CSCI 5541 NLP: Final Project Report

Team Members: Amoligha Timma, Gehad Abdelrahman, Malak Raafat, Abhi Bijlwan
Team Name: Audi Quattro
Advisors: Risako Owan, Junhan Wu

Abstract

Advancements in generative AI and large-scale language models have largely benefited widely spoken languages, leaving minority languages like Sourashtra underserved. We curated a custom Sourashtra dataset and evaluated the impact of cross-lingual transfer from Indo-Aryan and Dravidian languages. Among the models tested, the Augmented Indo-Aryan ensemble model outperformed others, demonstrating improved context and negation handling, while the Dravidian ensemble struggled due to limited lexical overlap. The augmented Indo-Aryan ensemble model achieved 67.08% accuracy, highlighting the effectiveness of targeted fine-tuning and cross-linguistic transfer. This work contributes to the inclusion of underrepresented languages in digital technologies, offering a framework for improving NLP tools for similar low-resource languages.

1 Introduction, Background, and Motivation

In recent years, there has been a boom in generative AI and large scale LLM models. With this, many popular languages, like English and Spanish, are able to get models that people can interact with - helping them bridge an online gap. Not only that, but having support in these languages allows users to further explore that language and make use of LLM features that were previously inaccessible.
Since the technology is fairly recent, there isn't much support for minority languages–such as the Sourashtra language, indigenous to India. Today, speakers of the Sourashtra face difficulties when using digital tools like automatic translation, social media platforms, or online learning systems because these tools are not built to recognize or work with their language. This is primarily due to the fact that there is very little data available for Sourashtra in the digital world–especially compared to larger, more common languages like English or Hindi. This leaves these speakers at a disadvantage when it comes to accessing online services, communicating on global platforms, and/or participating fully in the digital world.
Our aim is to ultimately enhance the way computers handle Sourashtra and potentially understand the language. Specifically, we aim to test whether models trained on closely related Indian languages such as Hindi, Gujarati, and Marathi (which have more resources available) can better classify sentiment/meaning in Sourashtra. We also create a dedicated, custom-made dataset in Sourashtra to help rigorously evaluate and fine-tune our model. Our broader goal is to contribute to the inclusion of underrepresented languages in the digital sphere, so that minority language speakers can benefit from the same technologies that support other major languages. This work not only addresses a technical challenge but also advocates linguistic diversity and equal access to technology.

1.1 Previous Work

Current approaches to sentiment analysis in low-resource languages commonly rely on large-scale multilingual transformer models such as XLM-RoBERTa (Conneau et al., 2020). In the FIRE2020 shared task on sentiment analysis of code-mixed Dravidian languages, researchers fine-tuned XLM-RoBERTa-base on Tamil-English and Malayalam-English datasets (Ou and Li, 2020). Their method included preprocessing techniques like token normalization, followed by 5-fold cross-validation and average ensemble, which contributed to achieving the top-ranking results in the competition.
Similarly, (Prytula, 2024) evaluated multilingual models including mBERT, DistilBERT and XLM-RoBERTa on the Ukrainian language through sentiment classification, discovering that XLM-RoBERTa-base achieved the highest accuracy (91.32%), outperforming both monolingual (Ukr-RoBERTa) and multilingual baselines. Despite

having minimal pretraining in Ukrainian, XLM-RoBERTa demonstrated strong generalization across review datasets from various domains. (Nuci et al., 2024) further showed the effectiveness of XLM-RoBERTa with mixed Albanian-English data, indicating that the model performed well even with minimal Albanian representation in pretraining.

Despite the growing adoption of multilingual models like XLM-RoBERTa, existing research still faces significant limitations. Many fine-tuning pipelines fail to consider language relationships when choosing training languages. (Dhamecha et al., 2021) showed that incorporating unrelated languages in fine-tuning can reduce performance, indicating choosing languages based on linguistic similarity is critical for effective transfer. However, employed models like in FIRE2020 often train on a mix of languages without considering those relationships. Additionally, although languages like Ukrainian and Albanian received some representation during the initial pretraining of XLM-RoBERTa, truly underrepresented languages like Sourashtra are entirely absent from pretraining corpora, which complicates generalization. Moreover, there are currently no publicly available labeled datasets for Sourashtra, which presents a major challenge for model development.

