Code Mixed Machine Translation

Advanced NLP Project - Monsoon 2022

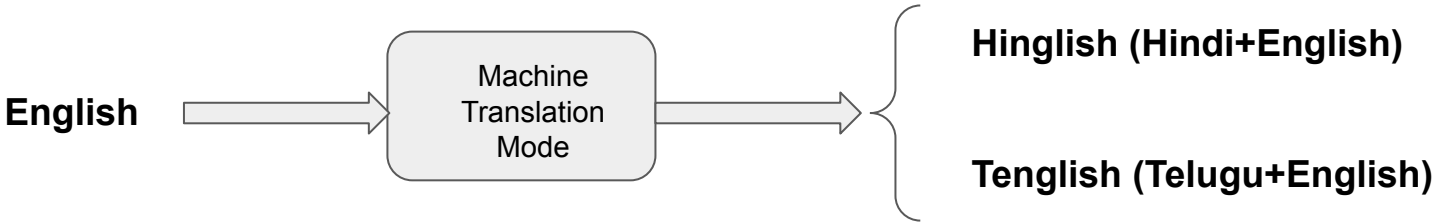
- Gopichand Kanumolu (2021701039)
- Lokesh Madasu (2021701042)

This presentation contains following sections

- Problem statement
- Dataset statistics & Plots
- Baseline Model
- Model Results
- Baseline + Model
- Comparison and Analysis of results

Problem Statement

- Code-mixing is using words from two or more languages in the same conversation.
- We encounter code-mixing in almost all the social media platforms, where users communicate informally using emojis, code mixing the text etc.
- Building a code-mixed machine translation system is not a trivial task, because the code-mixed languages do not follow a prescriptively defined structure.
- In this project we attempt to build a machine translation system that translates the English sentence into codemixed language Hinglish (combination of words from Hindi and English)
- In addition to that, we also attempt to build a machine translation system that translates English sentence into codemixed Telugu language



English-Hinglish Dataset Statistics

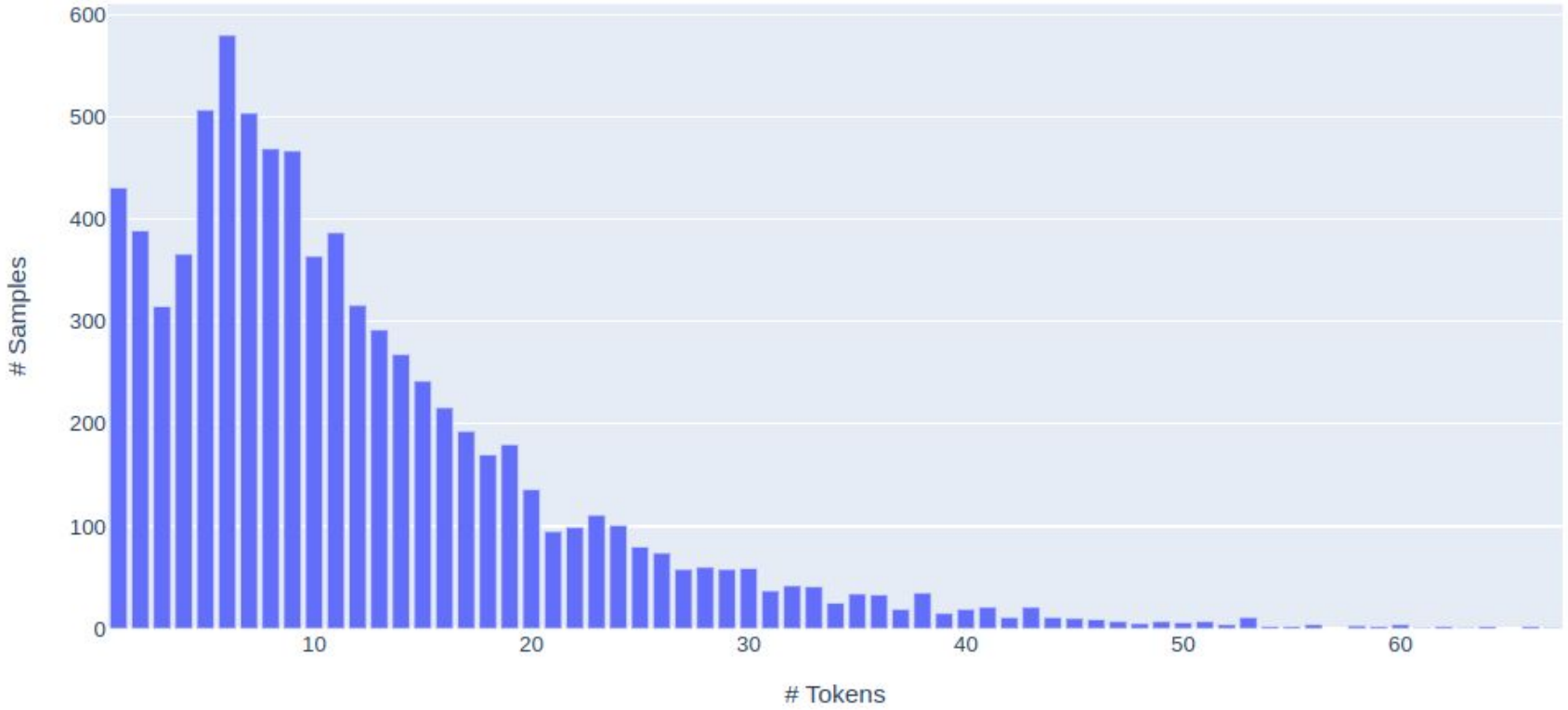
Field	English	Hinglish
Number of sentences in train data	8060	8060
Number of sentences in validation data	942	942
Number of sentences in test data	960	960
Minimum Number of Tokens	1	1
Maximum Number of Tokens	287	370
Average Number of Tokens	12.37	12.6

