Day 10: N-grams
Capturing Context in Urdu & Pashto

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

December 29, 2025

What are N-grams?

- Sequence of n consecutive words or characters.
- Unigram (1-word), Bigram (2-word), Trigram (3-word).
- Captures limited context.

Why N-grams are Important

- BoW ignores word order.
- N-grams preserve local context.
- Useful for sentiment and language modeling.

N-grams Example (Urdu)

Sentence:
وہ کتاب پڑھتا ہے

Unigrams: وہ | کتاب | پڑھتا | ہے

Bigrams: وہ کتاب | کتاب پڑھتا | پڑھتا ہے

N-grams Example (Pashto)

Sentence:
احمد کتاب لولي

Unigrams: احمد | کتاب | لولي
Bigrams: کتاب احمد | لولي کتاب

Word vs Character N-grams

- Word N-grams $\rightarrow$ meaning context
- Character N-grams $\rightarrow$ spelling variations
- Very useful for Urdu Pashto normalization

Limitations of N-grams

- Feature space grows rapidly.
- Sparse representations.
- Needs pruning or TF–IDF.

Key Takeaways

- N-grams bridge the gap between BoW and deep models.
- Extremely helpful for low-resource languages.
- Best combined with TF–IDF.

#Day10 #Ngrams #UrduNLP #PashtoNLP