Another limitation is that current research often reports only overall performance metrics like accuracy or F1 scores, with limited analysis of which types of sentences or structures are handled effectively or poorly. This leaves open questions about the performance of multilingual models in various contexts, especially in low-resource or cross-lingual scenarios with languages like Sourashtra.

Our work would be of particular interest to researchers, developers, and language advocates who are concerned with the under-representation of minority and low-resource languages in natural language processing. Individuals and communities who recognize the digital marginalization of languages like Sourashtra, and who seek to bridge this linguistic gap, would find our project highly relevant.

By proposing a scalable fine-tuning methodology, we provide a practical blueprint for addressing the common challenge of data scarcity. Researchers aiming to support other endangered or indigenous languages can adapt our approach by identifying historically and linguistically related languages to supplement model training, even when direct resources for the target language are minimal or nonexistent. Our findings demonstrate that even the addition of a small number of carefully selected target-language samples, such as 10 in our case, can meaningfully boost model performance. This highlights the value of minimal but strategic data augmentation.

Specifically, for those interested in advancing the use of Sourashtra in NLP applications, our project offers immediate tools. We have curated and annotated a new Sourashtra sentiment analysis dataset containing 250 samples, providing a valuable resource for future experimentation. This removes the need for researchers to start from scratch by collecting their own data, which can be both time-consuming and difficult, especially for non-native speakers. Furthermore, because translation APIs like Google Translate do not reliably support Sourashtra, our dataset fills a critical gap in accessible and authentic linguistic data.

2 Approach

Our project aimed to address the challenges of building effective sentiment analysis models for Sourashtra, a minority Indian language with extremely limited digital resources. We recognized that off-the-shelf multilingual models like XLM-RoBERTa, while strong (Carreras Timoneda and Vallejo Vera, 2024), often struggle with low-resource languages without targeted adaptation. Our hypothesis was that since Sourashtra is historically and linguistically similar to Indo-Aryan languages such as Gujarati, Marathi, and Hindi, fine-tuning XLM-RoBERTa on large-scale datasets from these related languages would help the model better capture the semantics and structures relevant to Sourashtra. We anticipated that this targeted cross-linguistic transfer would outperform general multilingual pre-training alone, as shown in studies such as the Catalan assessment, where language-specific fine-tuning significantly outperformed general multilingual models (Armengol-Estapé et al., 2021). Furthermore, to validate the importance of linguistic similarity, we fine-tuned the model in Dravidian languages, which are historically unrelated to Sourashtra, to observe whether improvements were specific to Indo-Aryan connections, based on insights from multilingual fine-tuning studies, which found that carefully selecting related languages maximizes performance, while unrelated languages can even harm it (Dhamecha

et al., 2021).

To execute this plan, we first curated and annotated a new Sourashtra sentiment analysis dataset, consisting of 250 balanced samples (125 positive and 125 negative). The curated dataset served two key purposes: establishing a zero-shot baseline for XLM-RoBERTa without any fine-tuning, and evaluating the effectiveness of our fine-tuned models. We then fine-tuned XLM-RoBERTa separately on datasets from Indo-Aryan languages and Dravidian languages and compared the results against the baseline model. Initially, fine-tuning on Indo-Aryan languages led to modest improvements of about 5% over the baseline, while fine-tuning on Dravidian languages yielded little to no gains. To further enhance performance, we augmented the Indo-Aryan datasets by adding just 10 labeled Sourashtra samples (5 positive and 5 negative). This strategic augmentation led to significant improvements, with accuracies rising to the 65.5%-69.1% range compared to the baseline accuracy of approximately 50%. Finally, we implemented an ensemble method to combine the strengths of individual models fine-tuned on different languages, resulting in a final accuracy of 67.08Our approach builds on and extends prior research, including findings from low-resource languages like Albanian, where fine-tuning led to significant improvements over using general multilingual models (Nuci et al., 2024), and highlights the importance of targeted fine-tuning for low-resource languages, the role of language relatedness in transfer learning, and the effectiveness of multilingual models like XLM-RoBERTa when adapted thoughtfully. By combining historical linguistic knowledge with modern transfer learning techniques, we demonstrated a practical pathway for improving NLP tools for underrepresented languages like Sourashtra. This strategy not only offers a solution for Sourashtra but also provides a replicable framework for improving sentiment analysis and language understanding for other minority and endangered languages facing similar resource limitations.

