Cross-Lingual Learning and Dravidian Languages

Amrita Nair
Universität des Saarlandes
amritahnair@gmail.com

Abstract

Despite their richness and diversity, low-resource languages have not received as much attention from NLP researchers as high-resource languages like English and Spanish. However, recent progress in transfer learning, unsupervised learning, and data augmentation techniques show promise for improving NLP systems for low-resource languages. Leveraging the latent symmetry learned by multilingual language models through joint training, this report explores how cross-lingual learning can benefit the understanding of Dravidian languages, specifically, Telugu, Malayalam and Tamil. The report covers tasks related to question answering, transliteration, code-switching, and hate speech detection. This non-exhaustive survey aims to facilitate further research in these important and socially beneficial tasks.

1 Introduction

The Dravidian language family comprises over 70 languages spoken primarily in southern India, as well as in parts of Sri Lanka, Pakistan, Nepal, and Bangladesh Britannica (2023). These languages are further classified into subgroups based on their respective regions.

Despite the diversity of this language family, this report will focus only on the three major ones. These languages are Telugu, Malayalam, and Tamil spoken predominantly in the Indian states of Andhra Pradesh/Telangana, Kerala, and Tamil Nadu, respectively.

The primary reason for selecting these languages is the availability of data, which is easier to obtain for these three languages compared to the rest. Moreover, these languages are among the most widely spoken in the Dravidian language family and have rich literary and cultural traditions, making them ideal candidates for research and analysis.

Recent advances in Natural Language Processing in various tasks like Name Entity Recognition(NER) Li et al. (2020), Machine Translation Stahlberg (2020), Question Answering Zhu et al. (2021) tend to focus on high-resource languages like English, Spanish and so on. This is due to the abundance of high quality annotated data available for these languages, which makes developing systems for them easy and accessible.

On the other hand, low-resource languages, like the Dravidian language family, have not received the same attention. These languages lack the same amount of high-quality annotated data, which makes it difficult to develop effective NLP systems for them. As a result, there is a significant gap between the performance of NLP systems in high-resource languages and those in low-resource languages like Dravidian.

Despite the challenges, there has been some progress in NLP research on low-resource languages. Researchers have been exploring ways to improve the performance of NLP systems in low-resource languages using transfer learning, unsupervised learning, and data augmentation techniques. These methods have shown promising results and are expected to further enhance the performance of NLP systems for low-resource languages.

Advances in the understanding of word representations, especially in the way that monolingual-BERT Devlin et al. (2019) works, show that it is possible to align the word representations across different languages, which in turn makes it possible to use the word represen-

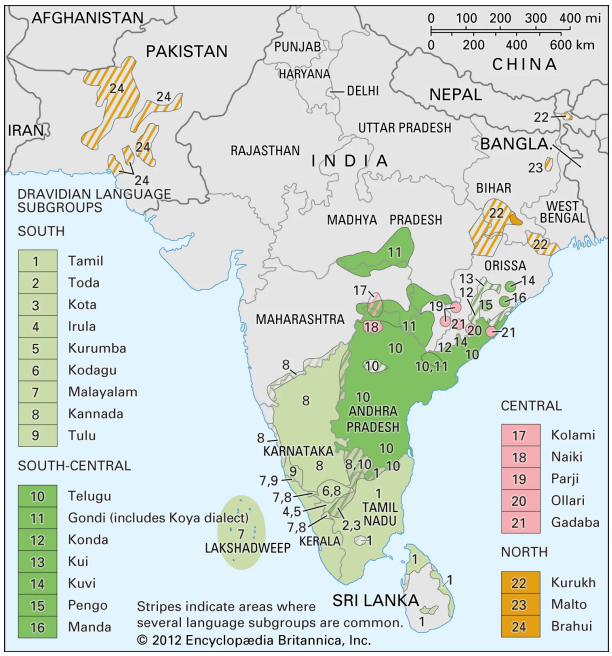

Figure 1: Britannica (2023) The distribution of Dravidian languages across the Indian subcontinent. The subgroups are south, south-central, Central and North.

tations of one language(high-resource) in solving tasks in other low-resource languages Wu et al. (2019b). This is possible because multilingual language models learn this latent symmetry during the joint training process. This process is known as *cross-lingual learning*.