Table 1: dataset statistics

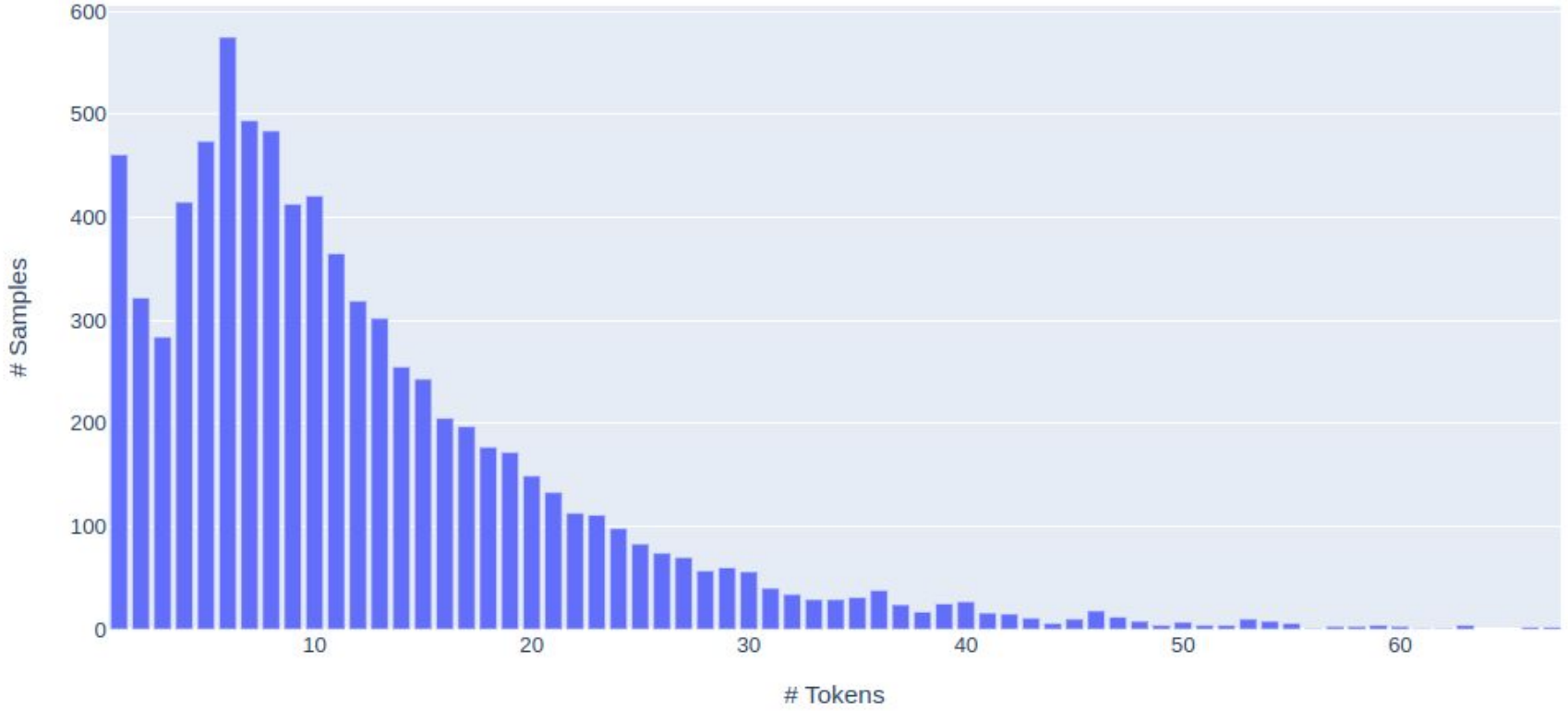
Field	Value
Duplicate English-Hindi Sentence Pairs in train set	519
Duplicate English-Hindi Sentence Pairs in validation set	12
Duplicate English Sentences in test_set	182
Number of Hindi tokens in target sentences	76364
Number if English tokens in target sentences	24269
Number of 'Other' tokens in target sentences	13730