2.1 Challenges

From the beginning of this project, we acknowledged that working with Sourashtra, a highly underrepresented minority language, would present unique challenges. One of the primary difficulties we anticipated was the uneven availability of comparable datasets across the six related Indian languages that we were exploring for Sourashtra. Although these languages have far more digital resources compared to our language of study, finding sentiment-labeled datasets of comparable size and quality proved to be a far more difficult than we expected. Some datasets we explored had small training and test sets, while others were sentiment based with additional emotions attached. Concurrently, developing our own dataset in Sourashtra, though limited to 250 sentences, required a considerable manual effort. Additionally, having a single annotator made it challenging to create a diverse range of topics and sentence structures, limiting the variety within the dataset. Furthermore, it was difficult to generate clearly positive or negative sentences while maintaining linguistic purity without mixing English words. This task underscored the broader challenge of building meaningful resources for minority languages, where even foundational datasets have to be created from the ground up.

When we first applied a baseline model to the curated Sourashtra data (using base XLM-RoBERTa) and then fine-tuned it on each of the six related languages, the results were more underwhelming than we had anticipated. Most fine-tuned models only yielded 1-2% increase in predicting Sourashtra sentiment, while one fine-tuned language model performed worse than the standard 50% observed in the baseline BERT model. We initially hypothesized that fine-tuning on historically linguistic similarity languages would lead to clear improvements, but the gains were limited. The early set-back was disappointing but vital as it encouraged us to reconsider initial assumptions and explore different approaches.

Subsequent efforts, including the use of an ensemble approach that combined different languages into a single model, also failed in delivering the improvements we anticipated. These challenges revealed the limits of simply increasing the volume or diversity of training data and further highlighted the complex nature of cross-language learning. Ultimately, these difficulties pushed us to think more creatively. We began exploring more targeted data augmentation strategies, such as adding a small number (around 10) of Sourashtra samples into the training sets. While seemingly a minor change, this adjustment had a measurable impact on model performance.

2.2 Novelty

The scientific novelty of our project lies in our approach to tackle the challenges we encountered lies

in how we adapted and extended existing multilingual learning techniques to address the challenges of working with a minority language. While prior research has explored cross-lingual transfer and fine-tuning multilingual models, much of that work has focused on languages with at least a moderate amount of training data.

Initially, we applied fine-tuning on individual languages and ensemble combinations, with the goal that linguistic similarity would provide the greatest results in accuracy and learning. When these early methods produced only minor gains, we introduced a more novel approach: augmenting the training process by adding a very small number of Sourashtra examples into the fine-tuning phase. This hybrid strategy allowed us to inject crucial language-specific signals into the training process, significantly boosting performance despite the tiny size of our Sourashtra dataset.

Our work demonstrated that even a small amount of native language data can play a disproportionately large role in improving model outcomes, especially when combined with well-resourced linguistically related languages. While other work focused primarily on fine-tuning models with the highest accuracy, our approach also included performing qualitative analysis on mislabeled sentences to understand key differences between various models. We also propose that in an augmentation approach, the Sourahstra samples selected impact overall model performance–with samples that cover diverse vocabulary, sentence structure, and negation often resulting in far better overall model accuracies.

3 Evaluation Metrics and Experiments

Success was measured by the model's accuracy in classifying Sourashtra sentences as positive or negative. The key indicator of success was improvement in sentiment classification accuracy after different stages of multilingual fine-tuning. Our goal was not only to achieve higher accuracy on this task, but also to validate whether certain language families improve transfer learning more effectively than others for a low-resource language like Sourashtra. The research questions we wanted to explore were the following:

1. Can existing multilingual models like XLM-RoBERTa be manipulated effectively for use with underrepresented languages with Sourashtra?
2. Does historical migration patterns of the Sourashtra people influence which language groups have the most linguistic influence on the underrepresented language?
3. Can even a small number of Sourashtra samples utilized in finetuning in conjunction with other mainstream languages improve overall model accuracy?