Leveraging this capability of multilingual large language models can provide researchers with invaluable tools, which will help attain higher accuracy's and performance across tasks in low-resource languages.

Research in this direction is necessary, since, according to the official 2001 census of India, more than 193 million people spoke Telugu, Tamil and Malayalam Wikipedia (2023). A growing number of people communicate on the internet in either romanized or the original scripts of these languages, some of which is code-switched (mostly with English).

Effectively translating and managing this data is essential in many cases, for example, in dealing with hate speech Priyadharshini et al. (2022). Moreover, the development of effective NLP systems for low-resource languages is crucial to enable access to digital services for a wider population, including those living in remote and underdeveloped areas where these languages are predominantly spoken. This would lead to greater language inclusivity and promote linguistic diversity, which is an essential component of human culture and heritage.

This report seeks to compile and survey a non-exhaustive list of advances made in this direction, i.e., answer the question, *How has cross-lingual learning benefited in the understanding of Dravidian languages?*

The report is divided into different sections, each section describes an aspect of dealing with low-resourced languages, specifically, Telugu, Tamil and Malayalam.

The first task is based on question answering, particularly in passage retrieval for dense models given a query. The next section deals with another important topic in the context of regional languages, that is, transliteration and code-switching. The last section deals with offensive and hate speech detection.

Overall, this report focuses on tasks that are deemed important from a practical perspective and necessary for social good. This report does not intend to be an exhaustive list of all tasks on which research has been conducted for these languages, rather, it provides a short survey to facilitate further research in these tasks.

2 What is cross-lingual learning, and why is it useful for under-resourced languages ?

As described in the introduction, cross-lingual learning is loosely defined as the situation when the knowledge gained through the training of one language is used/transferred to another language.

Pires et al. (2019) conducted a series of experiments to demonstrate that mBert (Devlin et al., 2019), which is a pre-trained language model trained on 104 languages, is able to perform *zero-shot* cross-lingual transfer, even when the scripts for the language pairs were not the same. Transfer works best for languages that are typologically similar, and the model tends to find pairs (among all the languages that it is trained on) between which cross-lingual transfer would be optimal.

This finding is further emphasised by Wu et al. (2019b) who assert that lexical overlap or domain similarity is not needed for transfer to occur, but there should be some shared parameters in the top levels of the multilingual encoder. They further show that, *word representations of vocabularies of different languages can be aligned efficiently post-hoc, suggesting latent symmetries in the learning of the large language models.*

This finding is particularly useful, especially for under-resourced languages, since it provides hope that tasks in these languages can be performed despite the lack of data, since the abundance of data in high-resource languages can be leveraged for learning embeddings and representations and patterns. Through this paper, data augmentation is also used as a method to supplement the lack of data.

3 Multilingual Dense Retrieval Models for Dravidian Languages

Retrieving a set of documents, given a query, is a standard NLP task. These documents further need to be ranked in order of relevance to the query. Traditionally, sparse retrieval models like bag-of-words would match occurrences of words in the query to the words in the document, completely disregarding the word order and treating each document as a collection of words (hence bag-of-words), Sparck Jones and Willet (1972).

Recent methods focus on "dense" retrieval methods, where, each document is mapped to a low-dimensional encoding. The query is also mapped to this low-dimensional space and the relevant documents are retrieved using algorithms like ANN (Approximate Nearest Neighbour) search Wu et al. (2019a) . These methods are considered to be more reliable for cases where the context/semantic similarity of the query/documents are considered paramount.

It is possible that a dense retrieval approach may not be a sure-shot solution. Luan et al. (2021) demonstrated that this is the case, since these dense retrieval models seem to suffer from the drawback of low performance for tasks that require word-overlap.