Table 2: other statistics

Plot of Number of Tokens in English Sentence Vs Number of Sentences



Plot of Number of Tokens in Hinglish Sentence Vs Number of Sentences

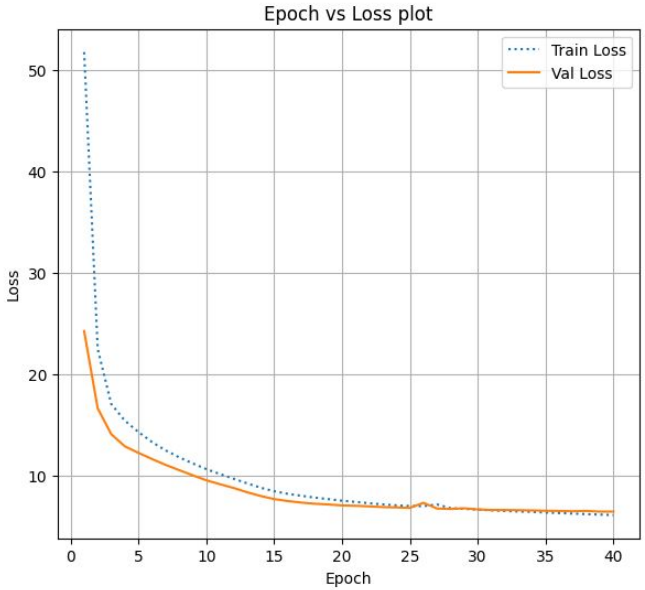


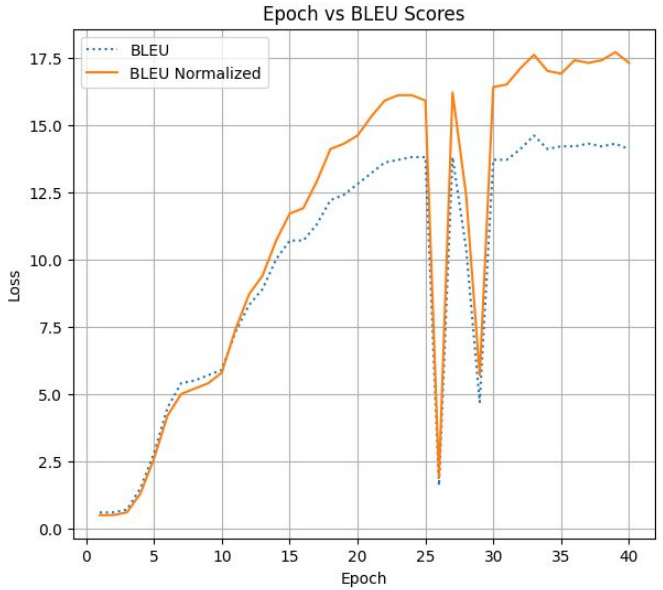
Baseline Model (mBART)

- We fine-tune mBART, which is a multilingual sequence-to-sequence transformer architecture
- It has 12 encoder and decoder layers each and hidden state dimension of 1024 and uses 16 attention heads resulting in ~680 million parameters.
- We use the following variations of mBART
    - **mBART-en:** In this model, we input the English sentence and the model has to generate Hinglish sentence
    - **mBART-hien:** In this variant, along with the English sentence we pass the parallel Hindi translated input to the encoder and the model has to generate the Hinglish sentence.
    - In both the cases, we use beam search decoder with a beam size of 5 for decoding

- Some parameter of the model
    - Loss : Categorical Cross Entropy
    - Learning Rate : 3E-05
    - Dropout : 0.3
    - Attention Dropout : 0.1
    - Number of Epochs : 40

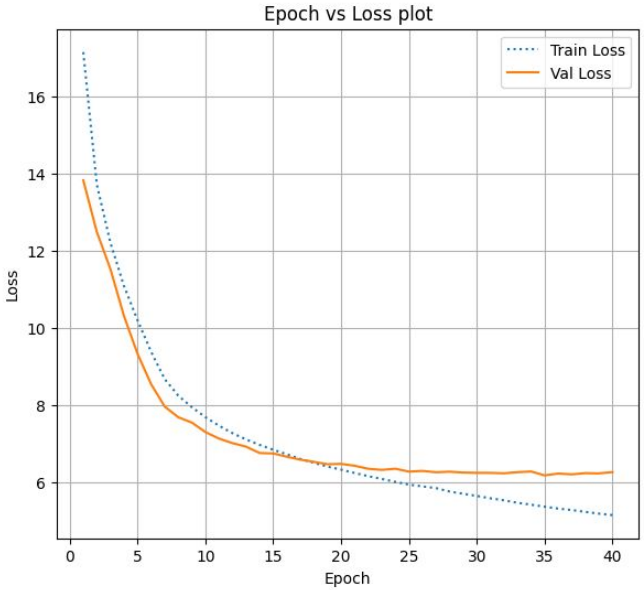
Baseline Results (mBART-en model)