The primary evaluation metric was classification accuracy, computed as the percentage of correct sentiment predictions on the Sourashtra dataset. We conducted several controlled experiments to compare model performance across training conditions as detailed below.

Baseline: XLM-RoBERTa with no additional fine-tuning

Indo-Aryan Ensemble: XLM-RoBERTa + Gujarati, XLM-RoBERTa + Marathi, XLM-RoBERTa + Hindi are finetuned individually and then combined ensemble style to develop an Indo-Aryan Ensemble Model.

Dravidian Ensemble: XLM-RoBERTa + Tamil, XLM-RoBERTa + Telegu, XLM-RoBERTa + Malayalam are finetuned individually and then combined ensemble style to develop a Dravidian Ensemble Model.

Sourashtra Finetuned: XLM-RoBERTa was finetuned (trained and tested) on exclusively the Sourashtra dataset we curated.

Augmented Indo-Aryan Ensemble: XLM-RoBERTa + Gujarati + 10 Sourashtra samples, XLM-RoBERTa + Marathi + 10 Sourashtra samples, XLM-RoBERTa + Hindi + 10 Sourashtra samples, are finetuned individually and then combined ensemble style to develop an Augmented Indo-Aryan Ensemble Model. The same 10 Sourashtra samples were used in each iteration.

4 Results

4.1 Quantitative Results

We began with a baseline XLM-RoBERTa model achieving 50.0% accuracy with no fine-tuning. Fine-tuning on single languages yielded small gains: Gujarati (50.4%), Marathi (56.8%), Hindi (52.4%), Malayalam (49.6%), Tamil (50.0%), and Telugu (56.8%). Combining the Indo-Aryan languages (Gujarati, Marathi, Hindi) into an ensemble produced 53.2%, while the Dravidian ensemble (Tamil, Telugu, Malayalam) reached 50.8%. Fully fine-tuning on our 250-sentence Sourashtra

dataset boosted accuracy to 64.0%. Finally, our augmented experiments—in which we added 10 Sourashtra examples into each Indo-Aryan fine-tuned model achieved 65.4% (Gujarati augmented), 66.3% (Marathi augmented), 69.2% (Hindi augmented), and a 67.1% accuracy for the combined augmented Indo-Aryan ensemble. These are all displayed in Figure 3 in the Appendix. We experimented with different ranges of learning rates (1e-5 to 3e-5) and different ranges of training epochs (3 to 12) and have listed the best performing parameters in Table 3 which can be found in the Appendix. We focused our efforts on determining best parameters for Indo-Aryan related experimentation as that proved to be of higher accuracy from initial experimentation.

In order to compare the performance of our best performing model (Augmented Indo-Aryan Ensemble), we took the test dataset of our fully-finetuned Sourashtra model and compared the distribution of points in a confusion matrix as well as overall accuracy in Figure 1 and Figure 2.

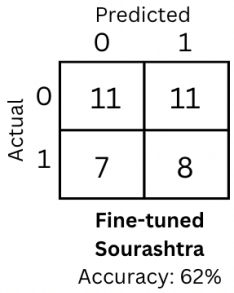

Figure 1: Confusion matrix for Sourashtra Fine-tuned model.

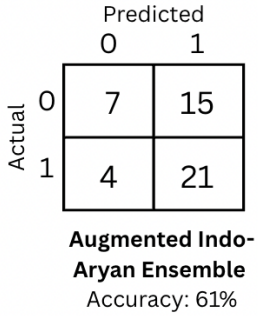

Figure 2: Confusion matrix for Augmented Indo-Aryan Ensemble model.

The fully Sourashtra fine-tuned model achieved an overall accuracy of 62%, with a balanced performance across both positive and negative classes (F1-scores: 0.55 for negative, 0.67 for positive). The Augmented Indo Aryan Ensemble Model model reached 61% accuracy, with better recall for the positive class (0.84) but poor performance on the negative class (F1-score: 0.42), indicating a recall-precision tradeoff.

