Does Machine Translation Impact Offensive Language Identification? The Case of Indo-Aryan Languages

[1]Alphaeus Dmonte, [2]Shrey Satapara, [1]Rehab Alsudais
[3]Tharindu Ranasinghe, [1]Marcos Zampieri

[1]George Mason University, USA
[2]IIT Hyderabad, India
[3]Lancaster University, UK

Abstract

The accessibility to social media platforms can be improved with the use of machine translation (MT). Non-standard features present in user-generated on social media content such as hashtags, emojis, and alternative spellings can lead to mistranslated instances by the MT systems. In this paper, we investigate the impact of MT on offensive language identification in Indo-Aryan languages. We use both original and MT datasets to evaluate the performance of various offensive language models. Our evaluation indicates that offensive language identification models achieve superior performance on original data than on MT data, and that the models trained on MT data identify offensive language more precisely on MT data than the models trained on original data.

1 Introduction

Social media platforms have seen rapid growth in popularity in recent years. Several types of content are shared across these platforms such as product reviews, discussions on topics ranging from entertainment to politics, and general commentary on topics like health, social issues, etc. However, some of this content shared on social media platforms can include offensive and toxic language, misinformation, etc., and can be harmful to individuals. Offensive content can include pejorative language (Dinu et al., 2021), cyberbullying (Rosa et al., 2019), hate speech (Röttger et al., 2021), targeted insults (Kwok and Wang, 2013; Basile et al., 2019), and many others. Machine-learning approaches have been used to moderate such content and mitigate its spread (Weerasooriya et al., 2023).

The content posted on these social media platforms is written in several languages and dialects. While most of the research on offensive language identification has been grounded around high-resources languages like English, German, and Spanish (Zampieri et al., 2019b; Risch et al., 2021; Basile et al., 2019), recently, this research has been extended to low-resource languages, including the Indo-Aryan languages (Ranasinghe and Zampieri, 2020; Ranasinghe et al., 2023). Models such as transformers have achieved state-of-the-art performance on this task in most of the languages (Ranasinghe and Zampieri, 2023).

Various social media platforms translate content between languages using machine translation models (Gupta, 2021). However, the non-standard nature of the user-generated content, the MT content is often mistranslated (Lohar et al., 2017). As most of the models and high- and low-resource languages are trained on the original instances from social media platforms, MT of such content can be challenging for offensive language detection models, especially in low-resource languages. Furthermore, there has not been much research on offensive language detection of machine-translated content.

In this paper, we fill this gap by evaluating the performance of state-of-the-art offensive language identification models on MT content in three low-resource Indo-Aryan languages: Hindi, Marathi, and Sinhala. These three languages have been spoken by millions of speakers each and have been widely used on social media platforms, despite their low-resource nature. The presence of language experts to validate the translations was another factor influencing the language choice. This work address the following research questions:

- **RQ1:** To what extent does MT impact the performance of existing state-of-the-

Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025), pages 460–468
January 20, 2025.

art offensive language identification models?

- **RQ2:** How does training on translated data affect the model with respect to the performance?

To answer these questions, we compare the performance of offensive language identification models, trained and evaluated on translated and non-translated data for the three aforementioned languages in various scenarios. The main contribution of this paper is a comprehensive evaluation of state-of-the-art offensive language identification models on machine-translated content in Indo-Aryan languages. We evaluate the performance of fine-tuned BERT-based models in four scenarios.

The remainder of this paper is structured as follows: Section 2 presents related work on MT and social media, Section 3 presents the datasets used in this paper, Section 4 presents the models used in our experiments and Section 5 presents the results of the experiments and a comprehensive discussion and analysis. Finally, Section 6 concludes this paper and presents avenues for future work.