These approaches seem relatively reliable for scenarios involving high-resources languages, and it is interesting to see if it would be possible to apply these same concepts for the multilingual case, such as cases where a dense retrieval model needs to be built for a low-resource language. Zhang et al. (2022) explore and discuss this angle. The paper mentions that retrieval systems that are based on multilingual models are able to generalize well across languages, i.e., *the model can be trained in one language and inference can be applied in another language for ranking. These transformer-based retrieval models (example, dual encoders) can perform cross-lingual transfer.*

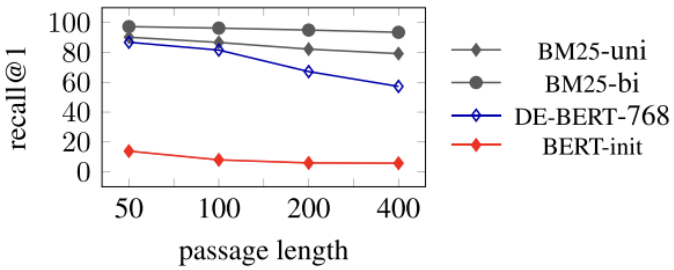

Figure 2: Luan et al. (2021) This figure compares the recall@1 (against the passage length) for retrieving a passage for a query by different models. BM25-uni and BM25-bi are sparse models, while the other two are dense models. Surprisingly, the sparse models perform better than the dense models.

The authors use a modified version of DPR(Dense Passage Retriever) Karpukhin et al. (2020) which is a dual encoder based passage retrieval model. A dual encoder based model encodes the passages and queries independently, and the similarity between the two representations is measured using the inner product. The modified version of DPR is called mDPR since it is initialised using the mBert model instead of the monolingual English BERT model.

The benchmark used is `Mr. TYDI` Zhang et al. (2021) which is a multilingual retrieval benchmark. It is partially derived from Wikipedia and covers languages that are typologically diverse.

		Ar	Bn	En	Fi	Id	Ja	Ko	Ru	Sw	Te	Th	Avg
(1)	BM25 (default)	0.368	0.418	0.140	0.284	0.376	0.211	0.285	0.313	0.389	0.343	0.401	0.321
(2)	BM25 (tuned)	0.367	0.413	0.151	0.288	0.382	0.217	0.281	0.329	0.396	0.424	0.417	0.333
(3)	mDPR (NQ pFT)	0.291	0.291	0.291	0.206	0.271	0.213	0.235	0.283	0.189	0.111	0.172	0.226
(4)	mDPR (MS pFT)	0.444	0.383	0.315	0.306	0.378	0.314	0.297	0.337	0.369	0.363	0.282	0.344
(5)	mDPR (MS pFT + in-lang FT)	0.691	0.651	0.489	0.551	0.562	0.488	0.453	0.485	0.640	0.876	0.619	0.591
(6)	mDPR (MS pFT + all FT)	0.695	0.623	0.492	0.560	0.579	0.501	0.487	0.517	0.644	0.891	0.617	0.600
(7)	mDPR (MS pFT + in-script FT)	–	–	0.473	0.555	0.563	–	–	–	0.635	–	–	–
(8)	mDPR (MS pFT + out-script FT)	–	–	0.476	0.563	0.565	–	–	–	0.644	–	–	
(9)	mDPR (in-lang FT)	0.678	0.638	0.418	0.516	0.544	0.447	0.383	0.448	0.580	0.860	0.597	0.555
(10)	mDPR (all FT)	0.695	0.659	0.476	0.550	0.565	0.496	0.453	0.515	0.633	0.891	0.607	0.594
(11)	mDPR (in-script FT)	–	–	0.444	0.535	0.560	–	–	–	0.622	–	–	–
(12)	mDPR (out-script FT)	–	–	0.457	0.543	0.573	–	–	–	0.624	–	–	
(13)	BM25 + row (6)	0.714	0.702	0.520	0.590	0.634	0.558	0.523	0.590	0.623	0.845	0.697	0.636
(a)	mono-ling DPR (in-lang FT)	0.678	–	0.426	0.573	0.545	–	0.476	–	–	–	–	–
(b)	mono-ling DPR (in-script FT)	–	–	0.412	0.540	0.488	–	–	–	–	–	–	–
(c)	mono-ling DPR (out-script FT)	0.682	–	0.426	0.522	0.540	–	0.454	–	–	–	–	–
(d)	mono-ling DPR (all FT)	0.682	–	0.448	0.540	0.533	–	0.454	–	–	–	–	–
(e)	English DPR (in-lang FT)	0.578	0.261	0.426	0.385	0.396	0.084	0.011	0.291	0.447	0.001	0.007	0.262
(f)	English DPR (MS pFT + in-lang FT)	0.592	0.318	0.497	0.423	0.439	0.218	0.182	0.298	0.499	0.001	0.030	0.318
(g)	AfriBERTa DPR (in-lang FT)	0.442	0.186	0.236	0.321	0.355	0.220	0.140	0.094	0.465	0.548	0.263	0.297
(h)	BM25 + row (f)	0.628	0.480	0.501	0.480	0.510	0.333	0.299	0.440	0.535	0.423	0.424	0.459

