Day 5: Stopwords in Urdu & Pashto NLP

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

December 25, 2025

What are Stopwords?

- High-frequency functional words
- Usually removed during preprocessing
- Improve efficiency of NLP models

Why Stopwords are Hard in Urdu & Pashto

- No standardized stopword lists
- Context-sensitive words
- Morphological variations
- Dialectal differences

Urdu Stopword Example

Sentence:

یہ کتاب میز پر ہے

Stopwords:

ہے، پر

Pashto Stopword Example

Sentence:
دا کتاب په میز باندې ده

Stopwords:
دا، په، باندې، ده

Related Research

- Hussain, S. (2008). Urdu linguistic processing.
- Hardie, A. (2003). Developing a corpus-based Urdu NLP toolkit.
- Rahman et al. (2023). Pashto text preprocessing challenges.

Key Takeaways

- English stopword lists do NOT work
- Custom lists are required
- Context-aware filtering is better

#Day5 #UrduNLP #PashtoNLP #Stopwords