Day 2: Challenges in Pashto & Urdu NLP

Sami Uddin Shinwari
Subject Specialist - IT | NLP Researcher
MS in Data Science (FAST NU Peshawar)
BS in Computer Science (UET Peshawar)

December 16, 2025

Why Pashto & Urdu are Challenging

- Pashto and Urdu are **low-resource languages**.
- Limited annotated datasets and pretrained models.
- Most NLP tools are designed for English.
- Impacts NER, MT, sentiment analysis, and summarization.

Script & Writing Challenges

- Arabic-based cursive scripts.
- Right-to-left (RTL) writing direction.
- No capitalization (NER becomes difficult).
- Multiple Unicode forms for same characters.

Urdu Example:
وی او اے

Tokenization & Segmentation Issues

- Whitespace tokenization often fails.
- Urdu problems:
  - Space omission: غلطالفاظ
  - Space insertion: الفاظ لط غ
- Pashto also suffers from inconsistent word boundaries.

Urdu Sentence Example:

یہ ایک جملہ ہے

Linguistic Challenges

- Rich morphology (prefixes, suffixes, inflections).
- Free or flexible word order.
- Strong dialectal variation.

Examples:

- Urdu verb forms:
  لکھنا □ لکھتا ہے، لکھ رہی تھی، لکھ چکا ہوں
- Pashto dialects:
  شمالی، جنوبی، مرکزی

Morphology & Word Variation

- Urdu words vary with gender, number, and tense.
- Pashto words inflect for case and agreement.
- Borrowings from Arabic and Persian increase variation.

Impact:

- Difficult stemming and lemmatization.
- One root appears in many surface forms.

Data & Tool Limitations

- Scarcity of labeled datasets.
- Few reliable tokenizers and POS taggers.
- Limited sentiment lexicons.
- English-trained models perform poorly.

Comparison with English NLP

- **English NLP**
  - Large datasets
  - Mature libraries
  - Strong pretrained models
- **Pashto & Urdu NLP**
  - Data scarcity
  - Custom preprocessing required
  - Language-specific solutions needed

Key Takeaways

- Pashto and Urdu require **custom NLP pipelines**.
- Data creation is as important as modeling.
- Research in these languages has high social impact.

#PashtoNLP #UrduNLP #LowResourceLanguages #NLP

References (APA)

- Ahmad, Z., & Hussain, S. (2007). Urdu corpus development. *Conference on Language and Technology*.
- Becker, M., & Riaz, K. (2012). Urdu morphology. *Journal of South Asian Linguistics, 4*(1), 23–45.
- Hardie, A. (2003). Corpus-based Urdu NLP. Lancaster University.
- Rahman, S., Iqbal, Z., & Khan, A. (2023). Pashto text processing challenges. *IJNLC, 12*(2), 55–70.