Data-Efficient Approaches for Hate speech Detection Tasks in Under-Resourced Languages

Anonymous ACL submission

Abstract

This paper addresses the challenges of hate speech detection in under-resourced languages, where the scarcity of labeled datasets limits the effectiveness of existing models. While most current hate speech detection models are designed for high-resource languages like English, they often fail to perform adequately for languages with fewer resources. This paper explores data-efficient approaches to bridge this gap by utilizing cross-lingual transfer learning, few-shot learning, zero-shot learning, and mixed-up data augmentation techniques. By leveraging these methods, the study demonstrates how knowledge from high-resource languages can be transferred to low-resource languages, improving model performance even in data-scarce scenarios. The experimental results, based on a variety of languages, show significant improvements in hate speech detection accuracy and generalization, offering a robust framework for expanding NLP capabilities to under-represented languages.

1 Introduction

Digital media platforms such as Facebook, Twitter, and Reddit have significantly influenced various aspects of our daily lives in recent years. This influence extends to the propagation of hate speech, the spreading of rumors, conspiracy theories, and the promotion of hatred, xenophobia, racism, and prejudice. These negative phenomena are not only societal concerns but have also been identified as contributing factors to conflicts and wars.

To address hate speech using natural language processing (NLP) techniques, researchers leverage computational linguistics and machine learning algorithms to detect, classify, and mitigate instances of hate speech in text data. NLP techniques involve the collection of large datasets containing both hate speech and non-hate speech examples, which are then used to train machine learning models. These models include a variety of approaches, from traditional machine learning algorithms to more advanced deep learning techniques (Schmidt and Wiegand, 2017).

Most models achieve high accuracy in downstream tasks when provided with sufficient training data. However, developing conversational AI systems or other specialized linguistic tools for many new languages is resource-intensive and not feasible for the over 6,500 languages spoken worldwide (Schuster et al., 2018). Therefore, methods enabling transfer learning from high-resource to low-resource languages are desirable. The lack of annotated multilingual datasets has hindered the development of sophisticated cross-lingual transfer methods for intent detection and slot filling.

Despite the advancements, these deep learning algorithms are often data-hungry, requiring substantial amounts of labeled data to achieve high performance. This reliance on extensive datasets motivates the development of data-efficient algorithms that can perform well even with limited labeled examples. Such approaches are crucial for applying hate speech detection technologies to low-resource languages and scenarios where annotated data is scarce. Efforts in transfer learning, semi-supervised learning, and few-shot learning are particularly promising in addressing these challenges, enabling the extension of state-of-the-art hate speech detection methods to a broader range of languages and contexts.

To tackle this challenge, exploring data-efficient approaches that leverage cross-lingual transfer learning have been proposed. Large multilingual language models can be fine-tuned on a source language and then applied to a target language using mixed-up data augmentation techniques. This involves taking a model fine-tuned on a high-resource language and applying it to an augmented dataset in the target language. These techniques can address the problem of language-specific data scarcity by effectively transferring knowledge from well-

resourced languages to low-resource languages.

Cross-Lingual Transfer Learning involves training a model on a high-resource language (source language) and then adapting it to perform well on a low-resource language (target language). This process leverages the shared characteristics and structures between languages, enabling the model to generalize learned features from the source language to the target language. When direct linguistic resources between the source and target languages are unavailable, transfer learning methods can be utilized. Since obtaining training data for all languages of interest is prohibitive, cross-lingual transfer learning offers the possibility of learning models for a target language using annotated data from other languages (source languages).

Transfer learning for cross-lingual cases is a type of transductive transfer learning, where the input domains of the source and target languages are different (Pan and Yang, 2009) due to each language having its own vocabulary space. When the input space is the same, lower layers of hierarchical models can be shared for knowledge transfer (Collobert et al., 2011; Kim et al., 2015), but this approach is not directly applicable when the input spaces differ.

2 Related Work

Deep learning-based methods have demonstrated impressive results in detecting hate speech. Badjatiya et al. (2017); Gambäck and Sikdar (2017); Zhang et al. (2018) and Kshirsagar et al. (2018) illustrate the effectiveness of various neural network architectures, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), in identifying and classifying hate speech within large datasets. These methods have proven capable of capturing complex patterns in textual data, significantly improving the accuracy of hate speech detection compared to traditional machine learning approaches.

In recent years, Transformer-based models have further advanced the state-of-the-art in hate speech detection. These models, including BERT, HateBERT, and their variants, leverage vast amounts of unlabeled data through self-supervised learning techniques, allowing them to understand context and nuances in language more effectively. Caselli et al. (2020); Rangel et al. (2021) and Yigezu et al. (2023) showcase how these Transformer-based models can outperform previous methods by utilizing sophisticated attention mechanisms to focus on relevant parts of the input text.