(a) MRR@100 on test set

Figure 3: Zhang et al. (2022), NQ: Natural Questions, pFT: pre-finetuned, MS: MS MACRO, all FT: finetune on all languages

Zhang et al. (2022) list various scenarios. For a target language, given the fact that the multilingual model that is used is mBert Devlin et al. (2019), there can be three major settings.

Assuming that the target language is L,

1. If $L \in mBert$ but there is no data available for the target language, the authors recommend that mDPR is pre-fine tuned on the MS MARCO passage dataset (which is in English), Bajaj et al. (2016), followed by retrieving passages in the target language (similar to a zero shot setting). The results are presented in figure 3. Rows 1,2,3 are the baselines against which the rest of the experiments are compared.

   The recommendation provided earlier is supported by the experimental results presented in row 4, which are superior to the baseline shown in row 3. According to the authors, the improved performance is likely attributed to the larger size of the MS MARCO dataset, despite the fact that the NQ dataset is more alike to `Mr. TYDI`.

   For Telugu, the score is 0.363(row 4) as compared to 0.111 in row 3, which is of course better as expected.

2. Next, we have a situation where the target language is present in mBERT and data from the target language is available for pre-training. This setup can be perceived in two ways:

   (a) The regular case, $L \in mBert$ and data is present in L. The authors of Zhang et al. (2022) recommend that mDPR is pre-finetuned on MS MARCO (similar to scenario 1) and then again finetuned on the data in the target language L. The authors assert that the model is exposed to cross-lingual learning this way. This recommendation can be supported by the results shown in figure 3, row 5 and can be compared with row 9. The difference between these experiments is that row 5 includes pre-finetraining on an English dataset while row 9 doesn't. The results of row 5 are, on average, better than row 9.

   Specifically for Telugu, the re-

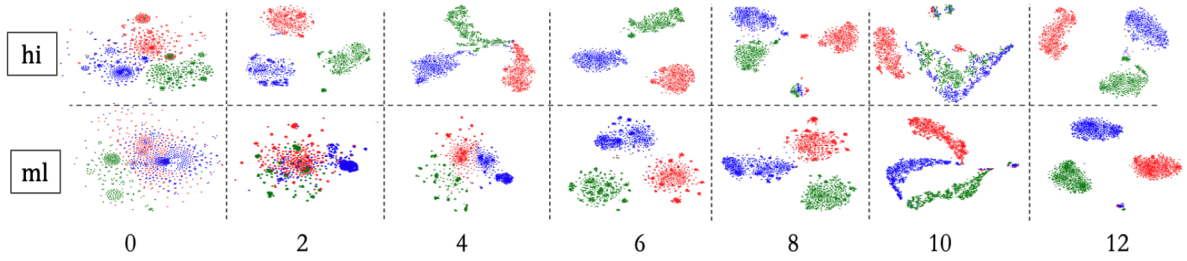

Figure 4: Krishnan et al. (2022), plot for XLM-R on Hindi(top) and m-BERT on Malayalam(bottom).

sults are promising, MRR@100 is 0.876(row 5) as compared to a score of 0.111 on the baseline mDPR model. Telugu has the same score for experiments when mDPR was trained on relevance judgements from all languages as well as the scenario when mDPR is first pre-finetuned on the MS MARCO dataset and then finetuned on data from all languages.

As alluded to in Luan et al. (2021), hybrid(dense+sparse) retrieval models, on average, perform better than non-hybrid varieties. This can be validated by the results presented in row 13 which are a combination of sparse(BM25) and dense(row 6) methods.

