LATIN-NASTALIQUE SCRIPT CLASSIFICATION SYSTEM
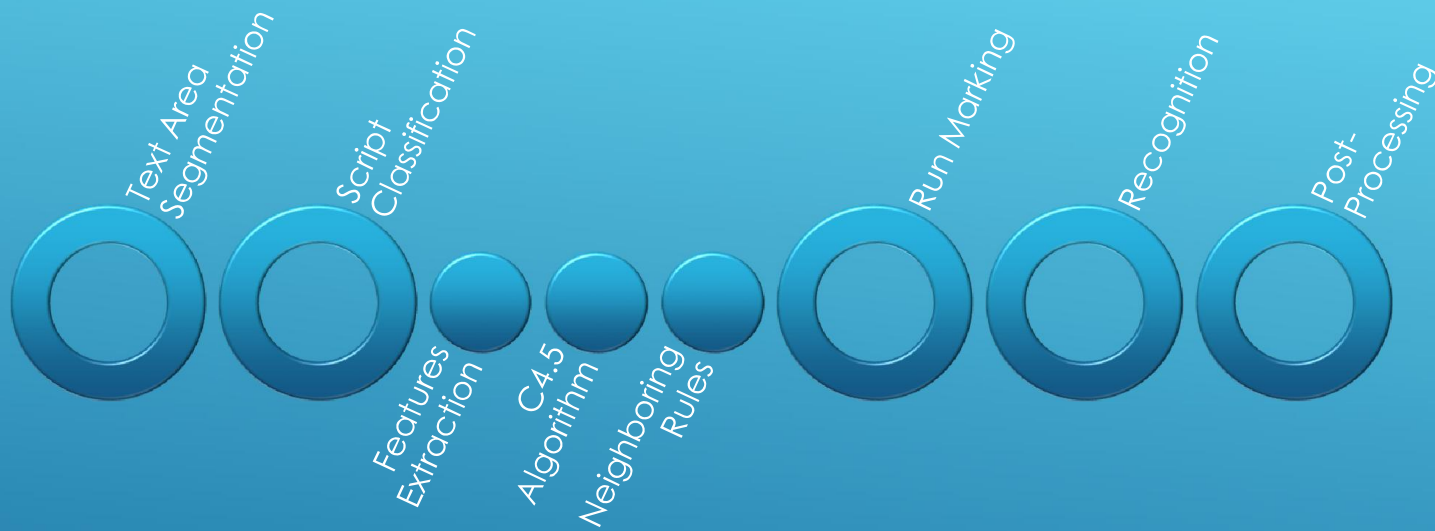
Presenter: Muhammad Usman Ghani

- Latin script is also used for terminology illustration or other purposes in Urdu books and Magazines.

- The script detection system isolates Nastalique and Latin script.

- The Nastalique script is recognized through Urdu OCR and Latin script is recognized by the Tesseract OCR.

- Font size independent approach is used.

مخصوص طرزِ تمدن اور بودوباش کی بنا پر ایک قطعی اور واضح تہذیب (sub-culture) کو جنم

INTRODUCTION

Latin-Nastalique Script Classification System

Text Area Segmentation · Script Classification · Features Extraction · C4.5 Algorithm · Neighboring Rules · Run Marking · Recognition · Post-Processing

SYSTEM OVERVIEW

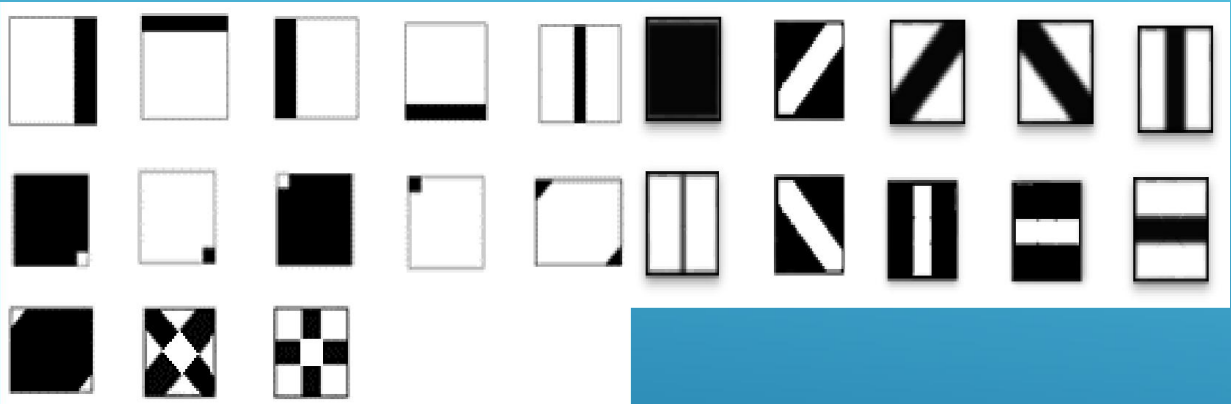15th July, 2014 | Center for Language Engineering (CLE)

- Features Extraction
    - Dimensional Features
    - Morphological Features
- Classification: C4.5 Decision Tree algorithm

SCRIPT CLASSIFICATION

- Dimensional Features
  - Height
  - Width
  - Area
  - Height-to-Width Ratio
  - Centroid Composite Value

FEATURES EXTRACTION (1)

- Morphological Features

FEATURES EXTRACTION (2)

▶ Script type of first ligature in a line is changed to script type of next two CCs, if these two CCs have same script type.

▶ Script type of last ligature in a line is changed to script type of previous two CCs, if these two CCs have same script type.

▶ If a ligature having script type Latin have Nastalique script CCs on its right and left, its script type would be changed to Nastalique.

▶ If a ligature having script type Nastalique have Latin script CCs on its right and left, its script type would be changed to Latin.

▶ If a Latin script ligature has a diacritic associated with it and it is placed below the MB or inside the MB, script type of such ligature would be converted to Latin.

NEIGHBORING RULES

99
Identity Crisis

(Collective Will
Nationality)

55
(Gallstones
(bile salts
cholesterol
calcium

RUN MARKING

- 99Identity Crisis

- (Collective WillNationality)

- 55(Gallstones(blle saltscholesterolcalcium£

RECOGNITION

99

Identity Crisis

(Collective Will

Nationality)

55

(Gallstones)

blle salts

Cholesterol

Calcium

£

POST-PROCESSING

QUESTIONS ?

Latin-Nastalique Script Classification System

THANK YOU ☺

15th July, 2014 | Center for Language Engineering (CLE)