Hate speech has emerged as a critical area of research due to its widespread impact on individuals and societies (Matamoros-Fernández and Farkas, 2021). The proliferation of digital media platforms has significantly accelerated the spread of hate speech (Jahan and Oussalah, 2023). We reviewed the existing studies on hate speech detection, exploring various approaches, challenges, and mitigation strategies to comprehensively understand the current state of the art, focusing on Cross-linguistic transfer learning (CLTL) and Cross-lingual hate speech detection.

CLTL is crucial in combating harmful content across diverse languages, making it essential for effective cross-lingual hate speech detection. The central idea of CLTL is to facilitate the transfer of linguistic knowledge across different languages, leveraging models trained on high-resource languages (source languages) and adapting them for use in low-resource languages (target languages). Several multilingual pre-trained models, such as multilingual BERT (mBERT) (Pires et al., 2019) and XLM-Roberta (XLM-R) (Conneau et al., 2019), have shown promising results in CLTL, facilitating the detection of cross-lingual hate speech across diverse languages and cultural contexts. Despite its successes, CLTL faces significant challenges. One major issue is the variability in linguistic structures across languages, which can hinder the transferability of model knowledge (He et al., 2020). This variability requires the design of adaptation techniques to maintain model effectiveness across different languages. Another challenge is the scarcity of annotated datasets in low-resource languages, which limits the training and evaluation of CLTL models (Abad et al., 2020). Innovative methods must be explored to compensate for the lack of this scarcity of annotated data (Shi et al., 2022).

Several studies have been conducted on hate speech detection in under-resourced languages over the past decade. Ranasinghe and Zampieri (2020) explores the use of cross-lingual contextual embeddings to tackle the challenge of offensive language identification in low-resource languages by leveraging transfer learning, focusing on three languages: Bengali, Hindi, and Spanish. Another study introduces a meta-learning approach for few-shot hate speech and offensive language detection across

ten low-resource languages. The study demonstrates that these models can quickly generalize and adapt to new languages with minimal labeled data by leveraging optimization-based and metric-based meta-learning models (Mozafari et al., 2022). Pamungkas et al. (2021) explore hate speech detection in low-resource languages by leveraging cross-lingual transfer learning from English within a zero-shot learning framework. The study proposes joint-learning models that utilize multilingual language representations, such as mBERT, and incorporate an external multilingual lexicon, HurtLex, to enhance model performance. The research findings indicate that joint-learning models outperform more straightforward machine translation approaches. Another study focuses on improving hate speech detection in low-resource languages, specifically Turkish and Arabic, by introducing a novel pooling strategy within a soft-voting ensemble framework. The study demonstrates the effectiveness of this method in enhancing hate speech classification for languages typically underrepresented in computational linguistics (Qachfar et al., 2024). In the study by Beddiar et al. (2021), a deep learning approach is employed to improve hate speech detection using data augmentation techniques, focusing on back translation and paraphrasing. The authors demonstrate significant results across multiple datasets by expanding datasets and applying CNN and LSTM models.

Our approach, CLTL with data augmentation, provides several advantages in hate speech detection for low-resource languages by generating diverse training data that enhances model generalization and reduces overfitting. It is especially valuable where labeled data is scarce, boosting model performance by increasing the training dataset. Its flexibility allows application across different languages, making it a versatile approach that complements techniques like transfer learning and joint-learning models, ultimately improving robustness and adaptability in detecting hate speech.

3 Datasets and Enhancement

Most studies on hate speech detection focus on high-resource languages, particularly English. This results in datasets and models primarily tailored for English-language content (Chen et al., 2018; Röttger et al., 2022; Ivison et al., 2022). Consequently, the development of effective models for under-resourced languages is hindered, affecting downstream tasks such as hate speech detection, sentiment analysis, and fake news detection (Röttger et al., 2022). Billions of non-English speakers are less protected against online hate, and major social media platforms exhibit clear language gaps in their content moderation systems.

Low-resource languages have received minimal investment and interest from the research community. This lack of attention results in a scarcity of datasets and resources tailored for specific computational tasks (Mussandi and Wichert, 2024; Adebara and Abdul-Mageed, 2022). Recognizing this significant gap, we have taken the initiative to compile and make available various datasets to support research in this area.

