## Building a better Bangla Part-of-Speech Tagger

## SAVANT Project

## April - June 2023

Core Commitments as of 5th April 2023

* Identify Bangla derivational affixes (*Thompson, H. R. (2020). Bengali: A comprehensive grammar. Routledge.*)
* Stemming : Rule-based-stemmer (append with derivational affixes)
* Cleaning and normalizing Raw Corpus
* Applying the stemming rules on a Bangla Corpora
* Validation : Stemmer Output
* Curating a sample dataset for POS-Tagging model learning
* Deploying the model on a larger corpora
* Analyze/Validate/Tweak the model to enhance accuracy
* Manual correction of the output
* Corpus Statistics : Stem freq, whole word freq, Transitional Prob etc