Investigating Cross Lingual Transfer Learning for Indian Languages

Arnav Goel **Medha Hira** **Siddharth Rajput** **Amil Bhagat**

1 Introduction

Languages are fundamental to human interaction, reflecting the intricate layers of culture, history, and social dynamics. Recent advancements in transformer-based Large Language Models (LLMs) have made significant progress in the field of language processing (Vaswani et al., 2017; Raffel et al., 2020; Brown et al., 2020; Touvron et al., 2023). These models have transformed the way machines comprehend and generate text, impacting a range of applications from sentiment analysis to question-answering (Zhao et al., 2023; Yi et al., 2024). Despite these advances, the distribution of benefits has been disproportionate. Many of the world's languages, especially several Indian languages, are underrepresented in digital language resources and technologies. Often classified as low-resource, these languages lack the extensive datasets required to develop robust computational models (Joshi et al., 2021).

Our work aims to provide a solution to this problem by leveraging cross-lingual transfer learning where task-specific annotations in one language are used to fine-tune the model for evaluation in another language (Pires et al., 2019). Specifically, cross-lingual transfer learning involves training a model on Language 1 (L1) and subsequently evaluating its performance on Language 2 (L2). Typically, L1 represents a high-resource language abundant in data, whereas L2 is characterized by limited labeled data or data relevant to the task at hand.

In this paper, we aim to investigate the effectiveness of cross-lingual transfer strategies by comparing their impact on languages within the same linguistic family against those from different families. This comparative analysis allows us to delve into both intra-family and inter-family knowledge transfer within the model's embedding space. We conducted an analysis involving five languages across three different tasks. The languages include three from the Indo-Aryan family (**Bengali**, **Hindi**, **Marathi**) and two from the Dravidian family (**Tamil**, **Telugu**). Our evaluation covered two sequence classification tasks—**Sentiment Analysis** and **Hate Speech Detection**—and one sequence labeling task, namely **Named Entity Recognition** (NER). We fine-tune a multilingual-BERT (Devlin et al., 2018) encoder model for this analysis. Additionally, we provide a detailed description and analysis of all the datasets used in our study.

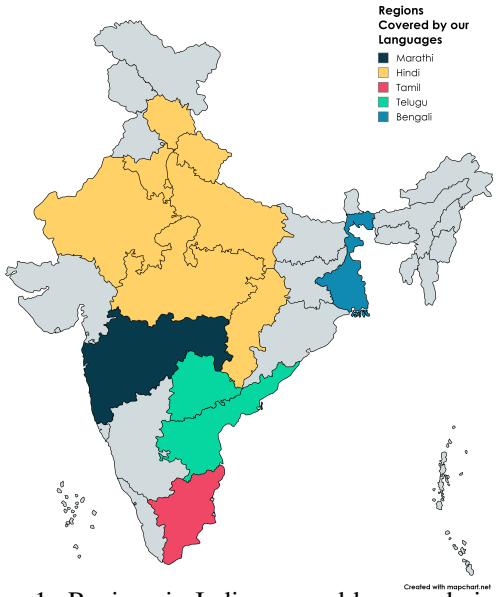


Figure 1: Regions in India covered by our choice of Languages

Figure 1 provides a visual representation of all the regions of India that have been covered by our chosen languages.

2 Related Work

2.1 Cross-Lingual Transfer Learning

While traditional NLP models were designed with language-specific features and computations, multilingual setups were engineered for explicit or latent machine translation (Bel et al., 2003; Wu et al., 2008). However the advent of neural word

embeddings and transformer-based masked language models (Devlin et al., 2018; Lample and Conneau, 2019; Ruder et al., 2019) paved the way for approaches based on multilingual representations of text, which leveraged distributional similarities across languages. This new approach involves preparing fixed semantic embeddings for input text and constructing tokenizers capable of processing tokens from multiple languages. Additionally, a joint-learning paradigm is employed, allowing simultaneous training across all languages. This strategy enhances the model's ability to generalize across linguistic boundaries, thus broadening its applicability. While not trained on a cross-lingual objective, models like MBERT have been shown to display state-of-the art results with respect to cross-lingual transfer learning on languages such as English, French and German (Pires et al., 2019).

Recent methods for cross-lingual transfer learning propose unsupervised methods to pre-train large models that show notable improvements in languages like Swahili and Urdu (Conneau et al., 2020), while more recent approaches explore supervised learning techniques, such as model fine-tuning, to enhance cross-lingual representations with reduced reliance on parallel data (Artetxe et al., 2020). Majority of the work in the domain has focused on evaluating performance of pre-trained multilingual models like mBERT, mBART and mT5 (Devlin et al., 2018; Pires et al., 2019; Xue et al., 2020) on cross-lingual setups for text classification tasks but a crucial gap can be found with respect to Indian languages. In his recent work, (Pawar et al., 2023) examined mT5 on cross-lingual transfer of morphological tags for verbs and nouns to provide a proxy for the quality of representations of word markings learned by the model.

2.2 m-BERT

Multilingual BERT (mBERT) is a language model designed to understand and generate text across multiple languages. It has been built on the architecture of the original BERT. mBERT has been pre-trained on a large corpus consisting of Wikipedia articles (unlabelled data) from 104 different languages. The studies cited below demonstrate mBERT's effectiveness in cross-lingual transfer tasks, underlining the reasons for selecting this model for our research.