In our efforts, we collected labeled corpora from a range of African, and Indian languages specifically curated for hate speech detection. This collection encompasses data from multiple sources and covers diverse linguistic and cultural contexts within these regions. By providing these datasets, we aim to bridge the resource gap and stimulate further research and development in NLP for underrepresented languages.

Table 1 presents the statistics for the datasets collected for hate speech detection in various low-resource languages, specifically African and Indian languages. These statistics include the number of entries, the proportion of hate speech vs. non-hate speech entries, and the sources from which the data were collected.

As observed from Table 1, there is a substantial imbalance between the two defined labels: hate speech and non-hate speech. The dataset contains a significantly larger number of non-hate speech entries compared to hate speech entries.

Input-level mixed-up and hidden-level mixed-up are two broad categories of mixed-up-based augmentation techniques. Input-level mixed-up techniques, such as CutMix (Yun et al., 2019; Kim et al., 2020; Walawalkar et al., 2020), focus on augmenting the data at the input level by combining different data samples and their corresponding labels. On the other hand, hidden-level mixed-up, such as Manifold Mixed-up (Verma et al., 2019), operates on the feature representations within the model rather than the raw input data.

By employing Mixed-up, we aim to enhance the dataset with synthetic examples that preserve the semantic labels of the original instances. This method not only increases the volume of the train-

Language	ISO 639 [1].	Hate Speech	Non-Hate Speech	Data source
Algeria ∗	ar-DZ	567	2934	(Lanasri et al., 2023)
Amharic	am	9447	5554	(Ayele et al., 2023)
Assamese	as	2347	1689	(Ghosh et al., 2023)
Bodo	brx	997	681	(Ghosh et al., 2023)
Chad ∗	ar-TD	1996	5007	(Sanoussi et al., 2022)
Egypt ∗	ar-EG	6034	2135	(Ahmed et al., 2022)
Hausa	ha-NG	2459	4008	(Tonneau et al., 2024)
Igbo	ig	4299	1153	(Tonneau et al., 2024)
Kikuyu	ki	46933	3242	(Ombui et al., 2021)
Oromo	om-ET	6636	6176	(Ababu and Woldeyohannis, 2022)
Tunisia ∗	ar-TN	2204	3820	(Haddad et al., 2019)

Table 1: Statistics of the datasets, with * indicating a language that belongs to the Arabic (ar) dialects

ing data but also promotes better generalization and improved performance of the model, especially in scenarios where the dataset is limited or imbalanced.

Mixed-up is an input-level data augmentation technique designed to enhance the diversity of training data and improve model performance. The technique is based on the observation that the label of a hateful sample remains unchanged when concatenated with another sample, whether it is hateful or non-hateful. Similarly, concatenating a non-hateful sample with another non-hateful sample results in a new non-hateful sample. Formally, let's denote (s_i, y_i) as a sentence and its corresponding label $y \in \{\text{hate}, \text{non-hate}\}$ in a minibatch S. The entire dataset is represented by D, and the minibatch S is defined as:

$$S = \{(s_0, y_0), (s_1, y_1), \ldots, (s_n, y_n) \mid (s_i, y_i) \in D\}$$

For each sample $s_i \in S$, we randomly select another sample $(s_j, y_j) \in D$ such that $s_j \neq s_i$. We then apply an augmentation probability p_{aug} to create a new augmented sample:

$$s_i^{\text{aug}} = \phi(s_i, s_j)$$

$$y_i^{\text{aug}} = y_i \vee y_j$$

The function ϕ is defined as:

$$\phi(s_i, s_j) = \begin{cases} \text{concat}(s_i, s_j) & \text{if } p > p_{\text{flip}} \\ \text{concat}(s_i, s_j) & \text{otherwise} \end{cases}$$

In this context, concat() refers to the concatenation of two sentences, and p_{flip} is the probability of flipping the sentence order. Flipping the sentences introduces further augmentation and prevents the model from learning positional biases. Finally, the minibatch S is updated by replacing the original samples with the augmented samples $(s_i^{\text{aug}}, y_i^{\text{aug}})$. This augmented minibatch provides the model with a richer and more varied training set, which helps improve its ability to generalize from the training data to unseen entries. By leveraging Mixed up, we aim to overcome the limitations of data scarcity and imbalance, ultimately leading to more robust and effective models for hate speech detection.

3.1 Fine-tuning Source and Target Language

Fine-tuning with a source language leverages the knowledge and patterns learned from a large source language dataset. This pre-training stage allows the model to capture general linguistic features and semantic representations that can be beneficial for the target language (Howard and Ruder, 2018). By pre-training on a high-resource language, the model can develop a robust understanding of syntax, semantics, and other linguistic structures, which are then transferred to the target language during the fine-tuning process (Ruder, 2019).