2 Related Work

Multiple studies have focused on offensive language identification of multilingual content (Ranasinghe and Zampieri, 2021; Nozza, 2021; Jiang and Zubiaga, 2024; Mnassri et al., 2024). While there have been various competitions organized for the task (Basile et al., 2019; Zampieri et al., 2020; Satapara et al., 2022; Ranasinghe et al., 2023). Transformer models such as BERT (Devlin et al., 2019) have achieved state-of-the-art performance in these datasets. Studies have also been conducted to understand the generalizability of the offensive language models (Dmonte et al., 2024). Zampieri et al. (2023a) introduces a subsequent task that identifies offensive spans and their associated targets. The language understanding capabilities of the recent Large Language Models (LLMs) have highlighted their ability to identify offensive content. These models, when evaluated with in-context learning approaches on English offensive language datasets, achieve comparable performance to the fine-tuned BERT-based models (Zampieri et al., 2023b). However, these studies do not consider the machine-translated content. MT is however used to augment offensive language datasets with more instances to train machine learning models (Beddiar et al., 2021). El-Alami et al. (2022) translate English tweets to Arabic using Google machine translation API and train models to classify offensive tweets. Dmonte et al. (2025) develop machine-translated datasets for five high- and low-resource languages and use these to understand the impact of MT on offensive language detection.

Lohar et al. (2018) and Saadany et al. (2021) explored the impact of MT on sentiment analysis task, while a Naive Bayes model was trained for sentiment analysis of English reviews translated to Greek and Italian (Bilianos and Mikros, 2023). Other works that utilized MT for sentiment analysis include a metric to assess the sentiment similarity between MT content and reference content () to identify the mistranslations of sentiments. While Si et al. (2019) propose a Neural Machine Translation(NMT) model that utilizes a two-step approach which first generates a sentiment label and then uses this sentiment label to train an NMT model. The work by Saadany and Orăsan (2020) translates Arabic to English using three NMT models, where two of these models are sentiment sensitive.

3 Data

The original datasets used in the paper are presented in Table 1. For English, we use the OLID (Zampieri et al., 2019a), a popular offensive language dataset in English, described in Section 3.1. While for Hindi, Marathi, and Sinhala, the datasets described in Section 3 were used. We further use the OLID dataset translated to the three languages from the MT-Offense (Dmonte et al., 2025) benchmark. The translated (MT) and original versions of the datasets used to evaluate the performance of offensive language identification systems in the following scenarios:

1. **ORIG-ORIG:** Original training - original test;
2. **TRAN-TRAN:** MT training - MT test;
3. **ORIG-TRAN:** Original training - MT test;

Language	Dataset	Training Inst.	Training OFF %	Testing Inst.	Testing OFF %	Sources	Reference
English	OLID	13,240	0.33	860	0.27	T	Zampieri et al. (2019a)
Hindi	HASOC-2019	4,665	0.53	1,318	0.46	T, F	Mandl et al. (2019)
Marathi	MOLD	2,499	0.35	626	0.33	T	Gaikwad et al. (2021)
Sinhala	SOLD	7,500	0.42	2,500	0.41	T	Ranasinghe et al. (2024)

Table 1: The datasets with the language, the number of instances (Inst.) in the training and testing sets, the % of offensive instances in each set (OFF %), the data source, and the reference. Data sources are represented by F - Facebook, I - Instagram, T - Twitter, and Y - YouTube.

4. **TRAN-ORIG:** MT training - original test.

3.1 English Dataset

We used OLID (Zampieri et al., 2019a), the official dataset of the SemEval-2019 Task 6 (OffensEval) (Zampieri et al., 2019b). The dataset consists of 14,100 tweets manually annotated according to the following hierarchical taxonomy:

Level A if the tweet is offensive (OFF) or not (NOT).
Level B if the tweet is offensive is it targeted (TIN) or untargeted (UNT).
Level C If the tweet is targeted is it targeted to a group (GRP), individual (IND), or others (OTH).

We only use OLID level A in our experiments. The labels on the language-specific datasets are mapped to OLID level A. This dataset was chosen due to the flexibility provided by its general three-level hierarchical taxonomy. The OFF class contains all types of offensive content, from general profanity to hate speech, while the NOT class contains non-offensive examples. The three languages we experiment with in this paper, Hindi, Marathi, and Sinhala, also follow the OLID taxonomy.

3.2 Indo-Aryan Languages

The following datasets for the Indo-Aryan languages were used in our experiments. These datasets are highly popular among the community and were released recently as part of different shared tasks.

Hindi (Mandl et al., 2019) For Hindi, we used the official dataset of the HASOC-2019 competition (Mandl et al., 2019). The hate speech and offensive language dataset was annotated with one of the two labels: Hate and Offensive (HOF) or Non Hate-Offensive (NOT). To maintain uniformity in all our datasets, we map the HOF label is mapped to the OLID Offensive label.