(Pires et al., 2019) the capabilities of multilingual BERT at zero-shot cross-lingual model transfer, in which task-specific annotations in one language are used to fine-tune the model for evaluation in another language. Further, (K et al., 2020) examine the cross-lingual abilities of m-BERT across Spanish, Hindi, and Russian, focusing on textual entailment and NER tasks. Key findings of this paper included the minimal impact of lexical overlap and the crucial role of network depth in M-BERT's cross-lingual effectiveness.

2.3 NLP Tasks Overview

Sentiment Analysis: Analyzes the emotional tone behind texts to gauge attitudes and emotions. Techniques for sentiment analysis in Hindi are surveyed in (Kulkarni and Rodd, 2021), while Bengali comments are analyzed using Word2Vec in (Al-Amin et al., 2017). **Hate Speech Detection**: Identifies offensive or harmful language in texts, vital for maintaining decorum on social platforms. Studies include hate speech detection in Hindi and Marathi (Velankar et al., 2021), and in Bengali (Romim et al., 2021). A related track at Fire 2020 focused on multiple languages (Mandl et al., 2020). **Named Entity Recognition (NER)**: Identifies and categorizes named entities in text, crucial for extracting structured information. A language-independent approach to NER in Indian languages is presented in (Ekbal et al., 2008). While there has been increasing interest in improving performance on Indic languages for these tasks, cross-lingual learning remains unexplored which we aim to fulfill through our study.

3 Dataset

For this study, we concentrate our analysis on five languages from two distinct families i.e. Indo-Aryan (Hindi, Bengali, and Marathi) and the Dravidian family (Tamil and Telugu). Each section detailed below provides an in-depth exploration of the data's characteristics, the sources from which the data was derived, and preliminary exploratory data analysis (EDA). While curating the dataset we tried to ensure a balanced representation of each language within the constraints of data availability.

3.1 Sequence Classification

These tasks involve the classification of the entire input sequence into a fixed number of classes. Both of our tasks are binary classification tasks.

3.1.1 Sentiment Analysis

This study utilized the IndicSentiment[1] dataset, which is provided by AI4Bharat. In this dataset, 1,000 samples of each language were allocated for training, while 156 samples were designated for testing. This allocation was consistent across all five languages examined. Each sample is associated with a sentiment label: '1' indicating positive, while '0' signifies negative sentiment for the given input. Details regarding the lexical composition of the Hindi dataset are illustrated in Figure 2.

Figure 2: WordCloud for Hindi Sentiment Analysis

3.1.2 Hate Speech Detection

Since we could not find any dataset encompassing all of the five languages, we curated a dataset by combining multiple sources. For Hindi, we utilized the HASOC 2020 and 2021 hate-speech detection sub tasks (Mandl et al., 2020, 2021). Details regarding the lexical composition of the Hindi dataset curated are illustrated in Figure 3.

The Bengali dataset was obtained from (Romim et al., 2020), while for Marathi, we employed the L3-Cube Marathi Hate Speech Detection dataset (Velankar et al., 2022), specifically using the 2-class annotation system for hate and non-hate. For Tamil, we used the dataset released in collaboration by collaboration between DreamSpace Academy, NYU CIC, and Omdena [2].

For Telugu, we combined transliterated datasets from the transliterated train and test data released by (Tabassum et al., 2024) with a dataset from IIIT-H's MT-NLP Lab[3]. Although the latter dataset was extensive, containing 35,142 entries, only a small fraction (less than 1%, or 370 entries) were classified as hateful indicating an extremely unbalanced dataset. This class imbalance was tackled to a sufficient extent by merging the two datasets.

[1]https://huggingface.co/datasets/ai4bharat/IndicSentiment

[2]https://github.com/dreamspace-academy/ai-tamil-hate-speech-project/tree/master

[3]https://ltrc.iiit.ac.in/download.php

To ensure quality, we pre-processed all datasets to eliminate duplicates, NaN values, and non-language specific characters like emojis and URLs. From the consolidated dataset, we selected a subset of 5,000 entries, dividing them into 4,000 for training and 1,000 for testing, maintaining class balance throughout.

Figure 3: WordCloud for Hindi Hate Speech Detection

3.2 Sequence Labelling

These tasks involve the labelling of the entire input sequence. For the NER task, the output sequence length equals the length of the input sequence.

3.2.1 Named Entity Recognition

The data for all five aforementioned languages has been sourced from WikiANN, a multilingual NER dataset consisting of Wikipedia articles annotated in 282 languages (Pan et al., 2017).

The dataset is annotated for three categories—Person, Location, and Organization—using the BIO encoding scheme, thus resulting in seven labels for sequence labelling. An insight regarding the distribution of these categories has been provided in Figure 4.

For the languages Hindi, Bengali, Marathi, and Tamil, we selected a subset of 4,000 samples from the training set and 1,000 samples from the test set. For Telugu, due to the limited availability of data, we utilized a smaller subset consisting of 2,000 training samples and 1,000 test samples.