Fine-tuning on the target language can effectively address the issue of data scarcity, which is a common challenge in NLP for low-resource languages. By leveraging transfer learning from high-resource languages and subsequently fine-tuning on the available data in the target language, models can achieve significant performance gains. This approach is particularly beneficial for languages with limited annotated data, as it maximizes the utility of existing resources and enhances the model's generalization capabilities (Conneau et al., 2019).

4 Models

For fine-tuning with the source language, we chose different LLMs, such as BERT (Devlin et al., 2018) and RoBERTa-base (Liu et al., 2019) for monolingual models and XLM-R (Conneau et al., 2019) and XLM-T (Barbieri et al., 2021) for multilingual models.

This choice implies that the monolingual models were specifically trained on English datasets, providing a strong baseline for English-language tasks. Meanwhile, the multilingual models, although designed to handle multiple languages, also included a significant amount of English data in their training corpora, making them robust for English as well. Table 2 indicate the experimental results of each model on the English dataset.

Models	Acc	F1	ΔF1
BERT-base	69.80	68.33	-
RoBERTa-base	77.82	75.69	+7.40
XLM-R	78.32	75.74	+0.05
XLM-T	88.43	84.58	**+8.84**

Table 2: Finetune SL: Accuracy and F1 score on English dataset. $\Delta F1$ represents the change or improvement in the F1-score compared to a previous baseline model.

5 Methodology

We used the English hate speech dataset (Vidgen et al., 2020) as the source language. For the initial fine-tuning phase, we selected a sample comprising 20,000 entries for training, 500 entries for development, and 2,000 entries for testing. To maintain a consistent proportion of hate content across the English datasets, we employed random sampling techniques.

During the experimental phase, we compared the performance of various models, as discussed in Section 4, to determine the most effective one for our specific use case. After comprehensive testing, we found that XLM-T outperformed the other models in terms of accuracy and generalization capabilities. Consequently, we selected XLM-T for further development, focusing on fine-tuning the target language data.

The superior performance of XLM-T was evident across all these metrics, demonstrating its robustness and reliability for handling diverse and complex hate speech data. Additionally, XLM-T's ability to generalize well across different datasets highlighted its potential for broader applications beyond our initial scope.

As a result, we decided to proceed with XLM-T, focusing on fine-tuning it with target language data to further enhance its performance and applicability. This decision aligns with our goal of developing a highly accurate and versatile model capable of effectively identifying and mitigating hate speech in various linguistic contexts.

We proceeded to train the model with a substantial dataset consisting of labeled hate speech entries. This dataset was carefully curated to ensure a comprehensive representation of various hate speech scenarios, capturing a wide range of expressions and contexts in which hate speech might occur.

To thoroughly assess the effectiveness of our model, we trained and evaluated the dataset both before and after applying data augmentation techniques. Data augmentation techniques, such as synonym replacement, back-translation, and paraphrasing, were employed to increase the diversity of the training data. This approach helped the model learn from a broader spectrum of hate speech expressions, enhancing its ability to generalize to new, unseen data.

We divided our target language dataset into three distinct subsets: training, validation, and test sets. The training set was used to train the model, providing it with a large amount of data to learn from. The validation set was used during the training process to fine-tune the model's parameters and prevent overfitting. Finally, the test set was used to evaluate the model's performance, providing an unbiased assessment of its ability to generalize to new data.

Throughout the training process, we continuously monitored key performance metrics, such as the F1-score. These metrics provided a comprehensive assessment of the model's ability to correctly identify and classify hate speech. By comparing the performance of the model before and after data augmentation, we were able to determine the impact of these techniques on the model's accuracy and generalization capabilities. Figure 1 depicts the architectural design of the proposed solution.

5.1 Data-Augmentation

For our experiment, we use an equal-sized dataset for all target languages, ensuring consistency across our analyses. Each dataset comprises 2,000 labeled data points, providing a robust foundation

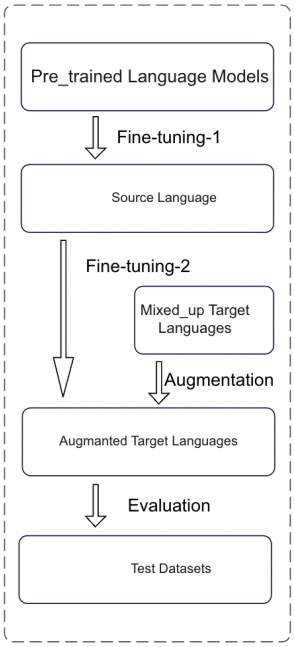

Figure 1: The architectural design of the proposed solution

for model training. To enhance the dataset and potentially improve model performance, we employ a mixed-up augmentation technique. This technique effectively increases the dataset size to 4,001 data points by combining data points in a way that introduces additional variability and richness to the training data.