Overall, the authors conclude that *all models seem to perform better when finetuned on all language data, the effects seem to be stronger without pre-finetuning*

(b) The second case is where $L \in mBert$ but no data is available in L. However, data is available in another language K where, $K \in mBert$. K may or may not be similar to L. The authors of Zhang et al. (2022), in this case, advise that mDPR is pre-finetuned on MS MARCO and then further finetuned on K. Based on empirical observations over combinations of pairs of languages in the `Mr. TYDI` dataset, it is concluded that there seems to be a "perfect" language K, pre-training on which gives good results for L. Additionally, the authors conclude that, in general, finetuning on K (after having already pre-finetuned) does not harm the accuracy.

3. $L \notin mBert$ but data in language L is present. The authors recommend that DPR trained on English should be finetuned on data in L. The reasoning provided by the authors of Zhang et al. (2022) is that "something is better than nothing". The results for this experiment are shown in row (e), figure 3. The results for most languages are quite bad, but the results for Thai and Telugu are atrocious. The results improve slightly for most languages(again, except for Thai and Telugu) when MS MARCO pre-finetuning is added.

   The results for Telugu only improve when DPR is trained on AfriBERTa Ogueji et al. (2021), which is a mBERT style model, trained on African languages as is evident by the name. The reason provided by the authors for using this model is that, the idea was to use a model that does not share any languages with `Mr. TYDI`, except for Swahili. The assumption is that this is an attempt at exposing the model to a wider linguistic diversity to enable better cross-lingual learning. This does improve the result for Telugu in particular. The result is not comparable to the previous two scenarios (scenario 1, 2) but it is better than the model based on English DPR(row e, f).

The results presented above for Telugu do seem promising, and it would be interesting to verify if these results pan across other Dravidian Languages like Malayalam, Kannada and Tamil.

4 Transliteration for Dravidian languages

Transliteration is the process of swapping text/syllables from one script to another while maintaining the phonetic pronunciation in the original script. It is a very common phenomenon on social media to have people communicate using the romanized version of the script instead of using the original script. From personal experience, this is due to the fact that it is much harder to type in the original script as compared to the Latin script which is ubiquitous. Hence, research in this area is essential as well. Krishnan et al. (2022) approach this topic in the context of Malayalam and Hindi transliterated text.

A three-way t-SNE plot of embeddings across the attention layers show that there is almost no overlap between the sentences written in the original script, the English translation and the romanized version, even through they all semantically mean the same. This is demonstrated by figure 4.

(Krishnan et al., 2022) address this problem of non-alignment in multilingual models using a combination of data augmentation techniques and a teacher-student method. We need a data augmentation technique, since there does not exist training data in the transliterated target.

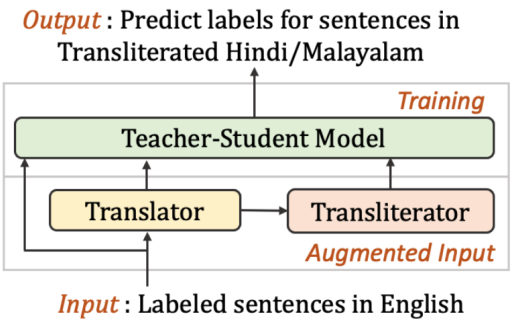

Figure 5: Krishnan et al. (2022)

The methodology can be described with an example, given a sentence S in the source language(English), it is translated to the target language(in its original script), Malayalam, and named T. Then, T is transliterated to the script of the source language. This way, we get the training data for the romanized sentences. A teacher-student method can then be used to learn an alignment between these two representations. An overview of the method is described in figure 5.

The authors also curate two datasets using this method, one, a binary sentiment(positive and negative) Malayalam movie reviews dataset, and the second one is a Hindi dataset that classifies tweets(from natural disaster crises) into relevant and non-relevant.

Figure 6 shows the classification performance on transliterated datasets using mBERT and XLM-R. The scores show that the performance of both models is significantly boosted when transliterated data is included in the training dataset. Specifically, there is a total increase of +5.6 % on mBERT and +4.7% on XLM-R in weighted F1 performance across the four romanized datasets. This demonstrates the importance of adding augmented and transliterated data in multi-lingual to improve their accuracy and effectiveness.