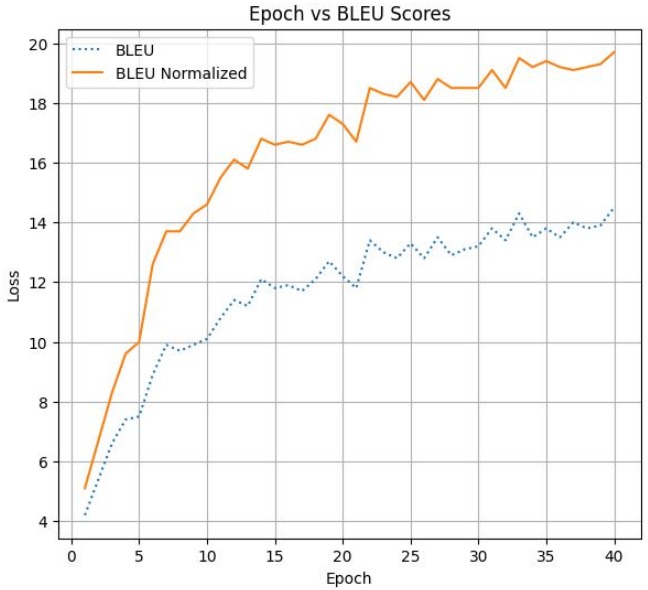




BLEU Score : 14.3
BLEU Normalized : 17.7

Baseline Results (mBART-hien model)