Once the augmented datasets are prepared, we proceed to train our models. This involves fine-tuning pre-existing models using the original 2,000 data points as well as the expanded dataset of 4,001 data points. Our objective is to evaluate the impact of data augmentation on model performance.

We conduct a comprehensive evaluation of the fine-tuned models at two key stages: Before Data Augmentation: Here, we assess the model's performance using the original dataset of 2,000 labeled data points. This provides a baseline measurement of how well the model performs without any augmentation.

After Data Augmentation: In this stage, we evaluate the model's performance after it has been trained on the augmented dataset of 4,001 data points. This allows us to determine the effectiveness of the mixed-up augmentation technique in enhancing the model's predictive capabilities. By comparing the performance metrics from these two stages, we can draw conclusions about the benefits and potential trade-offs associated with data augmentation in our specific context. This thorough approach ensures that we understand not only the immediate impact of augmentation but also its implications for model generalization and accuracy across different target languages. Figure 2 illustrates how the mixed-up data augmentation technique works.

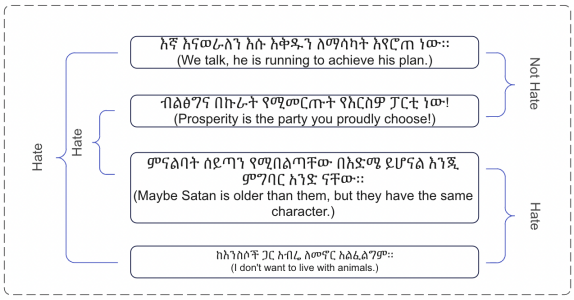

Figure 2: Mixed-up data augmentation sample combining Amharic and English versions.

6 Experimental Results

We use macro F1 to evaluate model performance because most of our test sets are imbalanced. The macro F1 score is particularly suited for this purpose, as it calculates the F1 score for each class independently and then takes the average, treating all classes equally. This approach ensures that the performance of the model is measured in a balanced manner, giving equal importance to both minority and majority classes.

We conducted three distinct experiments to explore the impact of various data augmentation techniques and learning strategies on model performance. Each experiment was designed to address a specific aspect of model generalization and robustness in multilingual settings.

The first experiment focused on evaluating the effectiveness of mixed-up data augmentation techniques. Data augmentation involves creating new training examples by applying various transformations to the original data, which can help the model generalize better to unseen data. In this case, we applied mixed-up data augmentation across all target languages to enhance the diversity and variability of the training dataset.

The results of this experiment, as shown in Table 3, demonstrated that the model trained on augmented data from all target languages significantly outperformed the model trained on non-augmented data. This indicates that the mixed-up augmentation technique was successful in helping the model

Sources	Before augmentation		After augmentation		
	Acc	F1	Acc	F1	**ΔF1**
Algeria	0.73	0.69	0.76	0.72	+ 0.03
Amharic	0.61	0.57	0.63	0.59	+ 0.02
Assamese	0.69	0.66	0.73	0.69	+ 0.03
Bodo	0.63	0.55	0.69	0.59	+ 0.04
Chad	0.77	0.69	0.89	0.70	+ 0.01
Egypt	0.75	0.69	0.79	0.73	+ 0.04
Hausa	0.79	0.72	0.82	0.76	+ 0.04
Igbo	0.53	0.46	0.60	0.50	+ 0.04
Kikuyu	0.87	0.69	0.89	0.74	+ 0.05
Oromo	0.72	0.67	0.80	0.69	+ 0.02
Tunisia	0.71	0.63	0.79	0.66	+ 0.03
Multilingual	0.63	0.55	0.67	0.56	+ 0.01

Table 3: Mixed-up: The Accuracy and Macro F1 scores on the respective test datasets for XLM-T were first fine-tuned on the source language, then evaluated on both augmented and non-augmented target languages. $\Delta F1$ represents the improvement in the F1-score when comparing the results before and after data augmentation.

to better generalize across different languages. The augmented data likely introduced more variability into the training process, allowing the model to learn more robust representations that are less sensitive to language-specific nuances.