To further investigate the impact of sample size on data augmentation, we ran multiple experiments using the same setup for the model with highest accuracy, which is the Augmented Hindi model with Sourashtra, varying the number of Sourashtra samples from 2 to 20 with equal numbers of positive and negative labels. Each configuration used identical hyperparameters: a learning rate of 3e-5, batch size of 16, 7 training epochs and model selection based on F1 score, differing only in the number of randomly selected Sourashtra samples. As shown in Table 1, the results revealed considerable inconsistency, where for many runs, accuracy remained fixed at 50%, while in others, particularly those with 10, 16 or 20 samples, the model achieved significantly higher accuracy. Interestingly, when the same configuration was rerun with a different random sample selection (e.g., at 10 or 20 samples), accuracy dropped back to 50%. This suggests that not all samples contribute equally to generalization and that certain samples may carry more informative linguistic features for the model. The results highlight the sensitivity of low-resource fine-tuning to sample selection and the need for more principled data selection strategies.

Number of Sourashtra Samples	Accuracy (%)
2	50
4	50
6	50
8	50
10	63.75
12	50
14	50
16	63.25
18	50
20	65.25

Table 1: Results of Augmented Hindi Model across varying Sourashtra sample sizes

The sampled sentences that provided the highest accuracy throughout our experimentation are indicated in table 2. 1 indicates negative sentiment and 0 indicates positive sentiment.

Sourashtra Text	Translation	Ground Truth
Elle ghommo mogo oppaaranee	I do not like this house	1
Esani kaayathee pod bhono mathiri hoyey	If you eat like this, your stomach will become like a pot	1
Mogo elle geedh oppai	I like this song	0
Moro dhollam tu kobeem onde chokkat meni ken rai	In my eyes, you will always be a good person	0
Phul vaasano chokkat avaras	This flower smells good	0
Pos podethe vel onde padam bithir saate chod laave	It would feel amazing to watch a movie inside while it is raining	0
Thele betko chokkot meni	That boy is a good man	0
Thone huditheе chodo vaatho avude	If you open your mouth, it's all lies	1
Tu kanadee thalyedi pisi madiri dekaaras	If you wear glasses, you look stupid	1
Tu kobeem thapu ken jai	You always go the wrong way	1

Table 2: Sourashtra examples with English translations and sentiment labels.

4.2 Qualitative Analysis Overview

Among all the models evaluated, the Augmented Indo-Aryan ensemble model emerged as the best-performing, demonstrating a stronger grasp of context and negation. The Dravidian ensemble underperformed due to limited lexical and syntactic overlap with Sourashtra, which reduced the effectiveness of cross-lingual transfer. The Indo-Aryan ensemble benefited from shared vocabulary and structure, resulting in improved performance over the baseline; however, it often failed to handle negation, misclassifying sentences with negative meaning due to overreliance on sentiment-bearing keywords. In contrast, the Indo-Aryan ensemble few-shot model, which included just ten labeled Sourashtra examples, captured negation and sentence context more accurately. For instance, it correctly differentiated "Umbad baathe ruchi ken se" (tastes good) as positive and "Umbad baathe ruchi nee" (does not taste good) as negative—something the ensemble-only model failed to do. Meanwhile, the Sourashtra fine-tuned model showed strength in detecting emotions and sentiment expressed using Dravidian-origin words, such as "kadapu" (anger) and "nombike" (trust), which were less accurately recognized by the other models. The next section, Successes and Failures, explores specific cases that failed and succeed between various models.

5 Successes and Failures

To better understand model behavior (specific successes and failures) beyond accuracy scores, we analyzed specific sentence-level predictions across four models: the baseline XLM-RoBERTa, the Indo-Aryan ensemble model, the Augmented Indo-Aryan ensemble model, and the Sourashtra-only fine-tuned model.

5.1 Indo-Aryan Ensemble vs. Baseline

The Indo-Aryan ensemble model outperformed the baseline primarily on positive sentiment examples that included words shared with Hindi, Marathi, or Gujarati. For example, *"haasate"* (Sourashtra for "laugh") closely resembles *"hasane"* in Marathi and *"hansana"* in Hindi. Likewise, *"chokkat dekaaras"* (looks good) mirrors the Marathi phrase *"cangale disate"*, reinforcing sentiment associations. This lexical overlap helped the ensemble model generalize better, showing improved recognition of positive sentiment when familiar Indo-Aryan vocabulary appeared. However, the ensemble model frequently misclassified negated phrases. Sentences like *"chokkat nee"* or *"chod nee"*, where *"nee"* (not) negates a positive word, were incorrectly predicted as positive. This suggests that without explicit Sourashtra negation training, the model relied too heavily on sentiment-bearing keywords.

