30-Day LinkedIn Posting Plan

Pashto, Urdu & English NLP

Practical, Educational & Engaging Content

Sami Uddin Shinwari
NLP Researcher | Subject Specialist IT

Overview

This 30-day content plan is designed to build awareness, skills, and engagement around Natural Language Processing (NLP) for low-resource languages, with a special focus on **Pashto and Urdu**. Each post combines theory, practical insight, and real-world relevance.

Week 1: Foundations & Basics

- **Day 1: What is NLP?**
  Explain NLP in simple terms.
  Post idea: Why NLP matters for low-resource languages.
- **Day 2: Challenges in Pashto & Urdu NLP**
  Script issues, lack of datasets, dialect variation.
  Post idea: Comparing Pashto/Urdu challenges with English.
- **Day 3: Tokenization**
  Urdu/Pashto tokenization problems.
  Post idea: Why whitespace tokenization fails.
- **Day 4: Normalization**
  Unicode issues, diacritics, character unification.
  Post idea: Why normalization is crucial for Urdu & Pashto.
- **Day 5: Stopwords**
  Why stopword removal is harder in Urdu/Pashto.
  Post idea: Examples of Urdu/Pashto stopwords.
- **Day 6: Stemming & Lemmatization**
  Morphological complexity of Urdu/Pashto.
  Post idea: Difference between stemmers and lemmatizers.

- **Day 7: Language Resources**
  Libraries, corpora, and open datasets.
  Post idea: Links to Pashto/Urdu NLP resources.

Week 2: Classical NLP Techniques

- **Day 8: Bag of Words (BoW)**
  Basic text representation.
  Post idea: Simple Urdu/Pashto sentence example.
- **Day 9: TF–IDF**
  Importance weighting for documents.
  Post idea: BoW vs TF–IDF comparison.
- **Day 10: N-grams**
  Capturing local context.
  Post idea: Why bigrams and trigrams matter in Urdu/Pashto.
- **Day 11: Part-of-Speech Tagging**
  Tools and challenges for Urdu/Pashto.
  Post idea: Why POS tagging is essential.
- **Day 12: Named Entity Recognition (NER)**
  Identifying names, places, organizations.
  Post idea: Use cases for Urdu/Pashto NER.
- **Day 13: Sentiment Analysis**
  Polarity detection challenges.
  Post idea: Why sentiment lexicons are lacking.
- **Day 14: Text Classification Pipeline**
  End-to-end multilingual NLP workflow.
  Post idea: Visual pipeline overview.

Week 3: Modern Machine Learning NLP

- **Day 15: Word Embeddings (Word2Vec, GloVe)**
  Dense word representations.
  Post idea: Urdu/Pashto embedding challenges.
- **Day 16: FastText for Low-resource Languages**
  Subword modeling advantages.
  Post idea: Why FastText suits morphologically rich languages.

- **Day 17: Recurrent Neural Networks (RNNs)**
  LSTMs and GRUs.
  Post idea: Sequence modeling explanation.

- **Day 18: Seq2Seq Models**
  Applications in translation and summarization.
  Post idea: Encoder–decoder intuition.

- **Day 19: Machine Translation Issues**
  Parallel corpus scarcity.
  Post idea: Urdu/Pashto MT limitations.

- **Day 20: Language Models**
  n-gram vs neural language models.
  Post idea: Evolution of language modeling.

- **Day 21: Data Augmentation**
  Back-translation, paraphrasing.
  Post idea: Boosting low-resource datasets.

Week 4: Deep Learning & Transformers

- **Day 22: Transformer Architecture**
  Attention-based models.
  Post idea: Transformers explained simply.

- **Day 23: BERT**
  Multilingual BERT for Urdu/Pashto.
  Post idea: Strengths and limitations.

- **Day 24: XLM-RoBERTa**
  Improved performance on low-resource languages.
  Post idea: Why XLM-R works better.

- **Day 25: LLMs for Pashto & Urdu**
  Challenges in building local LLMs.
  Post idea: Data and compute barriers.

- **Day 26: Fine-tuning Models**
  Fine-tuning mBERT/XLM-R.
  Post idea: Step-by-step overview.

- **Day 27: Evaluation Metrics**
  Accuracy, F1-score, BLEU.
  Post idea: Choosing the right metric.

- **Day 28: Real-world Applications**
  Education, media, governance.
  Post idea: NLP impact in South Asia.

- **Day 29: Creating Pashto/Urdu Datasets**
  Crowdsourcing and scraping.
  Post idea: Ethical dataset creation.

- **Day 30: Future of NLP in South Asia**
  LLMs, speech technology, dialect modeling.
  Post idea: Roadmap for the next decade.