These findings underscore the importance of data augmentation in training models for multilingual tasks, especially when dealing with limited or imbalanced datasets. By incorporating mixed-up data augmentation, we can create more diverse training sets that lead to improved model performance across a wider range of languages and scenarios.

Overall, data augmentation positively impacted the performance of the XLM-T model across all languages, with the largest gains in F1 scores for languages like Kikuyu, Bodo, and Hausa, suggesting these languages benefited the most from augmentation techniques. However, languages like Chad and Multilingual saw smaller improvements, implying that the effectiveness of augmentation may vary based on the language characteristics and the quality of the original dataset.

In the second experiment, we explored the capabilities of zero-shot learning by training a model using an augmented source dataset. Zero-shot learning is a technique that allows a model to make predictions on classes it has never seen during training by leveraging knowledge from related tasks or languages. Specifically, our focus was on assessing the model's ability to transfer knowledge across languages in a zero-shot cross-lingual setting.

As illustrated in Table 4, we measured the performance of the model trained on source languages with available training data and tested it on languages for which no training data was provided. This setup challenges the model to apply its learned representations and generalize to completely new languages and classes.

This experiment is particularly significant because it evaluates the model's robustness and flexibility in real-world scenarios where data for certain languages or classes might be scarce or unavailable. The successful performance of the model in these zero-shot tasks would demonstrate its ability to understand and categorize new, unseen classes without the need for specific examples of those classes in its training data. Basically, the model could use the features and patterns it learned during training to correctly sort data from languages and classes it has never seen before. This shows how useful zero-shot learning can be in situations with more than one language or class.

Overall, the results indicate that the XLM-T model is capable of transferring knowledge across linguistically and geographically related languages, as evidenced by strong performances in cases like Assamese → Bodo and Chad → Algeria/Tunisia. However, the performance drops in cases where the languages are more distinct, such as Amharic → Assamese. This suggests that linguistic proximity, shared cultural contexts, or similar alphabets might play a role in the effectiveness of zero-shot learning models in cross-lingual tasks.

In the third experiment, we evaluated the model's performance when trained on a limited dataset using few-shot learning techniques. Few-shot learning aims to enable models to perform well even when only a small number of training samples are available, making it particularly valuable in situations where data is scarce.

For this experiment, we trained the model with varying numbers of data samples—specifically 5, 10, 20, 30, 50, and 100 samples. The objective was to assess how well the model could generalize and perform on tasks with minimal training data. By progressively increasing the number of samples, we observed the model's learning curve and identify the minimum amount of data required to achieve acceptable performance.

This experiment is crucial for understanding the model's ability to learn from limited examples and its potential application in real-world scenarios where collecting large amounts of labeled data may be challenging. The results of this experiment provide insights into the effectiveness of few-shot

	Target Language										
Source Language	Algeria	Amharic	Assamese	Bodo	Chad	Egypt	Hausa	Igbo	Kikuyu	Oromo	Tunisia
Algeria	-	0.29	0.31	0.29	0.47	0.49	0.18	0.23	0.31	0.24	**0.54**
Amharic	0.31	-	0.15	0.17	0.28	**0.29**	0.18	0.19	0.24	0.24	0.27
Assamese	0.21	0.23	-	**0.56**	0.17	0.19	0.21	0.22	0.22	0.24	0.23
Bodo	0.25	0.23	**0.35**	-	0.21	0.16	0.29	0.24	0.22	0.24	0.25
Chad	**0.48**	0.26	0.29	0.24	-	0.41	0.28	0.26	0.28	0.29	0.37
Egypt	0.41	0.33	0.20	0.22	0.39	-	0.27	0.24	0.29	0.30	**0.45**
Hausa	0.24	0.24	0.25	0.26	0.27	0.26	-	**0.34**	0.29	0.24	0.19
Igbo	0.23	0.21	0.30	0.25	0.18	0.26	**0.39**	-	0.28	0.29	0.22
Kikuyu	0.29	0.33	0.30	0.31	0.23	0.29	0.33	0.24	-	**0.35**	0.29
Oromo	0.31	0.32	0.25	0.26	0.27	0.29	0.27	0.24	**0.34**	-	0.26
Tunisia	**0.39**	0.33	0.30	0.28	0.41	0.29	0.33	0.31	0.29	0.26	-

Table 4: Zero-shot: The Macro F1 score on the respective target languages was measured for XLM-T, which was initially fine-tuned on the source language. The best performance for each target language is highlighted in bold.