5.2 Indo-Aryan Ensemble vs. Augmented Indo-Aryan Model

The augmented Indo-Aryan model, which included just 10 Sourashtra sentences during fine-tuning, demonstrated clear improvements in understanding sentence context, especially with negation. For instance, the phrases *"Umbad baathe ruchi ken se"* (tastes good) and *"Umbad baathe ruchi nee"* (does not taste good) were correctly classified as positive and negative, respectively, by the augmented model—but flipped by the ensemble-only model. Another example, *"me chod neenathe bedki"* (I am not a good person), was correctly identified as negative by the augmented model, despite containing the positive root *"chod"*. This suggests the model moved beyond keyword spotting, learning basic syntactic structures and negation handling from even a handful of in-language examples.

5.3 Sourashtra Fine-Tuned vs. Augmented Indo-Aryan Model

The Sourashtra-only fine-tuned model showed stronger performance on culturally and linguistically specific vocabulary, particularly terms borrowed from Dravidian languages such as Tamil. Words like *"kadapu"* (anger), *"nombike"* (trust),

and *"porup"* (responsibility) were correctly classified by the fully fine-tuned model but misclassified by the augmented ensemble, which struggled with words less common in Indo-Aryan languages.
This indicates that the fine-tuned model learned deeper associations within the unique vocabulary and sentiment usage of Sourashtra, especially where Indo-Aryan models lacked exposure.

5.4 Potential Solutions to Failure Cases

From the above, it is clear there are still failure cases, particularly when dealing with negation, sarcasm, or ambiguous phrasing in Sourashtra. For example, some sentences use positive words like "chokkat" (good) but include subtle negations like "nee" (not), which the models occasionally miss, leading to incorrect sentiment predictions. This happens because the models may rely too heavily on keyword-based sentiment cues rather than fully understanding the sentence structure or semantics—especially in a low-resource language like Sourashtra with limited training data. Our approach also struggles with sarcasm or idiomatic expressions that are culturally specific and not well represented in the multilingual pre-training corpus. Additionally, code-mixing and non-standard spelling variations common in Sourashtra further contribute to misclassifications. Potential solutions include: 1) Expanding the labeled dataset to cover more sentence structures and edge cases, especially around negation and sarcasm; 2) Incorporating syntactic parsing or dependency structures to improve the model's handling of compositional semantics; 3) Using contrastive learning or data augmentation techniques to better teach the model how negation and sentiment interact.
However, overall, our model findings support the claim that linguistic similarity aids transfer, but domain-specific fine-tuning is essential for full coverage and nuance in minority languages like Sourashtra.

6 Discussion

Our use of publicly available pre-trained models, such as XLM-RoBERTa, and the application of widely used fine-tuned techniques on multilingual datasets means that the overall methodology of our final project is straightforward and easy to replicate in principle. Additionally, we have carefully documented our training parameters which can provide valuable information to those who would like to follow our approach. While the fine-tuning methodologies (such as using ensemble and augmented training approaches) themselves can be replicated, performance results can be sensitive to subtle differences in implementation, such as hardware configurations (GPUs and their capabilities), random seeds, and dataset preprocessing steps. One major barrier to replicability is the availability of Sourashtra datasets curated for sentiment analysis and model training. Since there are no large-scale, publicly available Sourahstra data repositories, researchers looking to reproduce our work would either need access to our custom dataset or would need to recreate it through their own data collection efforts. If someone does decide to create their own dataset, it can lead to significant changes in the accuracy of their model and final results.
While our project aims to enhance digital representation for Sourashtra, there are potential ethical concerns to consider. First, because the dataset was created by a single native annotator, there is a risk of introducing individual biases into the model's understanding of sentiment. To address this, future efforts should involve multiple annotators to ensure a more balanced and representative dataset. Additionally, there is a broader risk that sentiment analysis models could be misused for surveillance or monitoring of online speech without consent. As (Mohammad, 2022) highlights, automatic emotion recognition and sentiment analysis systems, while beneficial, can also be deployed in ways that suppress dissent or manipulate public opinion, making it critical to establish ethical safeguards. Clear disclaimers should accompany the release of our dataset and models, emphasizing that they are intended for educational and research purposes only. Furthermore, limited dataset diversity could lead to misrepresentation of the language's cultural nuances, highlighting the importance of expanding the dataset in collaboration with the Sourashtra community to ensure ethical and respectful language preservation. Minority languages often carry deep cultural, historical, and social significance, and careless or overly technical treatment risks reducing them to mere data points. In some cases, creating public datasets or models without community involvement could lead to misappropriation, exploitation, or inaccurate portrayals of the language and its speakers. Researchers must be mindful that language technology development should be done in collaboration with native-speaking communities, respecting their autonomy and cultural