Marathi (Gaikwad et al., 2021) MOLD, which was an official dataset in HASOC 2021 (Modha et al., 2021) is used for Marathi. This dataset has been annotated following the OLID taxonomy.

Sinhala (Ranasinghe et al., 2024) We use SOLD as our Sinhala offensive language detection dataset. SOLD is the official dataset in HASOC 2023 (Ranasinghe et al., 2023; Satapara et al., 2023). It has also been annotated following the OLID taxonomy.

3.3 *MT-Offense*

MT-Offense is a benchmark dataset comprising the OLID dataset machine translated into five low- and high-resource languages. In this work, we use the three Indo-Aryan languages: Hindi, Marathi, and Sinhala. The dataset was created by translating the OLID dataset into the aforementioned languages with three translation models for each language. opus-mt-en-* (Tiedemann and Thottingal, 2020), henceforth **Trans-1**, m2m100_1.2B (Fan et al., 2021), henceforth **Trans-2**, and nllb-200-3.3B (Costa-jussà et al., 2022), henceforth Trans-3 were used to translate the dataset. Since opus-mt does not support English to Sinhala translation, Mbart-50-English-sinhala-nmt, henceforth **Trans-1**, was used to translate the data set into Sinhala.

4 Approach

The following sections define the models and hyperparameters used in our experiments.

Language	Model	Source	Translation	Label	Score
Hindi	Trans-1	@USER Fist pump was for the troops.	- सेना के लिए FREARS पंप था.	NOT	0.547
	Trans-2	@USER They are really upset about this election loss.	@USER वे इस चुनाव के नुकसान के बारे में वास्तव में परेशान हैं।	NOT	0.751
	Trans-3	@USER She is gorgeous	@USER वह बहुत सुंदर है	NOT	0.755
Marathi	Trans-1	@USER She is beautiful	तिची आई सुंदर आहे	NOT	0.719
	Trans-2	@USER He is a Professional liar	@USER तो एक व्यावसायिक खोटारडे आहे	OFF	0.683
	Trans-3	@USER She is a nut case	@USER ती वेडी आहे	OFF	0.618
Sinhala	Trans-1	@USER @USER Oh noes! Tough shit.	අපොයි නෑ!	OFF	0.702
	Trans-2	@USER Brennan sure as hell is.	@Brennan සැබැවින්ම අපායක් ලෙස.	OFF	0.599
	Trans-3	@USER why do all crazy liberals have CRAZY EYES? LOL URL	@USER ඇයි හැම පිස්සු ලිබරල් කෙනෙකුටම පිස්සු ඇස් තියෙන්නේ?	OFF	0.680

Table 2: Example instances from the translated datasets. The model refers to the translation model used. The source language is English and the target is the translation for the corresponding language. The label is the gold standard label for the original OLID instance. A quality score for each instance is calculated using the TransQuest quality estimation model and displayed in the column Score.

4.1 Models

Several monolingual and multilingual models were fine-tuned using both the source language and the translated datasets. Models fine-tuned included ***XLM-R*** (Conneau et al., 2020), henceforth **xlm** and language-specific models ***hindi-bert-scratch*** (Joshi, 2022a), henceforth **mono**, ***marathi-bert-scratch*** (Joshi, 2022b), henceforth **mono**, and ***SinBERT-large*** (Dhananjaya et al., 2022), henceforth **mono**, for Hindi, Marathi, and Sinhala respectively. To investigate if the models pre-trained with the Twitter data affect the performance on the offensive language identification task, we experiment with ***bernice*** (DeLucia et al., 2022) and ***twhin-bert-base*** (Zhang et al., 2023), henceforth **twhin**, models that are specifically pre-trained on Twitter data.

4.2 Experimental Setup

Since the datasets do not include a predefined development set, we divide the training dataset into training and development sets using an 80-20 split. The hyperparameter values used to fine-tune the models are defined in Table 3.

4.3 Evaluation Metrics

To evaluate the performance of the models, we use standard evaluation metrics used in text categorization. The main metric used is the F1 score, which is the harmonic mean of precision and recall scores obtained in the test predictions. We employ the macro-averaged F1 score to account for the imbalanced nature of the datasets to ensure a balanced assessment across all classes.