The top baseline models for both mBERT and XLM-R in this study are `mBERTen+tr+tl` and `XLM-Ren+tr+tl`. These models use a combination of augmented data with the original data to fine-tune the model for the classification task.

The authors of Krishnan et al. (2022) postulate that the improvement produced by their model could be due to the fact that mBERT and XLM-R may not have seen many transliterations in these languages before. mBERT is trained using data from Wikipedia, while XLM-R uses Common Crawl. As such, XLM-R is likely to have been trained on at least some code switched(data that "switches" between two languages seamlessly) or transliterated data, which could explain its better performance in this study.

The practical significance of the model above is evident from its application on translit-

Test Data → Models ↓	hi_{ro}		ml_{ro}		hi_{nf}		ml_{kf}		AVG	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Monolingual LM$^{\diamondsuit}$	50.38	37.99	48.76	47.55	54.39	49.06	63.35	63.33	54.22	49.48
mBERT Baselines										
mBERT_{en}	47.55	41.23	48.23	41.63	56.67	55.27	57.84	57.72	52.57	48.96
mBERT_{tr}	51.81	48.21	51.32	45.62	52.89	47.89	58.26	57.54	53.57	49.82
mBERT_{tl}	55.29	54.18	61.72	56.47	56.14	55.78	58.09	57.79	57.81	56.06
$\text{mBERT}_{en+tr+tl}$	55.16	54.75	61.72	61.25	56.71	56.03	**63.74**	**63.42**	59.33	58.86
Our mBERT models										
mBERT-Joint	55.57	55.56	64.29	63.58	57.75	56.73	51.85	53.98	57.37	57.46
mBERT-Joint-TS	**57.37**$^{\spadesuit}$	**57.36**$^{\spadesuit}$	**65.15**$^{\spadesuit}$	**65.78**$^{\spadesuit}$	**63.22**$^{\spadesuit}$	**63.14**$^{\spadesuit}$	62.55	62.40	**62.07**	**62.17**
XLM-R Baselines										
XLM-R_{en}	50.57	45.76	50.86	47.13	58.11	56.77	61.74	60.98	55.32	52.66
XLM-R_{tr}	49.52	47.67	51.72	50.16	57.11	56.95	61.74	61.72	55.02	54.13
XLM-R_{tl}	54.81	53.72	51.67	54.71	51.45	51.24	59.84	59.23	54.44	54.73
$\text{XLM-R}_{en+tr+tl}$	55.57	54.19	62.46	61.52	56.40	55.67	63.24	63.10	59.42	58.62
Our XLM-R Models										
XLM-R-Joint	56.09	55.40	62.90	63.14	53.68	52.81	62.79	62.77	58.87	58.53
XLM-R-Joint-TS	**57.70**$^{\spadesuit}$	**57.03**$^{\spadesuit}$	**65.93**$^{\spadesuit}$	**65.71**$^{\spadesuit}$	**58.39**	**57.86**	**64.87**$^{\spadesuit}$	**64.87**$^{\spadesuit}$	**61.72**	**61.37**

Figure 6: Krishnan et al. (2022)

erated datasets of tweets that were published during the North India and Kerala flood crises. A model that generates embeddings in a comparable space as English tweets and can handle transliterated tweets quickly has the potential to be incredibly beneficial for emergency service information systems. It could make use of a broad range of English-trained crisis response models, improving the efficacy and accuracy of these systems.

5 Detecting offensive speech in Dravidian Languages

The increasing usage of social media for information dissemination has the potential of being a positive force, for example, in their use during natural disasters, movements for social good and so on. However, the same can also be used to spread negative and defamatory content against certain groups, specifically targeting them due to their differences. This content can be homophobic, trans-phobic, against a religion, sex or nationality. Controlling/managing/ this speech on any social media platform is the need of the hour.

The tools for regulating offensive speech in English and other high-resource language are available and quite advanced due to the abundance of high quality data. It is difficult to find the same for low-resource languages and situations, like code-switched data or romanized data.