values. Future projects should prioritize consent, representation, and long-term community benefit when working with minority languages to ensure that technological advancements support, rather than harm, linguistic and cultural diversity.

Although our models achieved promising results, especially with Indo-Aryan fine-tuning and data augmentation, various limitations restricted their performance. A key issue was the variability in results when small numbers of Sourashtra samples were included during training. Even though the sample size was fixed and model configurations stayed the same, performance fluctuated across different runs due to the random selection of samples. This suggests that certain samples might have been more linguistically informative or representative compared to others. Our qualitative analysis confirms this, indicating that models trained on Sourashtra examples with clear sentiment signals, negation structures or Indo-Aryan lexical overlap performed more consistently. These results suggest that it is not only the quantity of the data, but also the linguistic characteristics of training examples that affect generalization in low-resource settings. Understanding how to identify or prioritize such examples is a valuable next step for future research.

Beyond data variability, the total size and structure of our Sourashtra dataset presented further limitations. It included only 250 labeled, conversation-like sentiment sentences, all manually created and annotated by a single contributor. Although this effort provided a useful foundation for Sourashtra-specific evaluation, the models lacked exposure to a diverse range of sentiment expressions, negation patterns and domain-specific vocabulary. The dataset's limited scope and uniform author style likely limited the model's ability to generalize, especially for diverse input forms. Moreover, the additional language datasets used for fine-tuning (e.g., Hindi or Marathi) were themselves small, which restricted exposure to a wider range of sentence types during training.

Another constraint is the sensitivity to hyperparameters settings, where we noticed fluctuations in accuracy across different learning rates and training durations, with no uniform configuration performing consistently across all models. This indicates that fine-tuning XLM-RoBERTa in highly constrained settings may require adaptive optimization approaches.

To address these challenges, future research could explore principled sample selection strategies, to ensure that augmented examples meaningfully contribute to generalization. Collecting a larger and more domain-diverse Sourashtra (e.g., covering education, news or everyday conversation) would enhance the reliability of both training and evaluation. Fine-tuning on English-Sourashtra code-mixed data, proven successful in prior work on Tamil and Malayalam (Ou and Li, 2020), could further enhance model performance in mixed-language settings.

Ultimately, incorporating representational analysis, such as examining hidden state behavior using frameworks like the Linear Representation Hypothesis (Park et al., 2024), may provide deeper insight into how different fine-tuning strategies influence what the model learns. This could help identify where and why multilingual transfer fails or succeeds and guide more targeted approaches for low-resource languages like Sourashtra.

References

Jordi Armengol-Estapé, Casimiro Pio Carrino, Carlos Rodriguez-Penagos, Armentano-Oller Ona, Aitor Gonzalez-Agirre, Maite Melero, and Marta Villegas. 2021. Are multilingual models the best choice for moderately under-resourced languages? a comprehensive assessment for catalan. *arXiv preprint arXiv:2107.07903*.

Joan Carreras Timoneda and Sebastian Vallejo Vera. 2024. Bert, roberta or deberta? comparing performance across transformer models in political science text. *The Journal of Politics*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Tejas I. Dhamecha, Raghav Murthy V, Samarth Bharadwaj, Karthik Sankaranarayanan, and Pushpak Bhattacharyya. 2021. Role of language relatedness in multilingual fine-tuning of language models: A case study in indo-aryan languages. *arXiv preprint arXiv:2109.10534*.

Saif M. Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *arXiv preprint arXiv:2109.08256*.