Parameter	Value
epochs	3
batch size	8
learning rate	1e-5
adam epsilon	1e-8
warmup ratio	0.06
warmup steps	0
max grad norm	1.0
gradient accumulation steps	1

Table 3: Training parameter specifications for BERT-based models.

5 Results and Discussion

In this section, we present the results of the different models evaluated in the four aforementioned training and testing scenarios.

5.1 ORIG-ORIG

Table 4 shows the performance of the models trained and evaluated with the source language datasets. The results indicate that the models pre-trained on the Twitter data outperform other models in two of the three languages.

Language	xlm	mono	bernice	twhin
Hindi	0.808	0.794	**0.847**	0.823
Marathi	**0.889**	0.867	0.609	0.885
Sinhala	0.824	0.807	**0.834**	0.818

Table 4: F1 of the xlm, mono, bernice, and twhin models trained on **ORIG-ORIG**.

Language	Translated Training Dataset	Translated Evaluation Dataset											
		Trans-1				Trans-2				Trans-3			
		xlm	mono	bernice	twhin	xlm	mono	bernice	twhin	xlm	mono	bernice	twhin
Hindi	Trans-1	0.676	0.592	**0.679**	0.647	0.698	0.608	**0.712**	0.676	0.731	0.643	**0.741**	0.682
	Trans-2	0.645	0.600	**0.689**	0.636	0.712	0.641	**0.730**	0.719	0.707	0.631	**0.742**	0.684
	Trans-3	0.691	0.631	**0.691**	0.670	**0.727**	0.652	0.720	0.672	0.755	0.687	**0.775**	0.708
Marathi	Trans-1	0.419	0.419	**0.427**	0.419	0.419	0.419	**0.419**	0.419	0.419	0.419	**0.428**	0.419
	Trans-2	0.455	0.446	0.435	**0.489**	0.659	0.580	0.657	**0.726**	0.623	0.536	0.653	**0.708**
	Trans-3	0.480	**0.483**	0.451	0.479	**0.663**	0.580	0.660	0.666	0.723	0.653	0.716	**0.726**
Sinhala	Trans-1	0.712	0.423	0.720	**0.722**	**0.671**	0.430	0.578	0.640	**0.662**	0.475	0.581	0.632
	Trans-2	0.647	0.431	0.668	**0.693**	0.666	0.617	0.692	**0.697**	**0.679**	0.625	0.651	0.672
	Trans-3	**0.626**	0.432	0.539	0.545	**0.610**	0.559	0.579	0.600	0.705	0.657	0.671	**0.713**

Table 5: F1 score of all the models for **TRAN-TRAN**. The training dataset model and the evaluation dataset model refer to the OLID training and test dataset translated using the corresponding models.

5.2 TRAN-TRAN

The performance of all the models fine-tuned and evaluated with the translated datasets are shown in Table 5. The results indicate that the models pre-trained on the Twitter data perform better than the other models. Moreover, the models fine-tuned with the datasets translated with the Trans-3 models generally outperform the models that are trained with the Trans-1 and Trans-2 translation models.

5.3 ORIG-TRAN

In this scenario, the models are trained with the source language datasets and evaluated using the translated test datasets. We report the performance of the models in Table 6. As seen for Sinhala, the models pre-trained using the Twitter dataset and evaluated with datasets translated using Trans-2 and Trans-3 translation models underperform the Trans-1 translation model, while these models outperform or have comparable performance to the models evaluated with the Trans-1 dataset in Hindi and Marathi. The models evaluated using the Trans-3 translated test datasets generally had an overall better performance for all the languages.