Sources	5	10	20	30	40	50	100
Algeria	0.27	0.34	0.36	0.35	0.37	0.39	0.42
Amharic	0.25	0.29	0.26	0.29	0.31	0.32	0.32
Assamese	0.22	0.31	0.36	0.36	0.35	0.36	0.37
Bodo	0.19	0.22	0.25	0.26	0.27	0.28	0.33
Chad	0.18	0.16	0.19	0.23	0.24	0.35	0.38
Egypt	0.15	0.24	0.32	0.34	0.34	0.36	0.39
Hausa	0.24	0.24	0.25	0.27	0.28	0.32	0.36
Igbo	0.31	0.31	0.32	0.34	0.37	0.37	0.41
Kikuyu	0.24	0.27	0.28	0.29	0.31	0.36	0.36
Oromo	0.34	0.33	0.35	0.35	0.38	0.36	0.38
Tunisia	0.31	0.33	0.34	0.36	0.40	0.39	0.40
Multilingual	0.31	0.33	0.33	0.31	0.38	0.41	0.41

Table 5: Few-shot: The Macro F1 score on the respective test datasets for XLM-T was fine-tuned first on the source language, then fine-tuned again on N target-language entries, with results averaged across 5 random seeds for each N.

learning techniques and the model's capability to adapt to tasks with constrained data availability.

Overall, the few-shot learning results indicate that the XLM-T model benefits from having more target-language examples, with performance steadily increasing as N increases. Languages like Algeria, Oromo, and Tunisia perform well even with fewer examples, while Chad, Bodo, and Egypt show more significant gains as the number of examples increases. This highlights that some languages may require more data for the model to generalize effectively.

7 Limitations

The study mainly focused on detecting hate speech in low-resource languages, and the methods were specifically designed for this task. As a result, these methods may not be as effective for other types of NLP tasks.

8 Conclusion

This research addresses the challenge of hate speech detection in low-resourced languages, which refers to the task of identifying harmful, offensive, or discriminatory language in languages that lack substantial annotated datasets or computational resources. Hate speech detection in these languages faces significant limitations, primarily due to the scarcity of labeled data, the linguistic diversity, and the lack of sophisticated models tailored for these languages. Traditional hate speech detection models, predominantly trained on high-resource languages like English, struggle to generalize effectively to languages with fewer resources, leading to lower accuracy and reduced reliability in identifying harmful content.

To overcome these limitations, we employed data augmentation techniques, particularly mixed-up data augmentation, to artificially expand and diversify the training datasets. This approach allowed us to enhance the model's ability to generalize by generating additional synthetic training examples, thereby improving performance in low-resource settings.

During the experiments, we trained models using various techniques, including cross-lingual transfer learning, few-shot learning, and zero-shot learning. These techniques enabled the models to transfer knowledge from high-resource languages to low-resource ones, allowing for significant improvements in hate speech detection even with limited data. The results demonstrated that data-efficient approaches, when combined with data augmentation, significantly boost the performance of hate speech detection models in under-resourced languages, providing a more scalable and inclusive solution for addressing online hate speech across diverse linguistic contexts.

References

Teshome Mulugeta Ababu and Michael Melese Woldeyohannis. 2022. Afaan oromo hate speech detection and classification on social media. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 6612–6619.

Alberto Abad, Peter Bell, Andrea Carmantini, and Steve Renais. 2020. Cross lingual transfer learning for zero-resource domain adaptation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6909–6913. IEEE.

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric nlp for african languages: Where we are and where we can go. *arXiv preprint arXiv:2203.08351*.

Ibrahim Ahmed, Mostafa Abbas, Rany Hatem, Andrew Ihab, and Mohamed Waleed Fahkr. 2022. Fine-tuning arabic pre-trained transformer models for egyptian-arabic dialect offensive language and hate speech detection and classification. In *2022 20th International Conference on Language Engineering (ESOLEC)*, volume 20, pages 170–174. IEEE.

Abinew Ali Ayele, Seid Muhie Yimam, Tadesse Destaw Belay, Tesfa Asfaw, and Chris Biemann. 2023. Exploring amharic hate speech data collection and classification approaches. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 49–59.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2021. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. *arXiv preprint arXiv:2104.12250*.

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*, 24:100153.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2018. Multi-source cross-lingual model transfer: Learning what to share. *arXiv preprint arXiv:1810.03552*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12:2493–2537.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90.

Koyel Ghosh, Apurbalal Senapati, and Aditya Shankar Pal. 2023. Annihilate hates (task 4 hasoc 2023): Hate speech detection in assamese bengali and bodo languages. In *FIRE (Working Notes)*, pages 368–382.

Hatem Haddad, Hala Mulki, and Asma Oueslati. 2019. T-hsab: A tunisian hate speech and abusive dataset. In *International conference on Arabic language processing*, pages 251–263. Springer.

