My main research goal is to overcome the inherent limitations of current neural models on low-resourced and unseen languages. I am particularly interested in 1) incorporating efficient cross-lingual transfer learning capabilities, 2) exploring the effectiveness of reasoning on resource-limited settings, and 3) mitigating the drawbacks of cross-lingual transfer learning.

Integrating Efficient Meta-Learning To evaluate the comprehending capabilities of generalized pre-trained language models, we considered pre-trained transformer-based multilingual BERT (mBERT) as the baseline models for our sentiment analysis and emotion detection task datasets during my undergraduate thesis. Our sentiment analysis dataset, *SentNoB* [1], is a *single-label* classification dataset with each instance corresponding to one of *three* labels (positive, negative, neutral), whereas our emotion detection dataset, *EmoNoBa* [2], is a *multi-label* dataset with each instance corresponding to any of *six* labels (love, happy, surprise, sad, anger fear), with instances written in Bangla, a low resource language with over 100 dialects[1]. To measure the generalization toughness of the dataset, we considered the metric of the unique word (UW) percentage; a higher UW percentage refers to a more diverse dataset and vice versa. We identified that our sentiment analysis dataset has *12.91%* UW and our emotion detection has *18.81%* UW, in contrast to the 5.09% UW of the popular English language sentiment dataset IMDB review corpus [3]. Both of our datasets also attained a *moderate* level of Fleiss' kappa [4] inter-annotator agreement (IAA).

During our model evaluation, I found that our mBERT model, fine-tuned on the final output layer, performed worse than hand-crafted feature-based models on sentiment analysis tasks while for emotion detection tasks, it performed below random baselines. This was intuitive as the complexity, label size, and UW percentages of the emotion detection dataset increased in relation to the sentiment analysis dataset, leading to a decrease in performance.

One popular approach to alleviate the consequent limitation is to utilize existing resources developed in this language, also known as transfer learning. However, as we were the first to publicly release such a dataset, cross-lingual transfer learning is the obvious next step which is to learn from this task already available in other resource-rich languages. However, the translation dependency raises the concern of whether they can effectively bridge the contextualized representation gap between the resource-rich auxiliary language and the resource-scarce target language. While recent meta-learning methods have been introduced to mitigate these differences for cross-lingual transfer learning, there is a risk that these compute-intensive approaches could become even more compute-intensive when mitigating the contextualization gap between the source and target languages that do not belong to the same language family. Moreover, identifying only one source language within the language family that has sufficient resources to meet the target language's learning requirement is not always achievable. **(RQ1)** Therefore, my research question is can we consider multiple closely related languages as the source languages for more efficient and effective cross-lingual transfer?

Scope of Reasoning on Low-Resource Setting Both of our datasets consisted of instances written by social media users. Therefore, we employed active social media users to label our dataset. For our sentiment analysis task, we provided annotators with chunks of the dataset and instructed them to annotate on a scale of 1-5, with 1 indicating *Strongly Negative* sentiment, 3 *Neutral* sentiment, and 5 *Strongly Positive* sentiment. This was mainly to determine how often two annotators were at opposite extremes of the range and to consult with those who were struggling with the task before starting the next chunk of annotation. We also employed a similar approach for the emotion detection task and found very few such occurrences, indicat-

[1] `https://en.wikipedia.org/wiki/Bengali_language`

ing all annotators were contextually aligned. However, despite having *moderate* inter-annotator agreement, non-contextual hand-crafted lexical feature-based models dominated the top standings in the performance chart on both tasks against the likes of Bi-LSTM, and mBERT models.

One possible direction is to incorporate reasoning capabilities on the neural models by allowing annotators to describe their thought process on how they arrived at their annotation, also known as chain-of-thought (CoT). **(RQ2)** However, with a handful of work conducted on Natural Language Understanding (NLU) and Natural Language Generation (NLG) on low-resource languages, will the model understand and generate corresponding CoT for the resource-scarce language tasks?

Overcoming Cross-Regional Stereotypes While extrapolating the strongest feature n-grams of each label on both tasks, we observed certain types of stereotypes associated with some words. For instance, *Police*, *Monkey*, *Son of Dog* (also slang in this region) found themselves on negative sentiment polarities.

With incorporating cross-lingual transfer learning methods, concerns about these types of stereotypes getting injected into the new language resource will be evident. **(RQ3)** Therefore, can we explore methods beyond transfer-learning that will reduce stereotypical concepts of one culture getting injected into the other, and thus restraining cultures from (considering Bangla language dataset was used as a source language for cross-lingual transfer) hating *Police*?

References

[1] **Khondoker Ittehadul Islam**, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. "SentNoB: A Dataset for Analysing Sentiment on Noisy Bangla Texts". In: *Findings of EMNLP 2021*.

[2] **Khondoker Ittehadul Islam**, Tanvir Yuvraz, Md Saiful Islam, and Enamul Hassan. "EmoNoBa: A Dataset for Analyzing Fine-Grained Emotions on Noisy Bangla Texts". In: *AACL-IJCNLP 2022*.

[3] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. "Learning word vectors for sentiment analysis". In: *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*. 2011.

[4] Joseph L Fleiss. "Measuring nominal scale agreement among many raters." In: *Psychological bulletin* 76.5 (1971), p. 378.