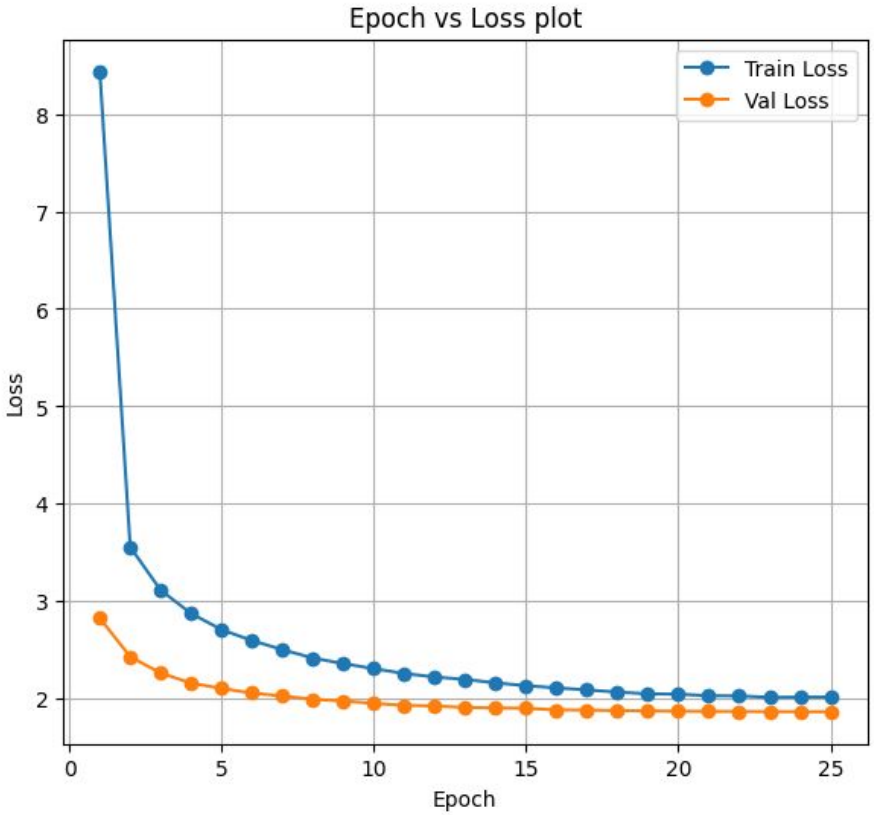


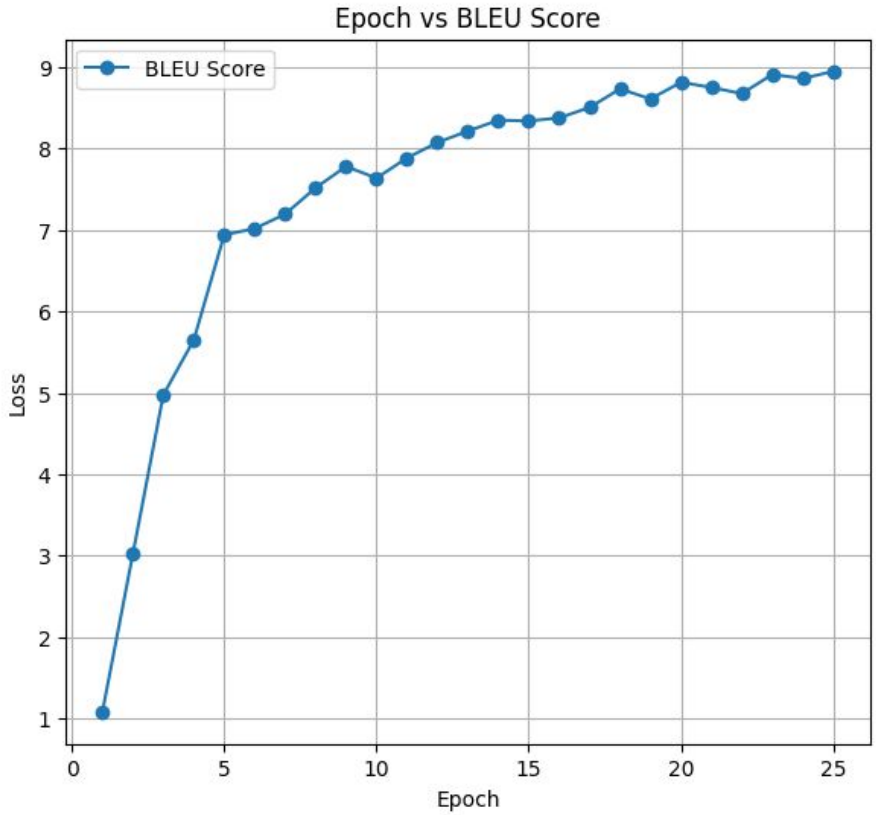
BLEU Score : 14.5
BLEU Normalized : 19.7

Results

mT5 (English-Hinglish)

