Day 6: Stemming & Lemmatization for Urdu & Pashto

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

January 10, 2026

What is Stemming & Lemmatization?

- **Stemming**: Removes suffixes mechanically.
- **Lemmatization**: Converts word to its dictionary form.
- English has strong tools — Urdu & Pashto do not.

Why Is It Hard for Urdu & Pashto?

- Rich morphology
- Gender, tense, number variations
- No standard morphological analyzers
- One root $\rightarrow$ many surface forms

Urdu Example

Root Verb:
لکھنا

Forms:
لکھتا ہے، لکھ رہی تھی، لکھ چکا ہوں

Challenge: All forms should map to one lemma.

Pashto Example

Root Verb:
لیکل

Forms:
لیکي، لیکلې، لیکلی دی

Challenge: Complex inflection rules.

Naive Stemming (Demo Only)

- Simple suffix stripping is often used
- Linguistically incorrect in many cases

Urdu Example:

کتابیں <– کتاب

But:

(Wrong) خوشی –> خوش

Existing Research

- Becker & Riaz (2012) — Urdu morphology
- Hardie (2003) — Corpus-based Urdu NLP
- Limited work for Pashto morphology

Impact on NLP Tasks

- Poor stemming affects:
  - Search
  - Topic modeling
  - Text classification
- Transformers reduce but do not eliminate the problem

Key Takeaways

- Stemming & lemmatization are unsolved problems
- Rule-based approaches are limited
- Data-driven solutions are needed

#Day6 #UrduNLP #PashtoNLP #Morphology

References

- Becker, M., & Riaz, K. (2012). Urdu morphology. *Journal of South Asian Linguistics*.
- Hardie, A. (2003). Urdu corpus-based NLP. Lancaster University.