Day 4: Text Normalization for Urdu & Pashto NLP

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

December 24, 2025

What is Text Normalization?

- Converting text into a standard form
- Unifying multiple Unicode variants
- Removing optional diacritics
- Essential preprocessing step

Unicode Challenges

- Same character, multiple Unicode forms
- Arabic, Persian, Urdu variants mixed
- Breaks tokenization and vectorization

Examples:

- ی | ي
- ک | ك
- ۀ | ہ | ھ

Urdu Normalization Example

Before:
اِسلام آباد

After:
اسلام آباد

Pashto Normalization Example

Before:
کابل کښې ژوند

After:
کابل کې ژوند

Diacritics Problem

- Optional in Urdu Pashto
- Create multiple word forms

Urdu: اِسلام اسلام

Pashto: کِتاب کتاب

Why Normalization Matters

- Smaller vocabulary
- Better TF-IDF features
- Improved embeddings
- Higher accuracy in NLP tasks

Related Work

- Zafar et al. (2019) — Urdu normalization
- Hardie (2003) — Urdu corpora processing
- Rahman et al. (2023) — Pashto preprocessing

Key Takeaways

- Normalization is mandatory for Urdu Pashto
- One word one Unicode form
- Preprocessing matters more than models

#Day4 #Normalization #UrduNLP #PashtoNLP