4 Methodology

In this section, we explain our cross-lingual transfer learning setup. We assess this setup across three downstream tasks: two sequence classification tasks—Sentiment Analysis and Hate Speech Detection—and one sequence labeling task, Named Entity Recognition (NER) as discussed above. Section 4.1 outlines the model configurations used for our evaluations, while section 4.2 details our baseline zero-shot experiments. Sections 4.3 and

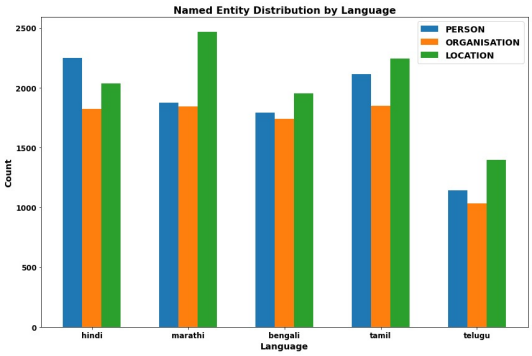


Figure 4: NER Dataset (Category-wise Distribution)

4.4 discuss our fine-tuning setups for one and two languages respectively, illustrating the efficacy of cross-lingual transfer learning.

4.1 Model Architectures

Figure 5 demonstrates our complete architecture for the entire study, across the three tasks. Through this section, we would try to expand upon the architecture relevant to each of the tasks:

- **Sequence Classification Tasks:** We utilize a consistent architecture for the sequence classification tasks—Sentiment Analysis and Hate Speech Detection—due to their similar task structure, where a single label is assigned to the entire input sequence. In this setup, as depicted in the upper branch of the figure, the input sequence is tokenized and processed through mBERT's multi-head attention encoder to generate embeddings. These embeddings are then normalised by a Batch Normalization layer, and finally fed into a classification network to output the predicted label.

- **Sequence Labelling Tasks:** For the sequence labelling task, we utilise the architecture depicted in the lower branch of Figure 5. The input sequence is tokenized, with sub-word tokenizations properly handled to ensure correct label mappings. These tokens are then processed through mBERT's multi-head attention encoder to generate embeddings. The embeddings are forwarded to a Gated-Recurrent Unit (GRU) layer, followed by normalization through a Batch Normalization Layer. To prevent overfitting, dropout is applied before the embeddings are sent to classification layers, which assign labels to each token in the sequence.

4.2 Zero-Shot Baseline Evaluation

For our baseline analysis, we examine the efficacy of pre-trained mBERT embeddings across three downstream tasks. As detailed in section 4.1, we initially **freeze the mBERT model by preventing gradient flow to its layers**. This step ensures that only the classification network following the mBERT model is trained on the respective downstream tasks and mBERT embeddings remain pre-trained. We regard this setup as our baseline because it assesses mBERT's zero-shot cross-lingual capabilities on Indic languages without any additional fine-tuning.

4.3 SFT-One Language

In our first setup for demonstrating cross-lingual transfer learning, we train on one language L_1 and subsequently evaluate the fine-tuned model on others languages L_i where $i \in [1, 2, 3, 4, 5]$. Our proposed methodology consists of three steps:

- Initialise the pre-trained mBERT model and unfreeze it to allow gradient to backpropagate till its input layer.

- Subsequently, we perform supervised fine-tuning on the entire network using the training data from language L_1, applying a task-specific training objective appropriate for the downstream task (thus fine-tuning the mBERT embeddings).

- Lastly, upon model training, we evaluate the model on testing splits of all the 5 languages L_1, L_2, L_3, L_4 and L_5 by simply passing the data through the trained network and performing inference.

For further analysis, we conducted experiments not only with the entire 100% training data split but also with 25% and 50% of the training split. This approach allows us to assess the impact of training data volume on facilitating cross-lingual learning more effectively. We call these **Reduced Training Data Experiments**.

4.4 SFT-Two Languages

Our final approach assesses the efficacy of cross-lingual transfer when training on a pair of two languages instead of just one as described above. We follow the same 3-step strategy as described in section 4.3 with a minor modification of curating a combined dataloader by mixing training splits of