Hande et al. (2021) conduct experiments and propose an approach to deal with the problem of lack of data in under-resourced languages by performing a type of data augmentation. The technique used by the authors comprises three essential components.

Firstly, they expand the dataset using data augmentation techniques. The datasets used to demonstrate the augmentation method are code-mixed comments from YouTube in Tamil, Malayalam and Kannada. The comments are regarding movie reviews. Each review is divided broadly into "not offensive" and "offensive" category. Psuedo-labels are generated for the transliterated data after which it is combined with the original data. This creates a bigger training dataset which attempts to solve the problem of the lack of data.

The effectiveness of this method is evaluated by conducting experiments on multiple multilingual large language models like mBERT, XLM-R (Conneau et al., 2020), DistilmBERT (Sanh et al., 2019), IndicBERT(BERT trained on Indic languages) (Siddhant et al., 2020), ULMFiT (Howard and Ruder, 2018), MuRIL (Khanuja et al., 2021). The experiments are conducted on primary, transliterated and com-

	CM-TRA								
	Malayalam			Tamil			Kannada		
	P	R	F1	P	R	F1	P	R	F1
mBERT	0.9468	0.9535	0.9478	0.6865	0.7432	0.7026	0.6048	0.6517	0.6188
XLM-R	0.9370	0.9410	0.9366	0.7284	0.7609	0.7427	0.6997	0.7455	0.7029
DistilmBERT	0.9582	0.9575	0.9537	0.7414	0.7516	0.7461	0.7008	0.7198	0.7037
MuRIL	0.7780	0.8821	0.8268	0.7081	0.7511	0.7045	0.6407	0.7249	0.6801
IndicBERT	0.9306	0.9465	0.9380	0.6867	0.7516	0.7057	0.5937	0.6671	0.6235
ULMFiT	**0.9649**	**0.9610**	**0.9624**	**0.8203**	**0.7719**	**0.7934**	**0.7576**	**0.7104**	**0.7306**

Figure 7: Hande et al. (2021) CM-TRA is the augmented dataset.

bined datasets. The authors wanted to check if the merged dataset performed better than the original one. The result corroborate this, since the weighted F1 scores for all three languages(Tamil, Kannada and Malayalam) are higher than the baseline model.

Figure 7 shows the results for the augmented dataset on different models. ULMFiT performs the best in terms of the F1 score across all three languages. Figure 8 in the Appendix shows all results.

6 Conclusion

This report focuses on dealing with some practical problems that are encountered while dealing with low-resource regional languages, such as code-switching and transliteration. Other major tasks such as dependency parsing (Tran and Bisazza, 2019), sentiment analysis (Puranik et al., 2021) have research in the major Dravidian languages. However, it is difficult to find research/data for other popular languages such as Tulu(spoken by nearly two million), Gondi(2 million speakers) and so on. Further, the major languages belong to the sub-categories of South-Central(Telugu) and South(Malayalam, Tamil) while Brahui is a language that belongs to the Northern subcategory and is the only Dravidian language spoken in Pakistan which makes it unique. Future work could be in this direction and cross-lingual learning, according to the studies above is a good direction to follow.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Encyclopædia Britannica. 2023. Dravidian languages: distribution. https://www.britannica.com/topic/Dravidian-languages#/media/1/171083/95886. [Online; accessed 09-April-2023].

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Adeep Hande, Karthik Puranik, Konthala Yasaswini, Ruba Priyadharshini, Sajeetha Thavareesan, Anbukkarasi Sampath, Kogilavani Shanmugavadivel, Durairaj Thenmozhi, and Bharathi Raja Chakravarthi. 2021. Offensive language identification in low-resourced code-mixed dravidian languages using pseudo-labeling. *arXiv preprint arXiv:2108.12177*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha

Talukdar. 2021. Muril: Multilingual representations for indian languages.

Jitin Krishnan, Antonios Anastasopoulos, Hemant Purohit, and Huzefa Rangwala. 2022. Cross-lingual text classification of transliterated hindi and malayalam. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 1850–1857. IEEE.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Karthik Puranik, Bharathi B, and Senthil Kumar B. 2021. Iiitt@dravidian-codemix-fire2021: Transliterate or translate? sentiment analysis of code-mixed text in dravidian languages.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4364–4374.