Keqing He, Weiran Xu, and Yuanmeng Yan. 2020. Multi-level cross-lingual transfer learning with language shared and specific knowledge for spoken language understanding. *IEEE Access*, 8:29407–29416.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Hamish Ivison, Noah A Smith, Hannaneh Hajishirzi, and Pradeep Dasigi. 2022. Data-efficient finetuning using cross-task nearest neighbors. *arXiv preprint arXiv:2212.00196*.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. 2020. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *International Conference on Machine Learning*, pages 5275–5285. PMLR.

Young-Bum Kim, Karl Stratos, Ruhi Sarikaya, and Minwoo Jeong. 2015. New transfer learning techniques for disparate label sets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 473–482.

Rohan Kshirsagar, Tyus Cukuvac, Kathleen McKeown, and Susan McGregor. 2018. Predictive embeddings for hate speech detection on twitter. *arXiv preprint arXiv:1809.10644*.

Dihia Lanasri, Juan Olano, Sifal Klioui, Sin Liang Lee, and Lamia Sekkai. 2023. Hate speech detection in algerian dialect using deep learning. *arXiv preprint arXiv:2309.11611*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ariadna Matamoros-Fernández and Johan Farkas. 2021. Racism, hate speech, and social media: A systematic review and critique. *Television & new media*, 22(2):205–224.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2022. Cross-lingual few-shot hate speech and offensive language detection using meta learning. *IEEE Access*, 10:14880–14896.

Joaquim Mussandi and Andreas Wichert. 2024. Nlp tools for african languages: Overview. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 2*, pages 73–82.

Edward Ombui, Lawrence Muchemi, and Peter Wagacha. 2021. Building and annotating a codeswitched hate speech corpora. *Int. J. Inf. Technol. Comput. Sci*, 3:33–52.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2021. A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. *Information Processing & Management*, 58(4):102544.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Fatima Zahra Qachfar, Bryan Tuck, and Rakesh Verma. 2024. Detectiveredasers at hsd-2lang 2024: A new pooling strategy with cross-lingual augmentation and ensembling for hate speech detection in low-resource languages. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 199–204.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. *arXiv preprint arXiv:2010.05324*.

Francisco Rangel, Gretel Liz de la Peña-Sarracén, María Alberta Chulvi-Ferriols, Elisabetta Fersini, and Paolo Rosso. 2021. Profiling hate speech spreaders on twitter task at pan 2021. In *Proceedings of the Working Notes of CLEF 2021, Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st to 24th, 2021*, pages 1772–1789. CEUR.

Paul Röttger, Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Data-efficient strategies for expanding hate speech detection into under-resourced languages. *arXiv preprint arXiv:2210.11359*.

Sebastian Ruder. 2019. *Neural transfer learning for natural language processing*. Ph.D. thesis, NUI Galway.

Mahamat Saleh Adoum Sanoussi, Chen Xiaohua, George K Agordzo, Mahamed Lamine Guindo, Abdullah MMA Al Omari, and Boukhari Mahamat Issa. 2022. Detection of hate speech texts using machine learning algorithm. In *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0266–0273. IEEE.

Anna Schmidt and Michael Wiegand. 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.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2018. Cross-lingual transfer learning for multilingual task oriented dialog. *arXiv preprint arXiv:1810.13327*.

Xiayang Shi, Xinyi Liu, Chun Xu, Yuanyuan Huang, Fang Chen, and Shaolin Zhu. 2022. Cross-lingual offensive speech identification with transfer learning for low-resource languages. *Computers and Electrical Engineering*, 101:108005.

Manuel Tonneau, Pedro Vitor Quinta de Castro, Karim Lasri, Ibrahim Farouq, Lakshminarayanan Subramanian, Victor Orozco-Olvera, and Samuel Fraiberger. 2024. Naijahate: Evaluating hate speech detection on nigerian twitter using representative data. *arXiv preprint arXiv:2403.19260*.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. 2019. Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, pages 6438–6447. PMLR.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2020. Learning from the worst: Dynamically generated datasets to improve online hate detection. *arXiv preprint arXiv:2012.15761*.

Devesh Walawalkar, Zhiqiang Shen, Zechun Liu, and Marios Savvides. 2020. Attentive cutmix: An enhanced data augmentation approach for deep learning based image classification. *arXiv preprint arXiv:2003.13048*.

Mesay Gemeda Yigezu, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023. Transformer-based hate speech detection for multi-class and multi-label classification.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 745–760. Springer.