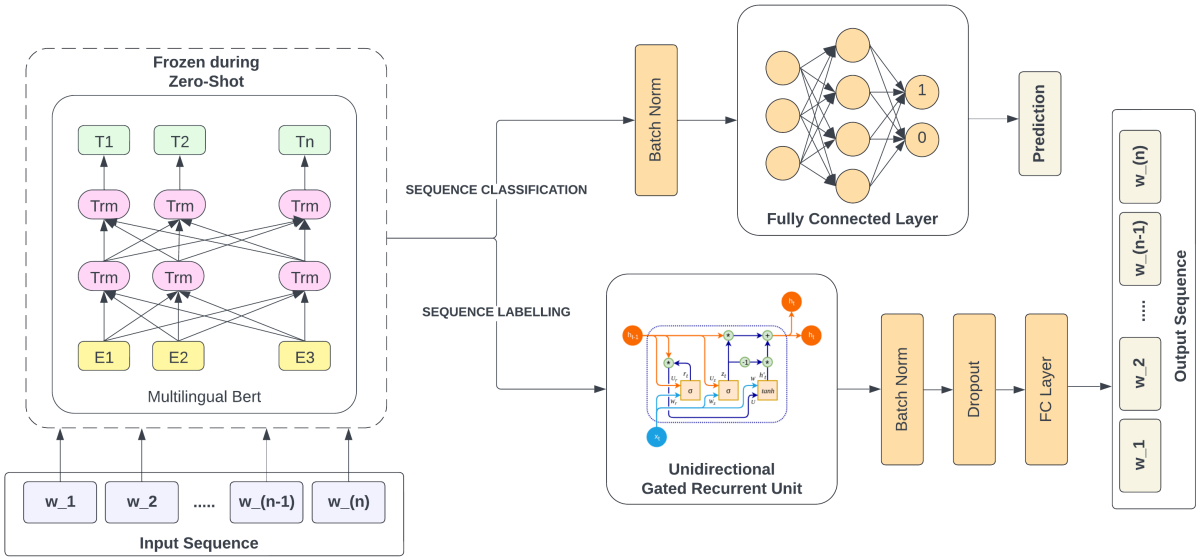


Figure 5: Model Architecture for the Study

Languages →	Hi (Hindi)			Mr (Marathi)			Bn (Bengali)			Ta (Tamil)			Te (Telugu)		
Fine-Tuning ↓	**Acc**	**W-F1**	**M-F1**	**Acc**	**W-F1**	**M-F1**	**Acc**	**W-F1**	**M-F1**	**Acc**	**W-F1**	**M-F1**	**Acc**	**W-F1**	**M-F1**
Zero Shot Evaluation															
Zero-Shot	0.417	0.417	0.416	0.564	0.559	0.557	0.622	0.619	0.617	0.500	0.494	0.492	0.526	0.521	0.523
Fine-Tuning on One Language															
Hi	*0.916*	*0.916*	*0.916*	0.788	0.786	0.785	**0.852**	**0.852**	**0.852**	0.807	0.807	0.807	0.782	0.781	0.782
Mr	0.634	0.585	0.590	*0.794*	*0.790*	*0.791*	**0.647**	**0.604**	**0.608**	0.634	0.600	0.604	0.602	0.537	0.543
Bn	**0.852**	**0.852**	**0.852**	0.769	0.769	0.769	*0.865*	*0.865*	*0.865*	0.750	0.748	0.748	0.705	0.689	0.692
Ta	**0.859**	**0.858**	**0.858**	0.737	0.729	0.727	0.807	0.807	0.806	*0.820*	*0.818*	*0.817*	0.775	0.773	0.774
Te	**0.846**	**0.846**	**0.845**	0.782	0.773	0.771	0.756	0.747	0.745	0.769	0.765	0.764	*0.846*	*0.846*	*0.845*
Fine-Tuning on Two Languages															
Hi-Bn	*0.878*	*0.878*	*0.878*	**0.775**	**0.775**	**0.775**	*0.859*	*0.858*	*0.858*	0.775	0.773	0.774	0.717	0.706	0.708
Hi-Mr	*0.878*	*0.877*	*0.878*	*0.859*	*0.858*	*0.858*	**0.826**	**0.826**	**0.826**	0.756	0.755	0.756	0.769	0.769	0.769
Hi-Ta	*0.826*	*0.823*	*0.824*	**0.794**	**0.794**	**0.794**	0.730	0.719	0.721	*0.807*	*0.804*	*0.805*	0.685	0.656	0.660
Hi-Te	*0.891*	*0.891*	*0.890*	**0.814**	**0.812**	**0.811**	0.788	0.788	0.787	0.788	0.788	0.788	*0.846*	*0.845*	*0.845*
Mr-Bn	**0.801**	**0.799**	**0.800**	*0.833*	*0.833*	*0.833*	*0.859*	*0.859*	*0.858*	0.769	0.766	0.767	0.750	0.744	0.745
Mr-Ta	**0.833**	**0.833**	**0.833**	*0.846*	*0.845*	*0.845*	0.826	0.826	0.826	*0.820*	*0.820*	*0.820*	0.769	0.767	0.767
Mr-Te	**0.788**	**0.788**	**0.788**	*0.859*	*0.858*	*0.858*	0.775	0.767	0.765	0.750	0.742	0.741	*0.878*	*0.877*	*0.876*
Bn-Ta	0.820	0.820	0.820	**0.833**	**0.833**	**0.833**	*0.852*	*0.850*	*0.851*	*0.794*	*0.790*	*0.791*	0.692	0.664	0.668
Bn-Te	**0.865**	**0.865**	**0.865**	0.762	0.758	0.756	*0.859*	*0.859*	*0.858*	0.782	0.779	0.778	*0.878*	*0.878*	*0.877*
Ta-Te	**0.871**	**0.871**	**0.871**	0.775	0.774	0.773	0.775	0.772	0.771	*0.814*	*0.814*	*0.814*	*0.859*	*0.859*	*0.858*

Table 1: Results for Cross-Lingual Experiments on Sentiment Analysis

2 languages. The entire network is fine-tuned in a supervised learning paradigm on this combined dataloader (thus fine-tuning the mBERT embeddings). This helps us visualise use of inter-family diversities in languages to help facilitate cross-lingual transfer.

5 Experimental Setup

In this section, we expand upon our experimental setup for each of the three tasks which helped us implement the setups described in section 4.

5.1 Sequence Classification Task Setup

For our Sentiment Analysis and Hate Speech Detection tasks, we employed a similar experimental setup, reflecting the tasks' comparable nature. We fine-tuned our network using a binary cross-entropy loss objective, coupled with an Adam optimizer set to a learning rate of 2×10^{-5}. The networks were fine-tuned on NVIDIA Tesla T4 GPUs over a period of 4 epochs. The fine-tuning setup outlined in section 4.3 was applied to train models on all five languages, with subsequent evaluations conducted on the remaining four languages. Additionally, models were trained using 25% and 50% of the total data set. This approach resulted in 12 unique model configurations for each downstream task.

