
# University of Engineering and Technology, Lahore

# (New Campus)

## Semester Project

### Group Members


| | |
|---|---|
| Hadeeba Javed | (2022-CS-709) |
| Umair Sajid | (2022-CS-719) |
| Ali Zahoor | (2022-CS-723) |
| Shabih Haider | (2022-CS-724) |


## Project Title

### Sentiment Analysis of Urdu Drama Transcripts

**Submitted to**

Maam Qurat-ul-Ain


**Department of Computer Science**

# Introduction

This project focuses on evaluating the accuracy of English-to-Urdu translation using state-of-the-art machine translation models. Our dataset consists of English transcripts of Pakistani dramas along with their Urdu translations. We aim to analyze the quality of these translations using automatic evaluation metrics, primarily BLEU (Bilingual Evaluation Understudy)..

# Objective

- To check the accuracy of existing Urdu translations from the dataset.
- To re-translate English sentences using a pre-trained model (facebook/m2m100_418M).
- To compare and evaluate the new translations using BLEU score metrics.
- To fine-tune and evaluate the Helsinki-NLP/opus-mt-en-ur model for improved translation performance.

## Dataset

Each team member was responsible for:

- Handling a part of the English-Urdu transcript data.
- Adding a new column to re-translate English to Urdu using facebook/m2m100_418M.

Later, we:

Merged individual Excel files into a single merged_specific_files.xlsx file.

- Focused on two main columns:
- English Sentence

Urdu Sentence (facebook/m2m100_418M)YouTube API.

- Each transcript was originally in English, including timestamps and speaker turns.

  To enhance the dataset, these English subtitles were automatically translated into Urdu using the YouTube subtitle translation feature via the YouTube API.

# Translation Model

- **Model Used for Re-translation**

  We use: facebook/m2m100_418M
- **Model Used for Fine-Tuning**

  We use: Helsinki-NLP/opus-mt-en-ur

  A model specifically trained for English to Urdu translation. We fine-tuned this

model on our cleaned dataset.

## Data Preprocessing

- Cleaned extra whitespaces and special characters from text.
- Removed rows with missing or empty translations
- Normalized both English and Urdu texts.
- Tokenized input and output using HuggingFace's tokenizer.
- Converted dataset into HuggingFace Dataset format.
- Split into training and evaluation sets (90/10 split).

# Training Model

- We used the HuggingFace `transformers` library for model training. Our base model was `Helsinki-NLP/opus-mt-en-ur`, a pre-trained machine translation model for English to Urdu.
- To fine-tune the model, we prepared a dataset of Pakistani drama scripts containing English dialogues and their corresponding Urdu translations.
- We tokenized the input and output sentences and trained the model for several epochs using the `Trainer` API.
- The model learned contextual translations relevant to the informal, conversational language typical in drama scripts.

# Evaluation Tools

- We evaluated the model performance using the BLEU (Bilingual Evaluation Understudy) score. BLEU measures the similarity between the machine-generated translation and one or more reference translations.
- We used the `sacrebleu` library to compute BLEU scores on a held-out test dataset.

# Implemented Tools

- Python (v3.10+)
- HuggingFace Transformers
- Datasets Library (HuggingFace)
- SacreBLEU (for evaluation)

# Results & Discussion

We successfully trained and evaluated our fine-tuned translation model on the merged dataset of Pakistani drama transcripts.

### Translation Quality Evaluation:

The model was evaluated using the BLEU score, which measures how closely a machine-generated translation matches a human reference.

### BLEU scores were calculated by comparing:

- The Urdu translations generated by the model.
- The original Urdu sentences present in the dataset.

## Final Evaluation Outcome:

Final BLEU score on 500 test examples: 58.92
This indicates a moderately high translation accuracy, especially considering the informal and conversational nature of the drama script dataset