# CS 689

# COMPUTATIONAL LINGUISTICS FOR INDIAN LANGUAGES

# PRIYANSHU KUMAR JHA

# 231110039

**SOLUTION 6:**

LEARNINGS FROM OBSERVATIONS:

1. **Effect of Vocabulary Size:**

   - Generally, increasing the vocabulary size tends to improve the precision, recall, F1-score, and accuracy of the model. This is evident from the comparison between UNIGRAM vocab size 1000 and 2000, where all metrics show improvement with a larger vocabulary.

   - However, the improvement may not be substantial, as seen in the comparison between BPE vocab size 1000 and 2000, where the metrics only slightly increase.

   - This suggests that having a larger vocabulary helps the model capture more diverse and nuanced tokens, leading to better performance.

2. **Effect of Maximum Sequence Length**:

   - The maximum sequence length appears to have little to no impact on the performance metrics, as evidenced by the consistent precision, recall, F1-score, and accuracy across different maximum sequence lengths for both mBERT and IndicBERT models.

   - This suggests that the tokenization method and vocabulary size may have a more significant influence on model performance compared to the maximum sequence length.

3. **Model Performance Comparison**:

   - Comparing the performance of mBERT and IndicBERT models, it's evident that mBERT consistently outperforms IndicBERT across all metrics.

   - IndicBERT shows notably lower precision, recall, F1-score, and accuracy compared to mBERT, indicating that mBERT, which is pre-trained on a multilingual corpus including Indic languages, better captures the nuances of the text.

4**. General Trend**:

   - The general trend observed is that larger vocabulary sizes tend to lead to better model performance, particularly in terms of precision, recall, F1-score, and accuracy.

   - Additionally, the choice of model architecture, such as mBERT versus IndicBERT, plays a significant role in determining performance, with mBERT generally outperforming IndicBERT.

On increasing the vocabulary size tends to improve model performance, the choice of model architecture also significantly influences the results. Additionally, the maximum sequence length appears to have minimal impact on performance in this context.