5.2 Sequence Labelling Task Setup

For our Named Entity Recognition (NER) task, special attention was required during the tokenisation phase as due to mBERT's tokeniser using sub-word tokenisation to handle out of vocabulary words. This meant ensuring that sub-words obtained from a token share the same label from the task to prevent misalignment. This was ensured

Languages →	Hi (Hindi)			Mr (Marathi)			Bn (Bengali)			Ta (Tamil)			Te (Telugu)		
Fine-Tuning ↓	**Acc**	**W-F1**	**M-F1**	**Acc**	**W-F1**	**M-F1**	**Acc**	**W-F1**	**M-F1**	**Acc**	**W-F1**	**M-F1**	**Acc**	**W-F1**	**M-F1**
Zero Shot Evaluation															
Zero-Shot	0.550	0.549	0.549	0.549	0.546	0.546	0.596	0.596	0.596	0.603	0.593	0.550	0.532	0.532	0.532
Fine-Tuning on One Language															
Hi	*0.734*	*0.734*	*0.734*	**0.736**	**0.734**	**0.734**	0.658	0.646	0.646	0.600	0.593	0.553	0.549	0.525	0.525
Mr	0.632	0.614	0.614	*0.863*	*0.862*	*0.862*	**0.639**	**0.628**	**0.628**	0.632	0.618	0.574	0.520	0.433	0.433
Bn	0.534	0.431	0.431	0.580	0.506	0.506	*0.885*	*0.885*	*0.885*	**0.655**	**0.591**	**0.520**	0.510	0.387	0.387
Ta	0.536	0.493	0.493	0.592	0.560	0.560	0.544	0.479	0.479	*0.767*	*0.758*	*0.730*	**0.557**	**0.475**	**0.475**
Te	0.568	0.564	0.564	**0.623**	**0.607**	**0.607**	0.617	0.585	0.585	0.561	0.567	0.540	*0.746*	*0.745*	*0.745*
Fine-Tuning on Two Languages															
Hi-Bn	*0.714*	*0.708*	*0.708*	**0.739**	**0.738**	**0.738**	*0.864*	*0.863*	*0.863*	0.635	0.617	0.571	0.560	0.515	0.515
Hi-Mr	*0.710*	*0.707*	*0.707*	*0.872*	*0.872*	*0.872*	**0.669**	**0.668**	**0.668**	0.619	0.622	0.594	0.575	0.560	0.560
Hi-Ta	*0.735*	*0.734*	*0.734*	**0.740**	**0.740**	**0.740**	0.584	0.567	0.567	*0.754*	*0.753*	*0.732*	0.563	0.532	0.532
Hi-Te	*0.717*	*0.716*	*0.716*	**0.700**	**0.692**	**0.692**	0.673	0.661	0.661	0.574	0.579	0.549	*0.756*	*0.755*	*0.755*
Mr-Bn	**0.664**	**0.658**	**0.658**	*0.834*	*0.831*	*0.831*	*0.819*	*0.814*	*0.814*	0.629	0.627	0.594	0.585	0.565	0.565
Mr-Ta	0.648	0.638	0.638	*0.866*	*0.865*	*0.865*	**0.653**	**0.645**	**0.645**	*0.764*	*0.766*	*0.749*	0.574	0.544	0.544
Mr-Te	0.631	0.630	0.630	*0.858*	*0.857*	*0.857*	**0.637**	**0.597**	**0.597**	0.558	0.566	0.551	*0.770*	*0.768*	*0.768*
Bn-Ta	0.577	0.568	0.568	**0.653**	**0.652**	**0.652**	*0.872*	*0.871*	*0.871*	*0.761*	*0.764*	*0.749*	0.574	0.559	0.559
Bn-Te	0.587	0.558	0.558	**0.684**	**0.680**	**0.680**	*0.890*	*0.890*	*0.890*	0.645	0.622	0.573	*0.758*	*0.757*	*0.757*
Ta-Te	0.561	0.560	0.560	**0.677**	**0.675**	**0.675**	0.666	0.660	0.660	*0.739*	*0.724*	*0.689*	*0.760*	*0.758*	*0.758*