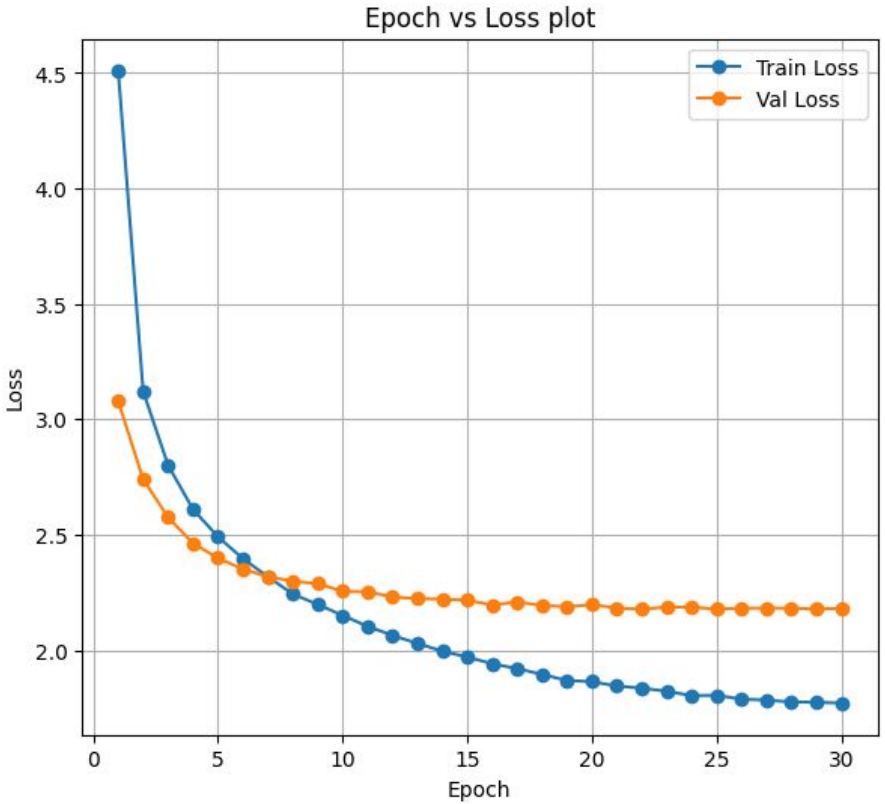
BLEU Score: 8.94

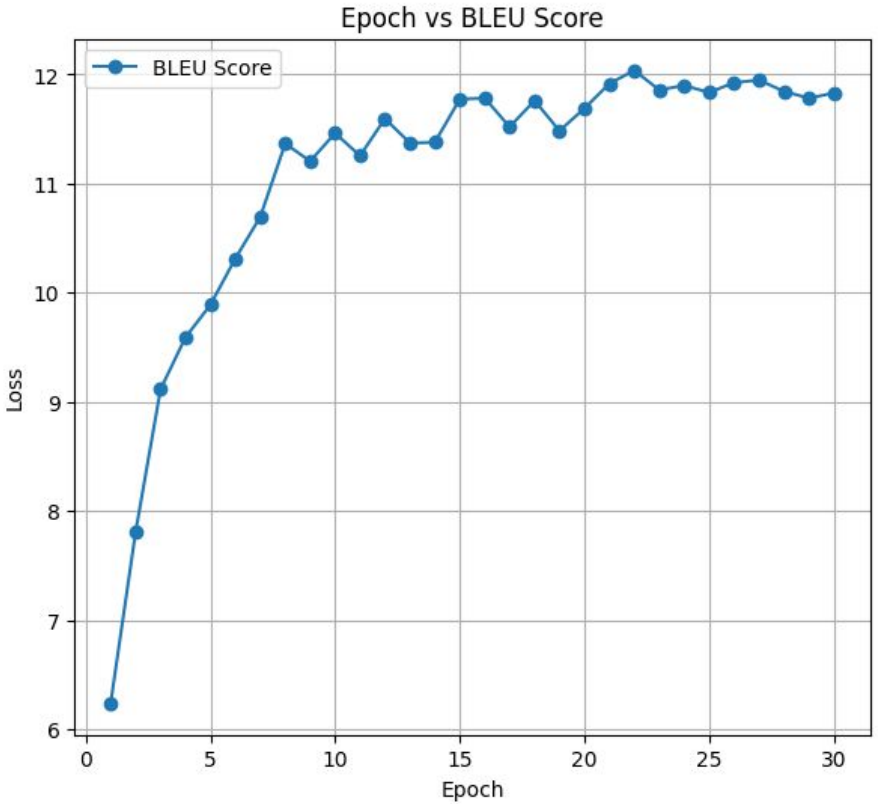




IndicBART (English-Hinglish)

BLEU Score: 11.83



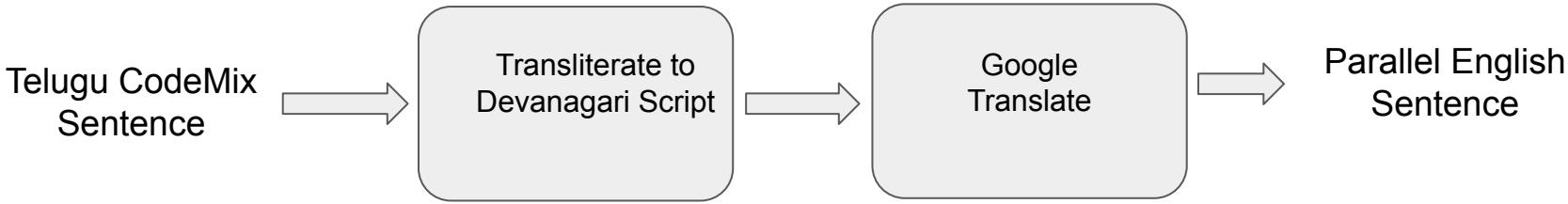


Analysis of English-Hindi Code Mix Outputs

English Sentence (Input)	Generated Sentence (CodeMix)	Comment
I think they have a strong defense to beat the Chitauri invaders.	मुझे लगता है कि उनके पास Chitauri invaders को beat करने का strong defense है	Perfect translation
the ratings are very low for both rotten tomatoes and metacritic	rotten tomatoes और metacritic की ratings बहुत कम है	Perfect translation
I do. I adore sci-fi. You?	मैं करता हूँ. मैं sci-fi adore करता हूँ. आप?	Perfect translation
I think you watched this movie also	मुझे लगता है कि मैंने यह movie भी देखी है	meaning changed
Me too. The more unexpected, the better!	मैं भी. ज्यादा unexpected है	Incomplete or missing info
Yeah, even though I love comedies, this wasn't 1 of my faves	हां, even though मुझे comedies पसंद हैं लेकिन मेरे fades में यह नहीं है	Factually incorrect