5.4 TRAN-ORIG

The models trained on the translated datasets and evaluated using the source language datasets were evaluated in this scenario. Similar to the other scenarios, the models pre-trained on the Twitter data outperformed other models, with ***bernice*** performing the best for all the languages. The performance

Language	Model	Translated Evaluation Dataset		
		Trans-1	Trans-2	Trans-3
Hindi	xlm	0.496	0.593	0.611
	mono	0.542	0.564	0.583
	bernice	0.569	0.561	0.626
	twhin	**0.576**	**0.621**	**0.683**
Marathi	xlm	**0.454**	0.448	0.481
	mono	0.449	0.432	0.459
	bernice	0.418	0.486	0.511
	twhin	0.443	**0.518**	**0.528**
Sinhala	xlm	0.427	0.423	0.443
	mono	0.419	0.423	**0.479**
	bernice	**0.531**	**0.482**	0.475
	twhin	0.520	0.456	0.453

Table 6: F1 score of the models on **ORIG-TRAN**. The evaluation dataset model refers to the translated OLID test datasets translated using the corresponding models.

of the models in this scenario is reported in Table 7.

Overall, the multilingual models outperform the monolingual models in most of the scenarios. The models trained and evaluated with the source language datasets outperform the models trained on source language datasets and evaluated with the translated datasets. Mistranslations in the translated instances can contribute to the lower performance of the models.

The results also indicate that the Trans-2 and Trans-3 translation models perform better than the Trans-1 models for most languages. The translation quality of the Trans-3 model is superior to the other two models. Hence, the models trained with this data generally tend to outperform the models trained on the other

Language	Model	Translated Training Dataset		
		Trans-1	Trans-2	Trans-3
Hindi	xlm	0.687	0.655	0.743
	mono	0.552	0.542	0.577
	bernice	**0.827**	**0.830**	**0.822**
	twhin	0.779	0.776	0.767
Marathi	xlm	0.402	0.661	0.717
	mono	0.402	0.580	0.668
	bernice	**0.402**	**0.802**	**0.788**
	twhin	0.402	0.782	0.753
Sinhala	xlm	0.628	0.641	0.649
	mono	0.418	0.593	0.612
	bernice	**0.689**	**0.670**	**0.662**
	twhin	0.640	0.664	0.649

Table 7: F1 score of the models on **TRAN-ORIG**. The training dataset model refers to the translated OLID training datasets translated using the corresponding models.

translated datasets.

The TRAN-TRAN and ORIG-TRAN scenarios indicate that the models, when trained with the translated datasets, significantly outperform the models trained on the source language datasets. This performance difference can be attributed to the discrepancies during the translation of the data from one language to the other, like mistranslations or irrelevant translations that are captured by the models trained with the translated data. Hence, such models, when evaluated with the translated data, achieve a better performance. Furthermore, the multilingual models generally outperform the monolingual models for the offensive language detection task. The multilingual models can better capture the discrepancies in data, especially translation to some other language, compared to the monolingual models, as these models are pre-trained with data from several languages. The performance difference of the models in the ORIG-TRAN and TRAN-ORIG scenarios can also be attributed to this behavior of the models as indicated in Tables 6, 7.

6 Conclusion

In this work, we presented an evaluation of the impact of MT on offensive language detection from English to three Indo-Aryan languages; Hindi, Marathi, and Sinhala. Unlike the previous work that used MT to augment the existing datasets in low-resource languages, in this work, we study the effects of MT on offensive language detection.

We answer the two research questions posed in the introduction. For **RQ1**, the ORIG-TRANS experiments show that the models trained on the source language dataset and evaluated with the translated datasets misclassify the offensive text. This can be attributed to the cultural stereotypes and translation quality, including mistranslations, irrelevant translations, etc. Words that are considered offensive in one language may not be offensive in another, and the models trained on the translated language datasets may not identify such offensive words, leading the models to misclassify the text.

The *TRANS-TRANS* and *ORIG-TRANS* experiments answer the **RQ2**. The results indicate that models trained on the translated data outperform the models trained on the original data when evaluated with the translated datasets. This behavior can be attributed to the translation patterns learned during training, which, in turn, improves performance.

As mentioned in Section 5.4, the translations generated by the automatic machine translation models consist of several inaccuracies and errors. Such errors are propagated to the offensive language models trained using this translated data. Hence, content can be misclassified, limiting the purpose of deploying the models for social media moderation. Moreover, cultural stereotypes and language-specific contextual differences might prompt the models to misclassify the text. These factors need to be considered when training online content moderation models.

The inaccuracies and errors from MT are propagated to the offensive language detection models that are trained on the translated data. This can cause the models to misclassify the content, limiting the use of such models for social media content moderation. Furthermore, cultural stereotypes, along with language-specific contextual differences, might prompt misclassification. Factoring these issues during model training can improve the performance of the online content moderation models.