Table 2: Results for Cross-Lingual Experiments on Hate Speech Detection

Languages →	Hi (Hindi)			Mr (Marathi)			Bn (Bengali)			Ta (Tamil)			Te (Telugu)		
Fine-Tuning ↓	**Acc**	**W-F1**	**M-F1**	**Acc**	**W-F1**	**M-F1**	**Acc**	**W-F1**	**M-F1**	**Acc**	**W-F1**	**M-F1**	**Acc**	**W-F1**	**M-F1**
Zero Shot Evaluation															
Zero-Shot	0.668	0.638	0.526	0.726	0.691	0.516	0.664	0.658	0.639	0.680	0.632	0.424	0.674	0.567	0.199
Fine-Tuning on One Language															
Hi	*0.933*	*0.933*	*0.914*	**0.828**	**0.824**	0.725	0.793	0.789	**0.790**	0.799	0.801	0.700	0.816	0.823	0.668
Mr	0.802	0.804	0.755	*0.935*	*0.935*	*0.897*	0.765	0.760	**0.759**	0.816	0.817	0.715	**0.822**	**0.828**	0.678
Bn	0.769	0.774	**0.721**	0.771	0.780	0.670	*0.940*	*0.940*	*0.940*	0.776	0.777	0.664	**0.810**	**0.810**	0.639
Ta	0.802	0.802	0.752	0.839	0.834	0.731	0.761	0.753	**0.754**	*0.913*	*0.865*	*0.913*	**0.840**	**0.844**	0.701
Te	0.762	0.763	0.713	**0.816**	**0.811**	0.708	0.727	0.721	**0.720**	0.787	0.785	0.671	*0.930*	*0.930*	*0.867*
Fine-Tuning on Two Languages															
Hi-Bn	*0.930*	*0.930*	*0.911*	**0.853**	**0.854**	**0.764**	*0.956*	*0.956*	*0.955*	0.805	0.808	0.710	0.812	0.821	0.670
Hi-Mr	*0.928*	*0.906*	*0.928*	*0.934*	*0.934*	*0.894*	0.791	0.785	**0.785**	**0.817**	0.819	0.722	0.814	**0.824**	0.665
Hi-Ta	*0.931*	*0.930*	*0.909*	**0.872**	**0.870**	**0.795**	0.778	0.777	0.777	*0.911*	*0.911*	*0.855*	0.840	0.843	0.701
Hi-Te	*0.929*	*0.929*	*0.909*	**0.846**	**0.844**	0.754	0.783	0.778	**0.776**	0.809	0.812	0.711	*0.927*	*0.929*	*0.867*
Mr-Bn	**0.828**	**0.827**	**0.775**	*0.935*	*0.934*	*0.894*	*0.945*	*0.945*	*0.945*	0.815	0.816	0.718	0.816	0.824	0.676
Mr-Ta	0.818	0.818	0.772	*0.940*	*0.939*	*0.902*	0.787	0.780	**0.781**	*0.915*	*0.916*	*0.869*	**0.839**	**0.847**	0.715
Mr-Te	0.830	0.831	0.782	*0.936*	*0.936*	*0.896*	0.795	0.790	**0.788**	**0.857**	**0.857**	0.776	*0.933*	*0.933*	*0.871*
Bn-Ta	0.775	0.780	**0.727**	0.793	0.794	0.681	*0.952*	*0.952*	*0.950*	*0.911*	*0.912*	*0.861*	**0.829**	**0.836**	0.691
Bn-Te	0.789	0.792	**0.740**	0.792	0.798	0.685	*0.937*	*0.937*	*0.938*	**0.802**	**0.806**	0.697	*0.922*	*0.923*	*0.856*
Ta-Te	0.795	0.794	0.735	**0.849**	**0.847**	**0.759**	0.725	0.712	0.710	*0.908*	*0.908*	*0.853*	*0.931*	*0.932*	*0.871*

Table 3: Results for Cross-Lingual Experiments on Named Entity Recognition

by using offset mappings returned by the tokeniser which helped map the sub-words to parent token labels and special and padded tokens with -100 labels.

Subsequently, for training the network we employed a Cross Entropy Loss objective coupled with a AdamW optimiser set to a learning rate of 2×10^{-5}. The padded indexes were ignored during evaluation and loss propagation in the network. The network was fine-tuned on NVIDIA Tesla T4 GPUs over a period of 6 epochs.

5.3 Evaluation Metrics

Since we are primarily performing text classification tasks, we employ the standard classification metrics i.e. **Accuracy and F1-score**. For taking class imbalance into account we report the weighted and macro-F1 scores as well. For the NER task, while we report the accuracy we will only analyse the weighted and macro-F1 due to the higher

6 Results and Analysis

Results for our experiments are summarized in Tables 1, 2, and 3 for the tasks of Sentiment Analysis, Hate Speech Detection, and Named Entity Recognition (NER), respectively. The rows in these tables represent the languages or language pairs on which the networks were fine-tuned, while the columns indicate the languages on which the fine-tuned mod-

els were evaluated. In this section, we provide a detailed analysis of our results, beginning with an examination of cross-lingual transfer within linguistic families—specifically Indo-Aryan and Dravidian (Section 6.1). We then explore the transfer between these two families (Section 6.2).

Subsequently, we analyse results of experiments wherein we fine-tune our models on a pair of two languages (Section 6.3). Lastly, we focus on Hindi and Tamil, representing each linguistic family, to analyse the impact of training data size on transfer quality by experimenting with reduced training splits (Section 6.4).

6.1 Intra-Family Transfer

Sentiment Analysis Experiments: In zero-shot evaluations, pre-trained mBERT models showed limited effectiveness on Sentiment Analysis, achieving a maximum accuracy of 62.2% on the Bengali split. Notably, within the Indo-Aryan family, Bengali showed substantial improvement, increasing by **almost 23 percentage points** in accuracy when fine-tuned on the Hindi training set. In contrast, using the Marathi dataset for training slightly reduced Bengali's performance. Hindi exhibited strong cross-lingual benefits when trained on the Bengali training split. Within the Dravidian family, Tamil and Telugu demonstrated significant intra-family transfer, with performance improvements of nearly **25 percentage points** when trained on each other's datasets.

Hate Speech Detection Experiments: On zero-shot evaluations, models again perform dismally with a highest accuracy of merely 59%. Within the Indo-Aryan Family, highest quality transfer is achieved when fine-tuned on Hindi as we observe a performance jump of almost **19 points and 7 points on Marathi and Bengali respectively**. Within the Dravidian family, Tamil and Telugu are not able to transfer effectively amongst each other likely due to the presence of code-mixed text in the data.

