Natural Language Processing – Assignment 1

Language Modeling for Urdu News Articles

Student ID: i23-2548 | Section: DS-A

1. Project Overview

This project presents a comprehensive Natural Language Processing (NLP) pipeline for modeling Urdu BBC news articles using statistical language modeling techniques. The system spans web scraping, preprocessing, morphological normalization, n-gram probability estimation, smoothing, interpolation, and constrained article generation. All linguistic tools—including tokenization, stemming, and lemmatization—were implemented from scratch to better understand Urdu's morphological richness and script complexity. The primary objective was to generate high-quality Urdu news text that resembles professional BBC Urdu journalism in structure and tone.

2. Data Collection and Preprocessing

A hybrid scraping strategy was employed using Selenium for dynamic pages and XML sitemap parsing for archived content. The final dataset contains 216 articles with a 90/10 training-testing split, comprising 7,682 processed sentences and 9,176 unique root tokens after normalization.

Preprocessing included regex-based noise removal, diacritic stripping, Urdu-specific sentence segmentation, custom tokenization preserving structural tokens (, ,), suffix-based stemming, and rule-based lemmatization to reduce vocabulary sparsity.

3. Language Model Implementation and Smoothing

Unigram, Bigram, and Trigram models were implemented using nested dictionaries. Add-k smoothing with a standardized k=0.0001 was applied across all models to eliminate zero probabilities without over-smoothing. For the Trigram model, linear interpolation ($\lambda1$=0.4, $\lambda2$=0.4, $\lambda3$=0.2) was used to combine contextual strengths from tri-, bi-, and unigram probabilities.

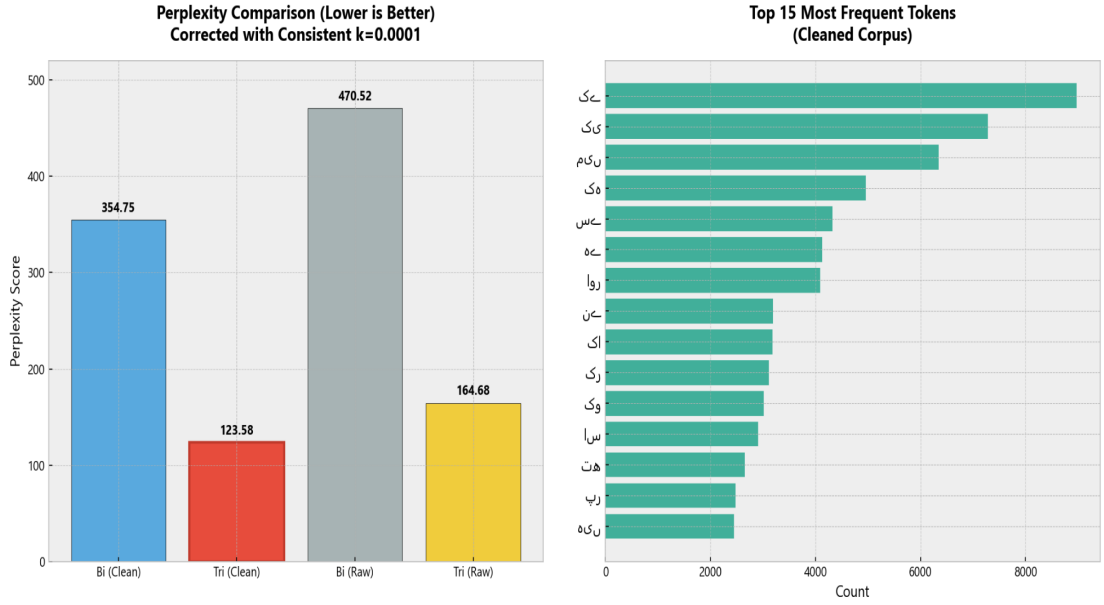
4. Article Generation Process and Constraints

Articles were generated using the interpolated Trigram model with Bigram and Unigram backoff. Generation began with tokens and continued probabilistically until reaching . Constraints included 200–250 words per article, a minimum of five sentences, and

maintenance of formal BBC Urdu tone.

5. Evaluation and Results Analysis

Perplexity was used as the primary evaluation metric. Using standardized smoothing (k=0.0001), the final optimized results are as follows: Cleaned Bigram = 354.75, Cleaned Trigram (Interpolated) = 123.58, Raw Bigram = 470.52, Raw Trigram (Interpolated) = 164.68. These results confirm that preprocessing significantly improves performance and that the interpolated Trigram model consistently achieves the lowest perplexity.



The left visualization compares perplexity across raw and cleaned models under a consistent smoothing configuration. The cleaned interpolated Trigram model achieves the lowest perplexity (123.58), demonstrating the effectiveness of morphological normalization and interpolation. The right visualization displays the top 15 most frequent tokens in the cleaned corpus, reflecting dominant news-domain vocabulary.

6. Qualitative Evaluation

Fluency and Grammar: The Trigram model produces more fluent and grammatically coherent Urdu phrases compared to the Bigram model due to extended contextual windows. Coherence and Meaning: While local coherence is strong, long-range semantic consistency remains limited due to the statistical nature of n-gram models. Preprocessing Impact: The cleaned pipeline reduces sparsity by collapsing morphological variants, enabling more reliable transition probability estimation.

7. Conclusion

Using standardized smoothing and interpolation, the optimized Trigram model achieved substantial improvements over raw configurations. The findings highlight that careful preprocessing and consistent hyperparameter tuning are critical for effective Urdu language modeling in low-resource settings.