# Project Outline : Team 2

iNLP | Spring 2024

SaTaN

Sanika Damle | Tanveer Ul Mustafa | Nanda Rajiv

## Problem Statement & Scope

### Problem Statement

1. We aim to implement contextual word embeddings using ELMO on Indian Languages (Hindi, Tamil, Marathi).

2. Algorithms for getting contextual embeddings are available, but there is a lack of evaluation metrics that are specifically designed for word representations in Indian languages.

### Scope

1. Implementation of ELMo: Adapting existing architectures and training the model on a cleaned corpus of text data in the chosen languages. Preprocessing steps like tokenization will be needed.

2. Evaluation Metrics: The evaluation metrics that we are using are word and semantic textual similarity, word analogy, sentiment analysis, named entity recognition, natural language inference and question answering. There are some languages for which datasets for all the tasks are not available, so we will not be computing that particular metric for that language.

3. Performance Analysis: The embeddings for each language will be evaluated and compared using the above metrics. We will log scores of each task and language.

## Datasets

### ELMo

These are the datasets that we will use for training the word embeddings. They are corpuses of sentences in each language.

1. Hindi : Corpus

2. Marathi : Corpus

3. Tamil : Corpus

## Word Similarity

To measure world similarity using ELMo, the context is needed. We datasets that have pairs of similar words. We will manually use these words in a couple of sentences and check the cosine similarity of the word embeddings in each case. The dataset just contains the pairs.

1. Hindi : Similar word pairs
2. Marathi : Similar word pairs
3. Tamil : Similar word pairs

## Semantic Textual Similarity

This is a dataset that has pairs of sentences in Hindi and Marathi. We were unable to find such a dataset for Tamil. We may be able to generate a few test cases for Tamil (by changing the voice, paraphrasing other sentences, etc). The original dataset has been formed by google translating an English STS dataset, which we may be able to extend to Tamil.

1. Hindi : STS dataset
2. Marathi : STS dataset

## Word Analogy

For this task, we will just consider Hindi. The data is a combination of 4 words in the form A:B::C:D. This indicates a relation between A and B which is repeated in the words C and D.

1. Hindi : Tuples of 4 words

## Sentiment Analysis

These datasets contain sentences along with their tones, that is, whether the sentences have a positive/negative/neutral/ambiguous tone for Hindi, Marathi and Tamil.

1. Hindi : Sentiment Analysis dataset (Tweets)
   Sentiment Analysis dataset (Movie Reviews)
2. Marathi : Sentiment Analysis dataset
3. Tamil : Sentiment Analysis dataset (Tweets)

## Named Entity Recognition

Named entity recognition is locating and classifying named entities in unstructured text into predefined categories like person names, organizations, etc. We will test how accurate this is with contextual embeddings.

1. Hindi : NER Dataset
2. Marathi : NER Dataset
3. Tamil : NER Dataset

### Natural Language Inference

This involves determining the relationship betweeen two pieces of text, a context and a hypothesis. If the hypothesis can be inferred from the context, it is marked as entailed, else it is non entailed. Only datasets for Hindi could be found for this task.

1. Hindi : NLI Dataset
2. Marathi : NLI Dataset

### Multiple Choice Question Answering

This has a hindi and marathi dataset for evaluating the question answering.

1. Hindi : Question Answers
2. Marathi : Question Answers

## Literature Review

Word embeddings are required for nearly all, if not all, NLP tasks. They enable passing words through neural language models, allowing for complex tasks such as prediction, classification, tagging, and many others.
These embeddings should be able to model both the complex characteristics of word use, as well as model how the same 'word' has different meaning in different linguistic contexts (known as polysemy) (Peters et al., 2018).

Peters et al. (2018) introduce a method for *deep contextualized word embeddings* which are learned functions of the internal states of a deep bidirectional language model pretrained on a large textual corpus. This is also known as ELMo (short for Embeddings from Language Models).

Prior alternates to contextual word embeddings were fixed word embeddings, which may or may not have been trained in context, but are used as is in inference. ELMo differs from this, because the embedding of the same word in a certain sentence, would differ from that in a different sentence. This, however, does not capture polysemy, which is the possibility of the same word to mean two different things. Another phenomenon which is homonymy, which is when two different words with the same sound or spelling have different meanings. These are not captured by fixed word embeddings. Other approaches which are indeed context-based do not generalise well to a range of NLP tasks, like ELMo does (Peters et al., 2018).

There are several potential metrics to measure the goodness of the embeddings. One of these methods in the literature is using semantic textual similarity, employed by Joshi et al. (2022). The dataset in the paper contains a similarity score, which can be used to test the implementation of the embeddings, based on the consistency of the score, as well as consistency with the labelled dataset.

Apart from this, word analogy would be a metric to measure the goodness of the embeddings model, and having analogies such as $\overrightarrow{master} - \overrightarrow{male} + \overrightarrow{female} =$

$\overrightarrow{mistress}$. This can be done even with an analogue of ’ratios and proportions’ for words. By comparing against an existing dataset provided by AI4Bharat (n.d.) of sets of four words with are such that A:B::C:D, the goodness of the embeddings can be measured.

Other evaluation metrics specifically for Indic lamguages, which are a combination of sentiment analysis, natural language inference, name entity recognition, MCQA, etc., can be employed to evaluate embeddings and many other NLP tasks centred around Indian languages.

# Implementation Timeline

1. Train ELMo for Hindi, Marathi, and Tamil (20th March)
2. Implement the word analogy evaluation metric (24th March)
3. Implement the Word Similarity and Semantic Textual Similarity metrics (10th April)
4. Implement the Sentiment Analysis and the Named Entity Recognition parts of the IndicGLUE metrics (14th April)
5. Implement the Natural Language Inference and Multiple Choice Question Answering parts of the IndicGLUE metrics (20th April)
6. Compilation of results and making of the report (24th April)

# References

1. Peters, M. E., Neumann, M. E., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). https://doi.org/10.18653/v1/n18-1202
2. Grave, É., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1802.06893
3. Mujadia, V., Agarwal, D., Mamidi, R., & Misra Sharma, D. (2022). Semantic textual similarity for Hindi. www.academia.edu/71473194/Semantic_Textual_Similarity_For_Hindi
4. Joshi, A., Kajale, A., Gadre, J., Deode, S., & Joshi, R. (2022). L3Cube-MahaSBERT and HindSBERT: Sentence BERT Models and Benchmarking BERT Sentence Representations for Hindi and Marathi. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2211.11187
5. indic_glue · Datasets at Hugging Face. (n.d.). https://huggingface.co/datasets/indic_glue