NER Experiments: On zero-shot evaluations, the pre-trained mBERT embeddings obtain the highest macro-F1 of 0.639 on Bengali but performs poorly on Tamil and Telugu. In the Indo-Aryan languages, results are promising when fine-tuned on Hindi seeing an average jump of **almost 18 points on macro-F1**. Similar results are observed when fine-tuned on the Bengali train split. Within the Dravidian family, fine-tuning on Tamil resulted in a notable **50-point** increase for Telugu, while Tamil saw an improvement of **25 points** when fine-tuned on Telugu, demonstrating the effectiveness of our methodology.

Thus, we can infer that our experiments demonstrate a high degree of intra-family cross lingual transfer across the two families.

6.2 Inter-Family Transfer

Upon comparing cross-lingual transfer abilities for sentiment analysis between the Indo-Aryan and Dravidian families, we find some interesting observations.

Sentiment Analysis Experiments: When fine-tuned on Tamil and Telugu, we observe the highest accuracy across all the 5 languages on the Hindi test split **with an increase of almost 43 points.** Similiarly, fine-tuning on Indo-Aryan languages shows an **average increase of 15 points** on Tamil and Telugu.

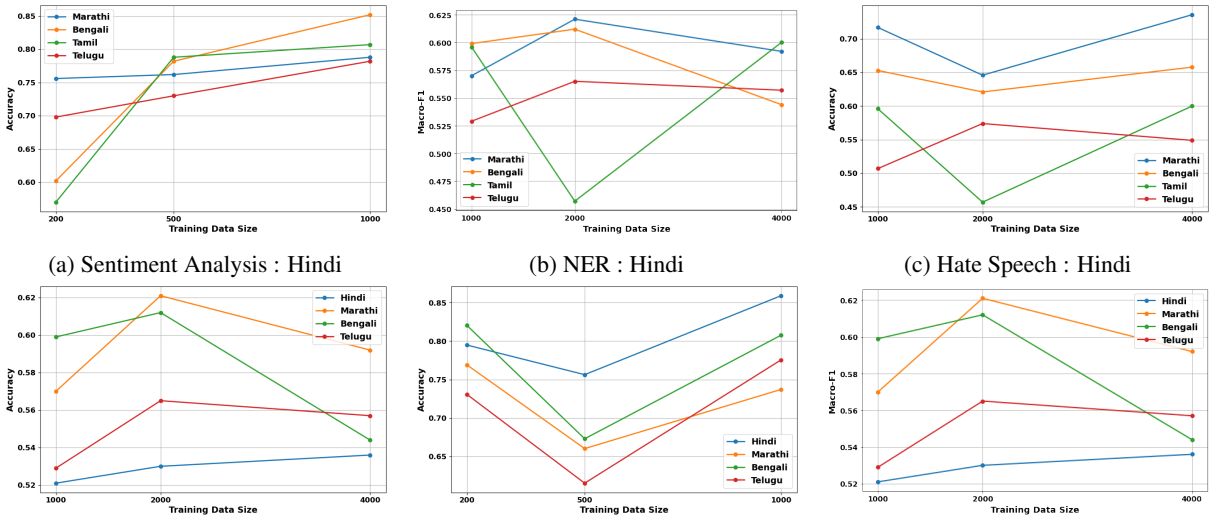
Hate Speech Detection Experiments: Fine-tuning on Dravidian languages leads to an dismal increase here as our dataset was transliterated into English. Similiar results can be seen when fine-tuned on Indo-Aryan languages. This holds with theoretical studies such as (Pires et al., 2019) which indicate poor cross lingual transfer when transliterated text is utilised.

NER Experiments: When fine-tuned on Dravidian languages, we observe an average jump in the macro-F1 scores of almost 18 points while when fine-tuned on Indo-Aryan languages, we observe a jump of almost **30 points on Tamil** and a notable jump of almost **50 points on Telugu.**

Our results indicate a very high inter-family cross lingual transfer between the two families. This is especially notable in tasks like NER which can be attributed to morphological transfer across scripts and better contextual understanding helping cross-lingual transfer in the task given at hand.

6.3 Two Language Fine-Tuning Results

In this section, we delve into the outcomes of fine-tuning across language pairs, examining both intra-family (same linguistic family) and inter-family (different linguistic families) language combinations. Our Sentiment Analysis results show exceptional performance when fine-tuning jointly on Tamil and Telugu, particularly when evaluated



(a) Sentiment Analysis : Hindi (b) NER : Hindi (c) Hate Speech : Hindi

(d) Hate Speech : Tamil (e) Sentiment Analysis : Tamil (f) NER : Tamil

Figure 6: Reduced Training Data Experiment Results

on the Hindi test split, achieving an **accuracy of nearly 87%**.

For Hate Speech Detection, the most effective results come from a joint Indo-Aryan language setup, such as Hindi-Bengali, which leads to the highest performance **(74% accuracy)** when applied to Tamil, Telugu, and Marathi, showcasing notable intra and inter-family transfer capabilities. Similarly, in Named Entity Recognition (NER), a joint language approach significantly enhances cross-lingual transfer, with an average improvement of roughly 20 percentage points for Indo-Aryan languages and nearly 40 points for Dravidian languages. These findings underscore the potent cross-lingual transfer potential when leveraging pairs of high-resource languages.

