

**Natural Language Processing**

**Assignment # 3**

**Muhammad Haziq Ijaz**

**I212692**

**A**

Report: Urdu Text Segmentation and Tokenization Using Custom SentencePiece Model

**Introduction**

The project focuses on segmenting and tokenizing Urdu text, with the aim of evaluating the effectiveness of a custom segmentation and tokenization algorithm. The text used in this study is extracted from a provided Urdu corpus, and several methods were employed to detect sentence boundaries and tokenize the text into subword units. The results of the segmentation and tokenization process were then evaluated based on character similarity, word preservation, and sentence count.

**Methods**

**1. Text Segmentation**

- Goal: To break down the Urdu text into sentences based on common sentence-ending patterns such as punctuation marks (e.g., '۔', '؟', '!') and common end words (e.g., 'ہے', 'تھا', 'ہوں').

- Approach: A predefined list of common sentence-ending words was augmented with frequently occurring words from the input text. Regular expressions were used to split the text into sentences.

**2. SentencePiece Tokenization**

- Goal: To break sentences into smaller subword units (tokens) for more efficient text processing.

- Approach: A simplified version of the SentencePiece model was implemented. The model was trained on the segmented sentences to identify frequent subword units, using subwords of length 1 to 4.

**3. Evaluation Metrics**

- Character Similarity: Percentage of matching characters between original and processed text.

- Word Preservation: Percentage of original words retained in the processed text.

- Sentence Count Ratio: Comparison of sentence count between original and segmented text.

- Overall Score: Average of character similarity, word preservation, and sentence count ratio.

**Results**

**1. Segmentation Evaluation**

Character Similarity: 18.20%

Word Preservation: 100.00%

Sentence Count Ratio: 100.00%

Overall Score: 72.73%

**2. Tokenization Evaluation**

Character Similarity: 100.00%

Word Preservation: 100.00%

Sentence Count Ratio: 100.00%

Overall Score: 100.00%

**Segmented Output**

بے چاری عوام چونکہ ہمیشہ سے دھوکہ کھانے کی عادی رہی ہے اس لئے ‘‘تبدیلی سرکار’’ کی چکنی چپڑی باتوں میں آگئی اور اپنے بہتر مستقبل کے لئے نئی حکومت کو اقتدار کے ایوانوں تک پہنچا دی

**Discussion**

1. Segmentation Accuracy: High word preservation and sentence count ratio (100%) indicate correct sentence boundary detection. However, character similarity (18.20%) was lower due to formatting changes.

2. Tokenization Performance: The custom SentencePiece model achieved a perfect overall score (100%). It effectively encoded and decoded subword units without loss of information.

3. Strengths: The combination of predefined common end words and frequent pattern detection improved segmentation accuracy.

4. Areas for Improvement: Character similarity could be enhanced by refining space and punctuation handling.