Day 9: TF–IDF for Urdu & Pashto

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

December 28, 2025

What is TF–IDF?

- TF–IDF = Term Frequency – Inverse Document Frequency.
- Measures how important a word is in a document.
- Penalizes very common words.
- Improves over Bag of Words.

TF–IDF Formula

$$TF\text{-}IDF(w, d) = TF(w, d) \times IDF(w)$$

- TF: Frequency of word in document.
- IDF: Rarity of word across documents.
- Rare + frequent = important.

Why TF–IDF is Important

- Reduces impact of stopwords.
- Highlights informative words.
- Improves text classification accuracy.
- Very effective for low-resource languages.

BoW vs TF–IDF

- BoW counts words only.
- TF–IDF assigns weights.
- Common words get low scores.
- Important words get high scores.

TF–IDF Example (Urdu)

Documents:

1) یہ ایک سادہ جملہ ہے

2) یہ جملہ این ایل پی کے بارے میں ہے

Observation:

- یہ، ہے → TF–IDF low
- این ایل پی → TF–IDF high

TF–IDF Example (Pashto)

Documents:

1) دا يو ساده جمله ده

2) دا جمله د اين ايل پي په اړه ده

Observation:

- TF–IDF low → دا، ده
- TF–IDF high → اين ايل پي

Limitations of TF–IDF

- Still ignores word order.
- No semantic understanding.
- Sensitive to spelling variations.
- Needs normalization beforehand.

Why TF–IDF Works for Urdu & Pashto

- No pretrained models required.
- Strong baseline for research.
- Suitable for small datasets.
- Widely used in academic NLP.

Key Takeaways

- TF–IDF improves BoW significantly.
- Highlights meaningful words.
- Essential classical NLP technique.

#Day9 #TFIDF #UrduNLP #PashtoNLP #ClassicalNLP