Krenare Pireva Nuci, Paul Landes, and Barbara Di Eugenio. 2024. Roberta low resource fine tuning for sentiment analysis in albanian. In *Proceedings of the 2024 LREC*, pages 14146–14151.

Xiaolei Ou and Hao Li. 2020. Ynu@dravidian-codemix-fire2020: Xlm-roberta for multi-language sentiment

analysis. https://ceur-ws.org/Vol-2826/T4-13.pdf.

Kimin Park, Yejin J. Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, PMLR, pages 39643–39666.

Mykhailo Prytula. 2024. Fine-tuning bert, distilbert, xlm-roberta and ukr-roberta models for sentiment analysis of ukrainian language reviews. *Artificial Intelligence*, 29(2):85–97.

A Example Appendix

Model	LR	Weight Decay	Epochs	Batch Size
Gujarati fine-tuned	5e-5	0.01	3	16
Marathi fine-tuned	3e-5	0.01	3	16
Hindi fine-tuned	3e-5	0.01	5	16
Augmented Gujarati	3e-5	0.01	10	16
Augmented Marathi	3e-5	0.01	12	16
Augmented Hindi	3e-5	0.01	7	16
Sourashtra fine-tuned	1e-5	0.01	10	16
Malayalam fine-tuned	1e-5	0.01	5	16
Tamil fine-tuned	1e-5	0.01	5	16
Telugu fine-tuned	1e-5	0.01	5	16

Table 3: Hyperparameters used for fine-tuning each language-specific model.

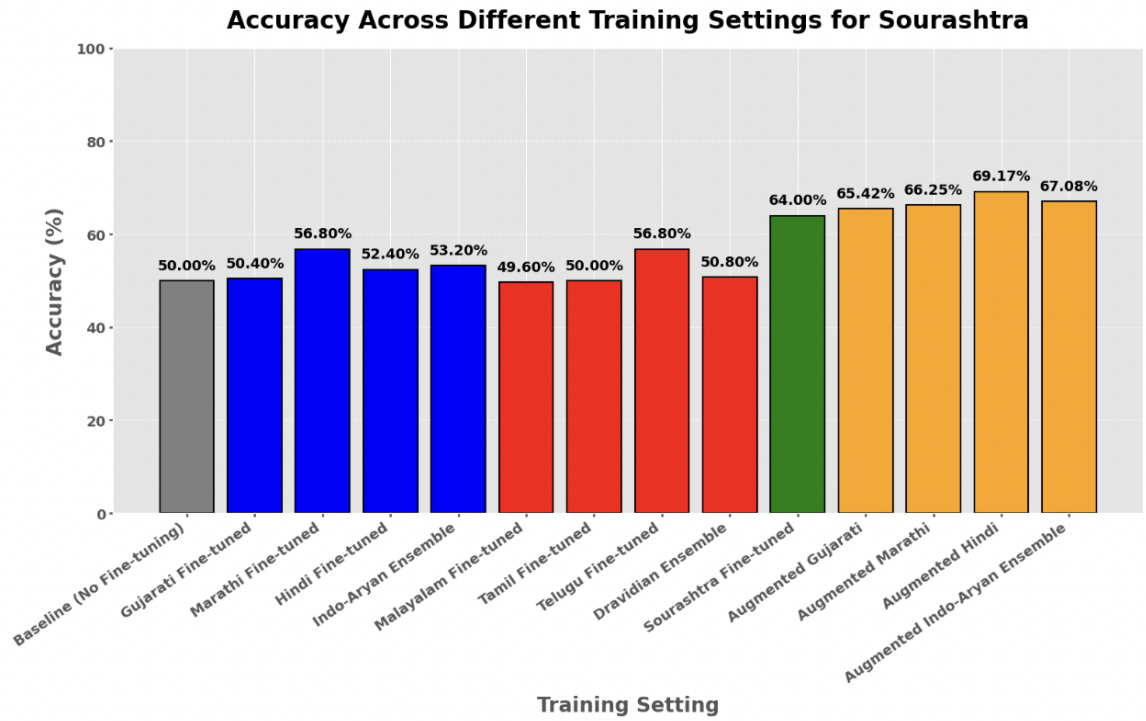

Figure 3: Accuracies across all fine-tuned models