Limitations

In our experiments, we utilize three open-source translation models. We acknowledge that the inclusion of proprietary translation models and APIs like Google Translate API, as well as advanced LLMs like GPT-4, could potentially produce accurate and higher-quality translations. This, in turn, may improve the performance of the models trained on the translated datasets. Using LLMs like Llama-3 or GPT-4, which have demonstrated superior performance on several NLP tasks, may yield better results. We also acknowledge that potential biases may be introduced with the use of the MT system. While this is certainly a limitation, this is a common scenario faced by social media users who have to resort to MT to be able to understand content in languages they are not proficient in.

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63.

Djamila Romaissa Beddiar, Md Saroar Jahan, and Mourad Oussalah. 2021. Data expansion using back translation and paraphrasing for hate speech detection. *Online Social Networks and Media*, 24.

Dimitris Bilianos and George Mikros. 2023. Sentiment analysis in cross-linguistic context: How can machine translation influence sentiment classification? *Digital Scholarship in the Humanities*, 38:23–33.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Alexandra DeLucia, Shijie Wu, Aaron Mueller, Carlos Aguirre, Philip Resnik, and Mark Dredze. 2022. Bernice: a multilingual pre-trained encoder for Twitter. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6191–6205.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Vinura Dhananjaya, Piyumal Demotte, Surangika Ranathunga, and Sanath Jayasena. 2022. BERTifying Sinhala-A Comprehensive Analysis of Pre-trained Language Models for Sinhala Text Classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7377–7385.

Liviu P Dinu, Ioan-Bogdan Iordache, Ana Sabina Uban, and Marcos Zampieri. 2021. A computational exploration of pejorative language in social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3493–3498.

Alphaeus Dmonte, Tejas Arya, Tharindu Ranasinghe, and Marcos Zampieri. 2024. Towards generalized offensive language identification. *arXiv preprint arXiv:2407.18738*.

Alphaeus Dmonte, Shrey Satapara, Rehab Alsudais, Tharindu Ranasinghe, and Marcos Zampieri. 2025. On the effects of machine translation on offensive language detection. *Social Network Analysis and Mining*.

Fatima-zahra El-Alami, Said Ouatik El Alaoui, and Noureddine En Nahnahi. 2022. A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model. *Journal of King Saud University-Computer and Information Sciences*, 34:6048–6056.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond English-Centric Multilingual Machine Translation. *The Journal of Machine Learning Research*, 22(1).

Saurabh Sampatrao Gaikwad, Tharindu Ranasinghe, Marcos Zampieri, and Christopher Homan. 2021. Cross-lingual Offensive Language

Identification for Low Resource Languages: The Case of Marathi. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 437–443.

Ananya Gupta. 2021. User-controlled content translation in social media. In *26th International Conference on Intelligent User Interfaces-Companion*, pages 96–98.

Aiqi Jiang and Arkaitz Zubiaga. 2024. Cross-lingual offensive language detection: A systematic review of datasets, transfer approaches and challenges. *arXiv preprint arXiv:2401.09244*.

Raviraj Joshi. 2022a. L3Cube-HindBERT and DevBERT: Pre-Trained BERT Transformer models for Devanagari based Hindi and Marathi Languages. *arXiv preprint arXiv:2211.11418*.

Raviraj Joshi. 2022b. L3Cube-MahaCorpus and MahaBERT: Marathi Monolingual Corpus, Marathi BERT Language Models, and Resources. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 97–101.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, pages 1621–1622.

Pintu Lohar, Haithem Afli, and Andy Way. 2017. Maintaining sentiment polarity in translation of user-generated content. *Prague Bulletin of Mathematical Linguistics*, (108):73–84.

Pintu Lohar, Haithem Afli, and Andy Way. 2018. Balancing Translation Quality and Sentiment Preservation (Non-archival Extended Abstract). In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 81–88.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation*, pages 14–17.

Khouloud Mnassri, Reza Farahbakhsh, Razieh Chalehchaleh, Praboda Rajapaksha, Amir Reza Jafari, Guanlin Li, and Noel Crespi. 2024. A survey on multi-lingual offensive language detection. *PeerJ Computer Science*, 10:e1934.

