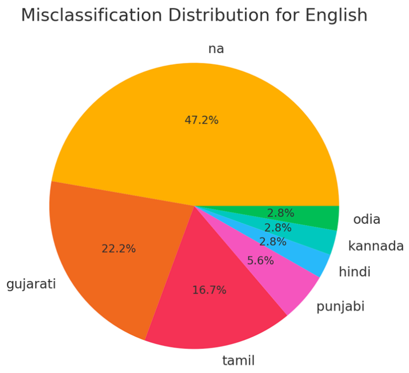
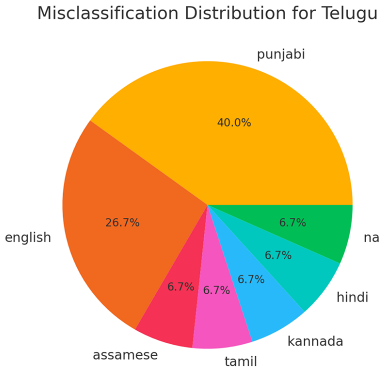
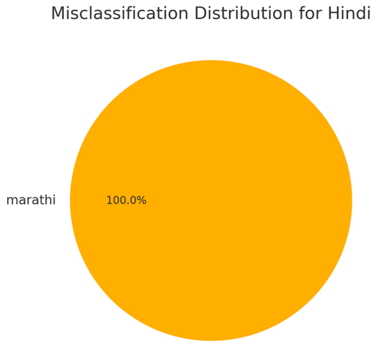
REPORT ON SCENE TEXT DATA COLLECTION & EVALUATION USING IndicPhotoOCR

Report by:
B24BB1040, B24CM1037,
B24CM1034,B24CM1084

# Analysis of Misclassification of *LANGUAGES* (English, Telugu, Hindi)


Misclassification Distribution for English


Misclassification Distribution for Telugu


Misclassification Distribution for Hindi

The pie charts illustrate the misclassification patterns of English, Hindi, and Telugu images by the OCR model, showing how each language is incorrectly predicted as other scripts and revealing the overall distribution of language recognition errors.

# Word recognition findings' analysis:

Language	Total samples	Correct recognitions (language)	Misclassifications (language)	Language recognition %	Correct word recognitions	Word accuracy	Precision (language) %	Recall (language)%
English	256	220	36	85.94%	205	80.08%	98.21%	85.94 %
Hindi	12	6	6	50.00%	11	91.67%	75.00%	50.00 %
Telugu	91	76	15	83.52%	53	58.24%	100.00%	83.52 %

The combined performance table shows that while English achieved high language recognition and strong word accuracy, its errors mainly stem from occasional confusion with Gujarati, Tamil, and unrecognized cases (na). (common errors include: misreading numbers, symbols like '&')

Hindi, despite excellent word accuracy, exhibited low language recognition due to frequent misclassification as Marathi (visually similar script could be a reason)

Telugu displayed perfect precision with no false positives, but lower word accuracy, often partially spelling words, difficulty in similar looking letters (న , ష) and a common error of missing '*virama*' (◌్) for the ending letter of some words (hugely impacts how sound of the word ends). The model sometimes misidentifies Telugu text as Punjabi, English, or other scripts
# Special observations: Words with smaller, trickier fonts were more prone to misclassification. Language misclassifications almost always led to wrong recognition of the word's text.