Preparing CodeMix Telugu-English Dataset for Machine Translation

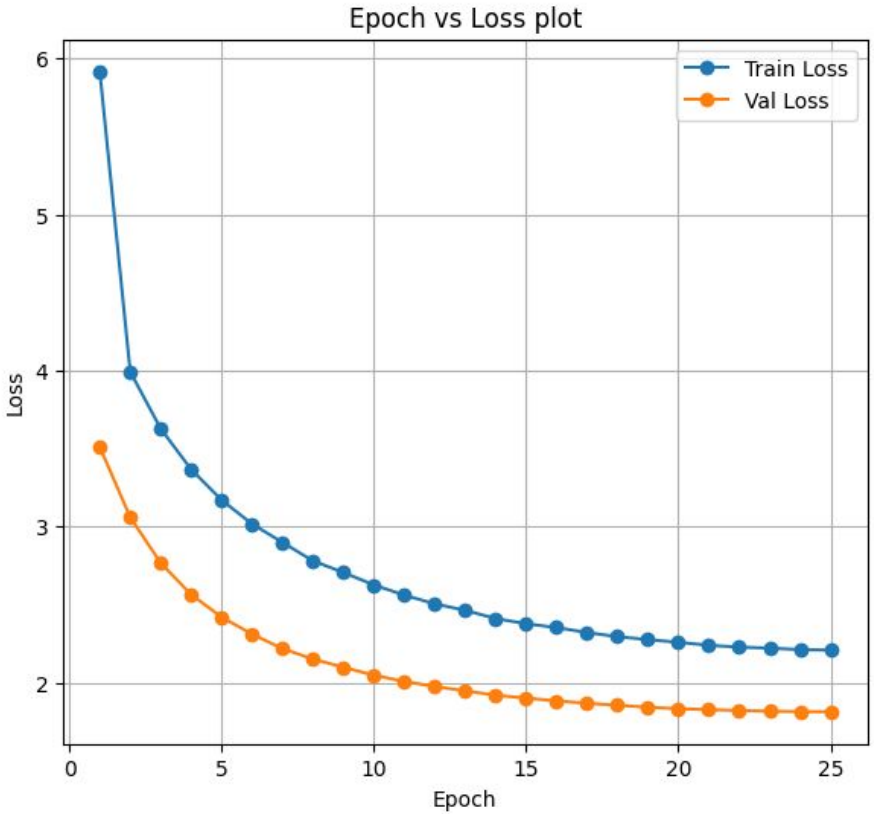
- Since there are no existing datasets available for Telugu-English CodeMixed Translation, we attempt to create a dataset by using existing monolingual (Telugu) dataset.
- We use sentiment classification dataset containing codemix Telugu sentences and generate parallel English sentences for training as follows.

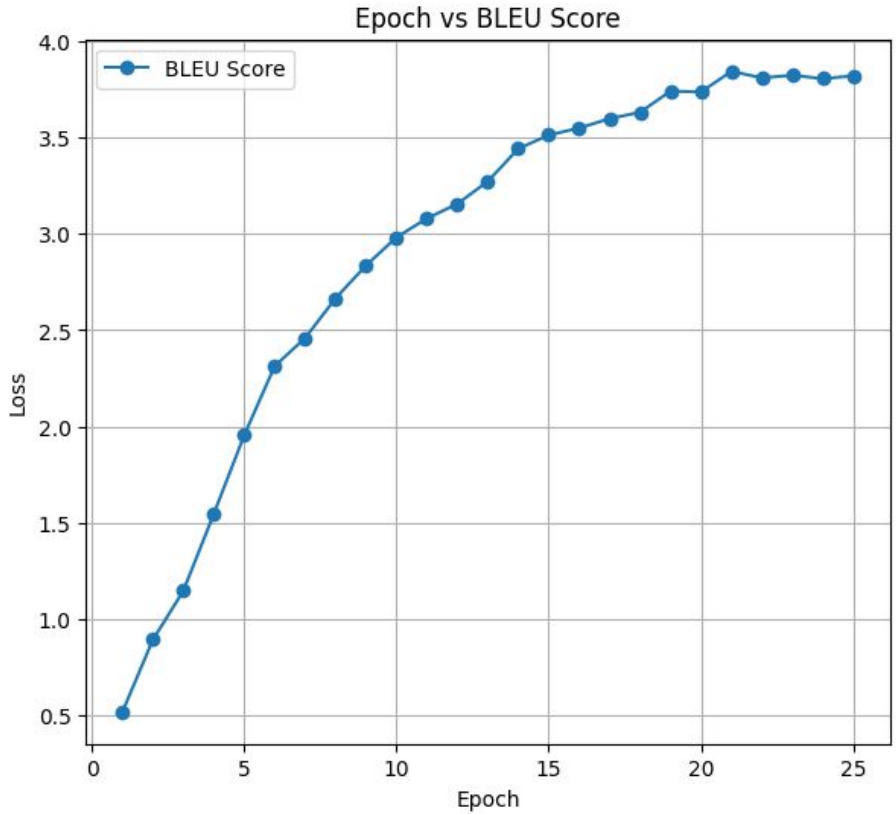


- To train the English-Telugu CodeMix model, we feed the Parallel English Sentence as input and the model has to generate Telugu CodeMix Sentence

mT5 (English-Telugu Codemix)

