). Data-efficient methods for improving hate speech detection. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 125–132, Dubrovnik, Croatia. Association for Computational Linguistics.

Safi Samghabadi, N., Hatami, A., Shafaei, M., Kar, S., and Solorio, T. (2020). Attending the emotions to detect online abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 79–88, Online. Association for Computational Linguistics.

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., and Stranisci, M. (2018). An italian twitter corpus of hate speech against immigrants. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Satapara, S., Modha, S., Mandl, T., Madhu, H., and Majumder, P. (2021). Overview of the hasoc subtrack at fire 2021: Conversational hate speech detection in code-mixed language. *Working Notes of FIRE*.

Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing.

In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Sharif, O., Hossain, E., and Hoque, M. M. (2021). Nlp-cuet@dravidianlangtech-eacl2021: Offensive language detection from multilingual code-mixed text using transformers. *CoRR*, abs/2103.00455.

Shi, X., Liu, X., Xu, C., Huang, Y., Chen, F., and Zhu, S. (2022). Cross-lingual offensive speech identification with transfer learning for low-resource languages. *Computers and Electrical Engineering*, 101:108005.

Sigurbergsson, G. I. and Derczynski, L. (2019). Offensive language and hate speech detection for danish. *arXiv preprint arXiv:1908.04531*.

Snell, J., Swersky, K., and Zemel, R. (2017). Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Stappen, L., Brunn, F., and Schuller, B. W. (2020). Cross-lingual zero- and few-shot hate speech detection utilising frozen transformer language models and AXEL. *CoRR*, abs/2004.13850.

Steinberger, J., Brychcín, T., Hercig, T., and Krejzl, P. (2017). Cross-lingual flames detection in news discussions. In *RANLP*, pages 694–700.

Takawane, G., Phaltankar, A., Patwardhan, V., Patil, A., Joshi, R., and Takalikar, M. S. (2023). Leveraging language identification to enhance code-mixed text classification.

Tita, T. and Zubiaga, A. (2021). Cross-lingual hate speech detection using transformer models. *CoRR*, abs/2111.00981.

Vadakkekara Suresh, G., Chakravarthi, B. R., and McCrae, J. P. (2022). Meta-learning for offensive language detection in code-mixed texts. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '21, page 58–66, New York, NY, USA. Association for Computing Machinery.

Vashistha, N. and Zubiaga, A. (2021). Online multilingual hate speech detection: Experimenting with hindi and english social media. *Information*, 12(1).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vidgen, B. and Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., and Margetts, H. (2019). Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Vinyals, O., Blundell, C., Lillicrap, T., and Wierstra, D. (2016). Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

Vitiugin, F., Senarath, Y., and Purohit, H. (2021). Efficient detection of multilingual hate speech by using interactive attention network with minimal human feedback. In *13th ACM Web Science Conference 2021*, WebSci '21, page 130–138, New York, NY, USA. Association for Computing Machinery.

Vu, X.-S., Vu, T., Tran, M.-V., Le-Cong, T., and Nguyen, H. (2020). Hsd shared task in vlsp campaign 2019: Hate speech detection for social good. *arXiv preprint arXiv:2007.06493*.

Wang, C.-C., Day, M.-Y., and Wu, C.-L. (2022). Political hate speech detection and lexicon building: A study in taiwan. *IEEE Access*, 10:44337–44346.

Wang, R., Tan, X., Luo, R., Qin, T., and Liu, T.-Y. (2021). A survey on low-resource neural machine translation. In Zhou, Z.-H., editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4636–4643. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Wang, S., Liu, J., Ouyang, X., and Sun, Y. (2020). Galileo at SemEval-2020 task 12: Multi-lingual learning for offensive language identification using pre-trained language models. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1448–1455, Barcelona (online). International Committee for Computational Linguistics.

Weller, O., Marone, M., Braverman, V., Lawrie, D., and Van Durme, B. (2022). Pretrained models for multilingual federated learning. *arXiv preprint arXiv:2206.02291*.

Wigand, C. and Voin, M. (2017). Speech by commissioner jourová—10 years of the eu fundamental rights agency: A call to action in defence of fundamental rights, democracy and the rule of law.

Yadav, A., Chandel, S., Chatufale, S., and Bandhakavi, A. (2023). Lahm: Large annotated dataset for multi-domain and multilingual hate speech identification. *arXiv preprint arXiv:2304.00913*.

Yang, K., Jang, W., and Cho, W. I. (2022). Apeach: Attacking pejorative expressions with analysis on crowd-generated hate speech evaluation datasets. *arXiv preprint arXiv:2202.12459*.

Yin, W. and Zubiaga, A. (2021). Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

zahra El-Alami, F., Ouatik El Alaoui, S., and En Nahnahi, N. (2022). A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model. *Journal of King Saud University - Computer and Information Sciences*, 34(8, Part B):6048–6056.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., and Çöltekin, Ç. (2020). Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*.

Zampieri, M., Ranasinghe, T., Chaudhari, M., Gaikwad, S., Krishna, P., Nene, M., and Paygude, S. (2022). Predicting the type and target of offensive social media posts in marathi. *Social Network Analysis and Mining*, 12(1):77.

Zhang, B., Williams, P., Titov, I., and Sennrich, R. (2020). Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Zia, H. B., Castro, I., Zubiaga, A., and Tyson, G. (2022). Improving zero-shot cross-lingual hate speech detection with pseudo-label fine-tuning of transformer language models. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):1435–1439.

Zoph, B. and Le, Q. V. (2016). Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*.