6.4 Reduced Training Data Experiment Results

In Figure 6, we illustrate the performance trends across varying training data sizes for sentiment analysis, hate speech detection, and named entity recognition tasks using m-BERT across Indo-Aryan and Dravidian languages. In sentiment analysis, Hindi displayed an upward trajectory in accuracy with larger datasets, while Marathi exhibited consistent improvement. Bengali initially declined before recovering with increased data, indicating sensitivity. Tamil demonstrated steady enhancement, while Telugu's accuracy fluctuated. In hate speech detection, Marathi showcased steady improvement, Bengali's performance fluctuated, and Tamil peaked before declining with larger datasets, while Telugu showed notable improvement. Named entity recognition exhibited diverse trends, with Bengali initially excelling but experiencing fluctuations, Marathi and Hindi showing steady improvement, and Tamil and Telugu displaying volatility.

7 Impact of our Work

This research on cross-lingual transfer learning for Indian languages has significant implications for computational linguistics, especially in tasks like Hate Speech Detection, where it can help combat the proliferation of harmful content online. By demonstrating effective methodologies for Indian languages with limited datasets, the study paves the way for the development of NLP tools that can be extended to other low-resource languages, promoting linguistic diversity and technological inclusivity. The impact is thus twofold: advancing language technology and fostering a safer, more equitable digital landscape.

8 Conclusion and Future Work

Thus we can reach the conclusion that our work demonstrates an effective cross-lingual setup on Indian languages on the given three tasks. We aim to further this work in the future by integrating more Indian languages and trying multi-task and unsupervised setups to further improve performance.

9 Ethics Statement

All data used in this paper has acquired ethically and the original data sources have been cited appropriately. The content of the Hate Speech and Sentiment Analysis data can have extreme language, which does not reflect the opinions of myself or my collaborators. Reader's discretion is advised while viewing the dataset.

References

Md Al-Amin, Md Saiful Islam, and Shapan Das Uzzal. 2017. Sentiment analysis of bengali comments with word2vec and sentiment information of words. In *2017 international conference on electrical, computer and communication engineering (ECCE)*, pages 186–190. IEEE.

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Nuria Bel, Cornelis HA Koster, and Marta Villegas. 2003. Cross-lingual text categorization. In *Research and Advanced Technology for Digital Libraries: 7th European Conference, ECDL 2003 Trondheim, Norway, August 17-22, 2003 Proceedings 7*, pages 126–139. Springer.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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. *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*.

Asif Ekbal, Rejwanul Haque, Amitava Das, Venkateswarlu Poka, and Sivaji Bandyopadhyay. 2008. Language independent named entity recognition in indian languages. In *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2021. The state and fate of linguistic diversity and inclusion in the nlp world. *Preprint*, arXiv:2004.09095.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. *Preprint*, arXiv:1912.07840.

Dhanashree S Kulkarni and Sunil S Rodd. 2021. Sentiment analysis in hindi—a survey on the state-of-the-art techniques. *Transactions on Asian and Low-Resource Language Information Processing*, 21(1):1–46.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Proceedings of the 12th annual meeting of the forum for information retrieval evaluation*, pages 29–32.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Prasenjit Majumder, Johannes Schäfer, Tharindu Ranasinghe, Marcos Zampieri, Durgesh Nandini, et al. 2021. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages. *arXiv preprint arXiv:2112.09301*.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Siddhesh Pawar, Pushpak Bhattacharyya, and Partha Talukdar. 2023. Evaluating cross lingual transfer for morphological analysis: a case study of Indian languages. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 14–26, Toronto, Canada. Association for Computational Linguistics.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam. 2020. Hate speech detection in the bengali language: A dataset and its baseline evaluation. *Preprint*, arXiv:2012.09686.

Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam. 2021. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In *Proceedings of International Joint*

Conference on Advances in Computational Intelligence: IJCACI 2020, pages 457–468. Springer.

Sebastian Ruder, Anders Søgaard, and Ivan Vulić. 2019. Unsupervised cross-lingual representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 31–38, Florence, Italy. Association for Computational Linguistics.

Nafisa Tabassum, Mosabbir Khan, Shawly Ahsan, Jawad Hossain, and Mohammed Moshiul Hoque. 2024. Sandalphon@DravidianLangTech-EACL2024: Hate and offensive language detection in Telugu code-mixed text using transliteration-augmentation. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 167–172, St. Julian's, Malta. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Abhishek Velankar, Hrushikesh Patil, Amol Gore, Shubham Salunke, and Raviraj Joshi. 2021. Hate and offensive speech detection in hindi and marathi. *arXiv preprint arXiv:2110.12200*.

Abhishek Velankar, Hrushikesh Patil, Amol Gore, Shubham Salunke, and Raviraj Joshi. 2022. L3cube-mahahate: A tweet-based marathi hate speech detection dataset and bert models. *Preprint*, arXiv:2203.13778.

Ke Wu, Xiaolin Wang, and Bao-Liang Lu. 2008. Cross language text categorization using a bilingual lexicon. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Zihao Yi, Jiarui Ouyang, Yuwen Liu, Tianhao Liao, Zhe Xu, and Ying Shen. 2024. A survey on recent advances in llm-based multi-turn dialogue systems. *Preprint*, arXiv:2402.18013.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *Preprint*, arXiv:2303.18223.