Day 8: Bag of Words (BoW) for Urdu & Pashto

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

December 27, 2025

What is Bag of Words (BoW)?

- A simple text representation technique.
- Text is converted into a list of word frequencies.
- Word order is ignored.
- Widely used in classical NLP.

Why BoW is Important

- Easy to understand and implement.
- Works well with small datasets.
- Strong baseline for low-resource languages.
- Useful for text classification and topic modeling.

BoW Example (English)

Sentence:
Ali reads a book

BoW Representation:

- Ali: 1
- reads: 1
- a: 1
- book: 1

BoW Example (Urdu)

Sentence:
علی کتاب پڑھتا ہے

Tokens:
علی | کتاب | پڑھتا | ہے

BoW Representation:

- علی: 1
- کتاب: 1
- پڑھتا: 1
- ہے: 1

BoW Example (Pashto)

Sentence:
احمد کتاب لولي

Tokens:
احمد | کتاب | لولي

BoW Representation:

- احمد: 1
- کتاب: 1
- لولي: 1

Limitations of BoW

- Ignores word order and context.
- Cannot capture meaning or semantics.
- Vocabulary grows very large.
- Sensitive to spelling variations (Urdu/Pashto).

Why BoW Still Works for Urdu & Pashto

- Works even without large pretrained models.
- Effective for small datasets.
- Easy baseline before deep learning.
- Helps understand core NLP concepts.

Key Takeaways

- BoW is the foundation of text representation.
- Very useful for low-resource languages.
- Should be combined with normalization and stopwords.

#Day8 #BagOfWords #UrduNLP #PashtoNLP #ClassicalNLP