Sandip Modha, Thomas Mandl, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Tharindu Ranasinghe, and Marcos Zampieri. 2021. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages and conversational hate speech. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 1–3.

Debora Nozza. 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.

Tharindu Ranasinghe, Isuri Anuradha, Damith Premasiri, Kanishka Silva, Hansi Hettiarachchi, Lasitha Uyangodage, and Marcos Zampieri. 2024. Sold: Sinhala offensive language dataset. *Language Resources and Evaluation*, pages 1–41.

Tharindu Ranasinghe, Koyel Ghosh, Aditya Shankar Pal, Apurbalal Senapati, Alphaeus Eric Dmonte, Marcos Zampieri, Sandip Modha, and Shrey Satapara. 2023. Overview of the HASOC subtracks at FIRE 2023: Hate speech and offensive content identification in assamese, bengali, bodo, gujarati and sinhala. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 13–15.

Tharindu Ranasinghe and Marcos Zampieri. 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.

Tharindu Ranasinghe and Marcos Zampieri. 2021. Multilingual offensive language identification for low-resource languages. *Transactions on Asian and Low-Resource Language Information Processing*, 21:1–13.

Tharindu Ranasinghe and Marcos Zampieri. 2023. A Text-to-Text Model for Multilingual Offensive Language Identification. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 375–384, Nusa Dua, Bali. Association for Computational Linguistics.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12, Duesseldorf, Germany. Association for Computational Linguistics.

Hugo Rosa, N Pereira, Ricardo Ribeiro, Paula Costa Ferreira, Joao Paulo Carvalho, S Oliveira, Luísa Coheur, Paula Paulino, AM Veiga Simão, and Isabel Trancoso. 2019.

Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93:333–345.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 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.

Hadeel Saadany and Constantin Orăsan. 2020. Is it Great or Terrible? Preserving Sentiment in Neural Machine Translation of Arabic Reviews. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 24–37.

Hadeel Saadany, Constantin Orăsan, Emad Mohamed, and Ashraf Tantavy. 2021. Sentiment-Aware Measure (SAM) for Evaluating Sentiment Transfer by Machine Translation Systems. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1217–1226.

Shrey Satapara, Hiren Madhu, Tharindu Ranasinghe, Alphaeus Eric Dmonte, Marcos Zampieri, Pavan Pandya, Nisarg Shah, Sandip Modha, Prasenjit Majumder, and Thomas Mandl. 2023. Overview of the hasoc subtrack at fire 2023: Hate-speech identification in sinhala and gujarati. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation (Working notes)*, pages 344–350.

Shrey Satapara, Prasenjit Majumder, Thomas Mandl, Sandip Modha, Hiren Madhu, Tharindu Ranasinghe, Marcos Zampieri, Kai North, and Damith Premasiri. 2022. Overview of the hasoc subtrack at fire 2022: Hate speech and offensive content identification in english and indo-aryan languages. In *Proceedings of the 14th annual meeting of the forum for information retrieval evaluation*, pages 4–7.

Chenglei Si, Kui Wu, Aiti Aw, and Min-Yen Kan. 2019. Sentiment aware neural machine translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 200–206.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT–building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480.

Tharindu Weerasooriya, Sujan Dutta, Tharindu Ranasinghe, Marcos Zampieri, Christopher Homan, and Ashiqur Khudabukhsh. 2023. Vicarious Offense and Noise Audit of Offensive Speech Classifiers: Unifying Human and Machine Disagreement on What is Offensive. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11648–11668.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.

Marcos Zampieri, Skye Morgan, Kai North, Tharindu Ranasinghe, Austin Simmmons, Paridhi Khandelwal, Sara Rosenthal, and Preslav Nakov. 2023a. Target-based offensive language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 762–770.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447.

Marcos Zampieri, Sara Rosenthal, Preslav Nakov, Alphaeus Dmonte, and Tharindu Ranasinghe. 2023b. OffensEval 2023: Offensive language identification in the age of Large Language Models. *Natural Language Engineering*, 29(6):1416–1435.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2023. TwHIN-BERT: A Socially-Enriched Pre-trained Language Model for Multilingual Tweet Representations at Twitter. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5597–5607.