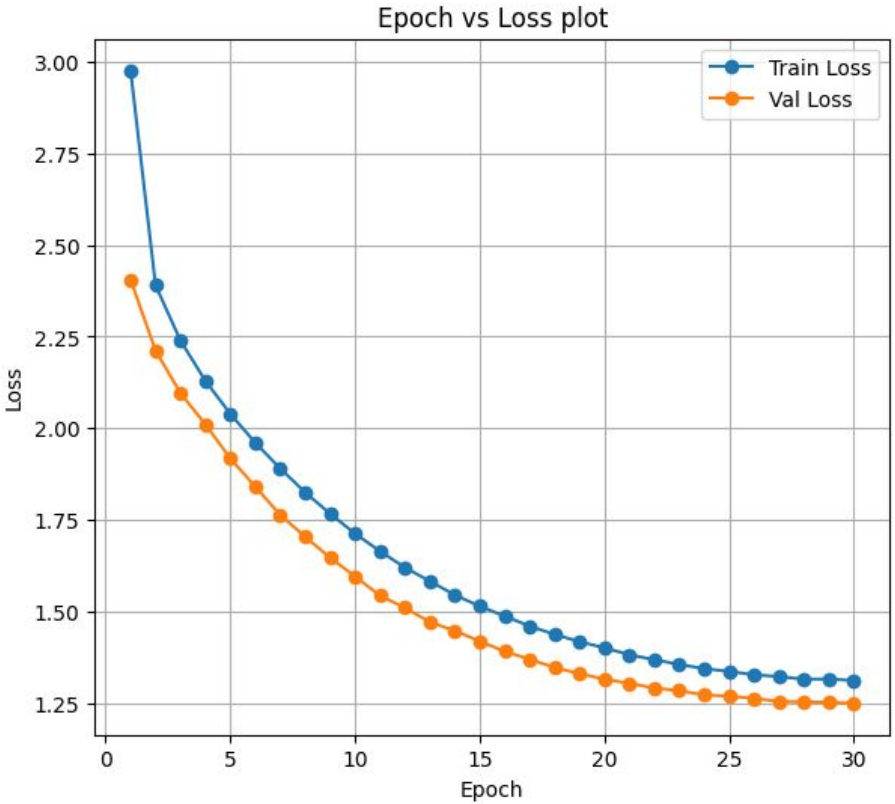
BLEU Score: 3.818

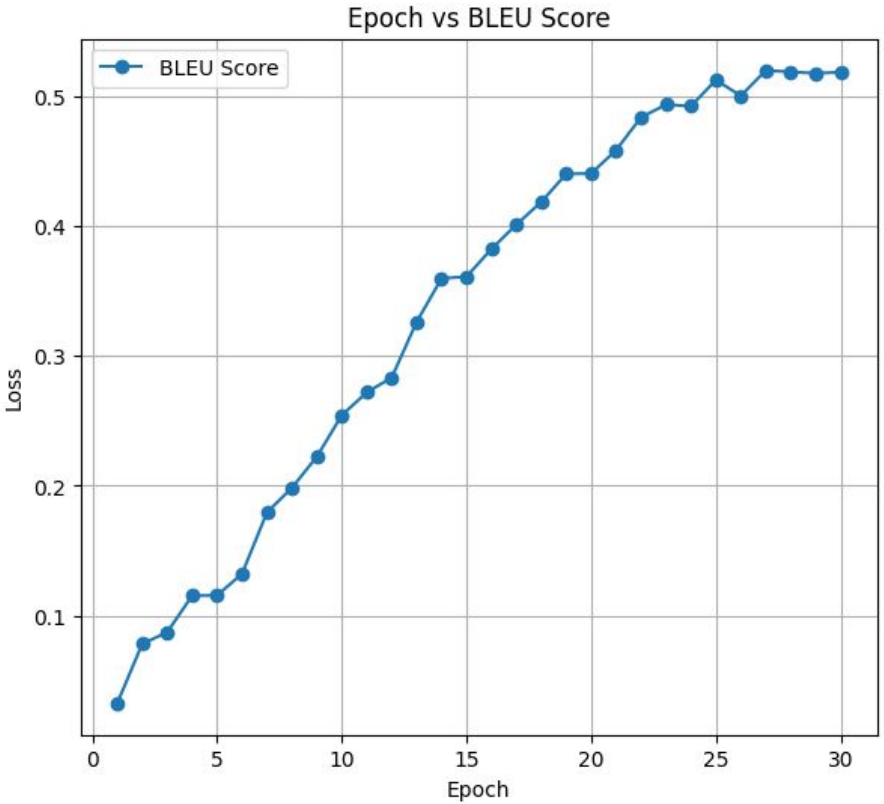




IndicBART (English-Telugu Codemix)

BLEU Score: 0.518





Analysis of English-Telugu Code Mix Outputs

English Sentence (Input)	Generated CodeMix (English-Telugu)	Comment
My voice in Telugu is very good.	తెలుగు లో నా వాయిస్ చాలా బాగుంది.	Perfect translation
Raj Tarun Power Play Movie Review Tell Bro	రాజ్ తరుణ్ పవర్ ప్లే మూవీ రివ్యూ చెప్పు బ్రో	Perfect translation
Hello Thyview. Waiting for Master Review	హెల్లో థైవ్యూ. మాస్టర్ రివ్యూ కోసం వెయిటింగ్	Perfect translation
Locations are the main plus points of the movie	మూవీ లోనే లొకేషన్స్ మాత్రం మెయిన్ పాయింట్స్	meaning changed
Bro talk in normal language	బ్రో లాంగ్వేజ్ లో మాట్లాడు బ్రో	missing info
@Chandler999999 I saw Knight and it was great.	@కౌంటర్ర్ర్ర్ నేను కోటిని చూసాను అది గ్రేట్ గా ఉంది.	Factually incorrect

Thank You

Queries?