Aditya Siddhant, Danish Aggarwal, Chaitanya Maheshwari, Manvi Kaul, Maitri Dubey, Nishtha Nain, Mahak Gambhir, and Rahul Gupta. 2020. Indicbert: A multilingual language model for indian languages. In *Proceedings of the 1st Workshop on Technologies for MT of Low Resource Languages*, pages 46–52.

Karen Sparck Jones and Peter Willet. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Felix Stahlberg. 2020. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418.

Ke Tran and Arianna Bisazza. 2019. Zero-shot dependency parsing with pre-trained multilingual sentence representations.

Wikipedia. 2023. Dravidian languages — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Dravidian%20languages&oldid=1148834674`. [Online; accessed 09-April-2023].

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2019a. Scalable zero-shot entity linking with dense entity retrieval.

Shijie Wu, Alexis Conneau, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Emerging cross-lingual structure in pretrained language models. *arXiv preprint arXiv:1911.01464*.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. tydi: A multi-lingual benchmark for dense retrieval. *arXiv preprint arXiv:2108.08787*.

Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. 2022. Towards best practices for training multilingual dense retrieval models. *arXiv preprint arXiv:2204.02363*.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*.

7 Appendix

Model	Code-Mixed Dataset								
	Malayalam			Tamil			Kannada		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
mBERT	0.9195	0.9410	0.9301	0.7461	0.7664	0.7556	0.6863	0.7082	0.6936
XLM-R	0.9206	0.9380	0.9288	0.5275	0.7263	0.6112	0.6449	0.7326	0.6851
DistilmBERT	0.9411	0.9520	0.9465	0.7368	0.7632	0.7489	0.6789	0.7249	0.7010
MURiL	0.7780	0.8821	0.8268	0.5275	0.7263	0.6112	0.3012	0.5488	0.3890
IndicBERT	0.9572	0.9600	0.9568	0.7150	0.7454	0.7287	0.6714	0.6992	0.6809
ULMFiT	0.9643	0.9580	0.9603	0.8220	0.7650	0.7895	0.7186	0.6864	0.7000
	Transliterated Dataset								
	Malayalam			Tamil			Kannada		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
mBERT	0.9023	0.9398	0.9202	0.7063	0.7648	0.7286	0.6779	0.7389	0.7002
XLM-R	0.8902	0.9265	0.9080	0.5538	0.7320	0.6290	0.6369	0.7198	0.6750
DistilmBERT	0.9089	0.9370	0.9199	0.7248	0.7571	0.7390	0.6789	0.7249	0.7010
MuRIL	0.9039	0.9405	0.9218	0.5275	0.7263	0.6112	0.6432	0.7249	0.6815
IndicBERT	0.9305	0.9445	0.9373	0.7194	0.7354	0.7263	0.6433	0.6722	0.6558
ULMFiT	0.9521	0.9505	0.9508	0.8033	0.7682	0.7842	0.7304	0.6979	0.7115
	CM-TRA								
	Malayalam			Tamil			Kannada		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
mBERT	0.9468	0.9535	0.9478	0.6865	0.7432	0.7026	0.6048	0.6517	0.6188
XLM-R	0.9370	0.9410	0.9366	0.7284	0.7609	0.7427	0.6997	0.7455	0.7029
DistilmBERT	0.9582	0.9575	0.9537	0.7414	0.7516	0.7461	0.7008	0.7198	0.7037
MuRIL	0.7780	0.8821	0.8268	0.7081	0.7511	0.7045	0.6407	0.7249	0.6801
IndicBERT	0.9306	0.9465	0.9380	0.6867	0.7516	0.7057	0.5937	0.6671	0.6235
ULMFiT	**0.9649**	**0.9610**	**0.9624**	**0.8203**	**0.7719**	**0.7934**	**0.7576**	**0.7104**	**0.7306**

Figure 8: Hande et al. (2021) Code-mixed dataset is the original dataset, Transliterated is the modified version and CM-TRA is the augmented version which is the combination of code-mixed and transliterated