DEBIASING TEXT USING STYLE TRANSFER

https://github.com/jerin-mx/Debiasing-Text-With-Style-Transfer

INTRODUCTION

Problem

- Address Subject Bias in Language
- Transform biased language into a more neutral form while preserving content.
- Falls under **Text Style Transfer**

Solution

Two-Stage Style Transformation Approach:

- Using BART model, fine-tuned for the task of transforming subjective text into a neutral style.
- For classification, employing fine-tuned BERT and SentenceBERT models to assess style changes and content preservation.

Baseline and Main Model Architectures

- Baseline Approach: LSTM seq2seq model with BERT tokenizer
- ML Architectures Used: BART for style transfer, BERT and SentenceBERT for evaluation metrics like CPS and STI.

Summary

Initial results show promising bias mitigation, with the BART model effectively neutralizing subjective language while retaining content. The STI and CPS scores from BERT-based evaluations provide quantitative insights into the efficacy of our approach.

DATA COLLECTION

Data Source

WNC (Wiki Neutrality Corpus)

181,473

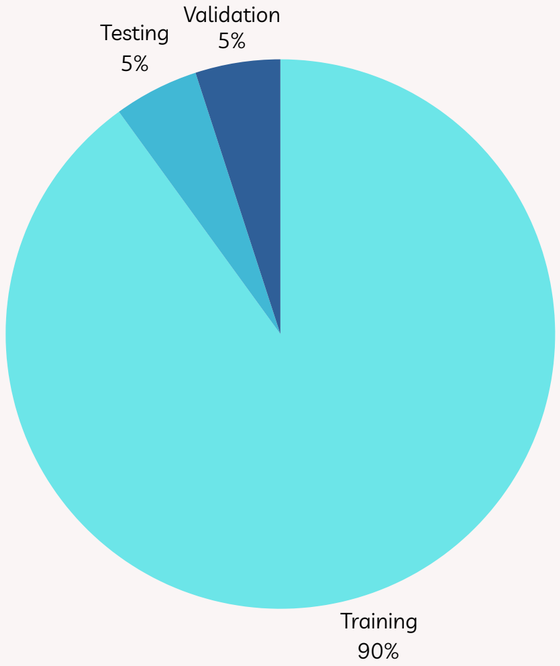
Selection & Accessibility

- Selected one-word edits from WNC : ~56,000
- Simplifies analysis
- Offers targeted insights into language biases
- Publicly available on Kaggle

Cleaning & Prep

- Minimal cleaning required
- Includes raw and tokenized texts, with part-of-speech and syntactic parse tags

Data Division

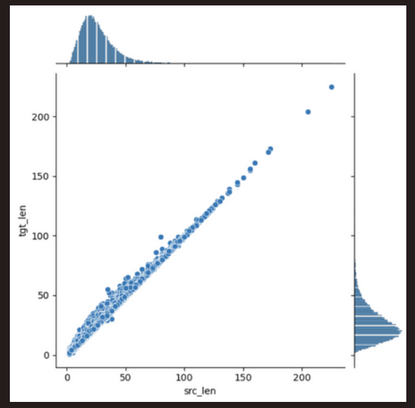


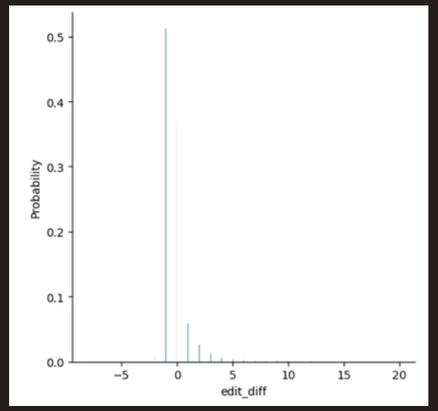
Data Sample

id	source	target
1	The company's new product is the best on the market.	The company's new product is one of the best on the market.
2	This book is a must-read for anyone interested in the topic.	This book is a valuable resource for anyone interested in the topic.
3	adequate testing for adverse health effects as well as performance data for these devices are seriously lacking.	adequate testing for adverse health effects as well as performance data for these devices are lacking.

EXPLORATORY DATA ANALYSIS

Objective: Understand dataset structure and text edits & inform data preprocessing and model design.





Text Length

- Most edits alter only 1-2 words
- 3-4% more than 1 word edits
- outliers identified : < 3 words & > 170. Drop texts with less than 3 word

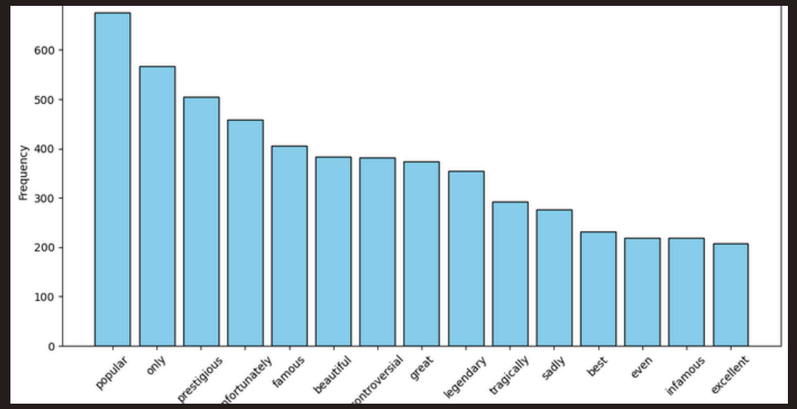
Text Quality

- Overall very good
- Some short texts have HTML tags and Color codes, but very few such samples
- A few duplicates

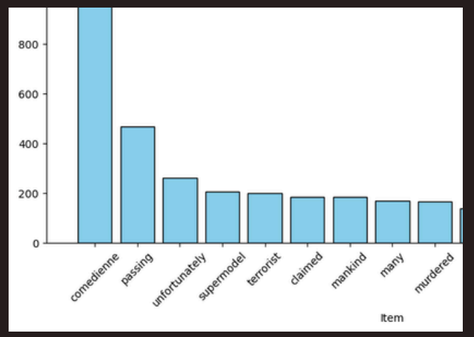
Source vs Target

- 50% subjectivity words removals
  - Mostly adjectives
- 40% replacement
  - Mostly adverbs
  - Replaced with lighter tones

Removed Words



Replaced words



Terrorist:
assailant
attacker
criminal
fighter
.
.

PREPROCESSING

Data Loading & Structuring

Duplicate Removal

Outlier Handling

Data Prep for BART Model

Conversion for BERT Model

- Cleaned and saved data in HuggingFace Dataset format
- Retained essential columns: 'id', 'source', and 'target'
- Prepared datasets: 51,961 rows (train), 672 (validation), 963 (test)

- Transformed dataset for binary classification with BERT
- Original sentences labeled as biased (1), style-transferred as unbiased (0)
- Doubled number of records, maintaining data distribution.

BASELINE MODEL

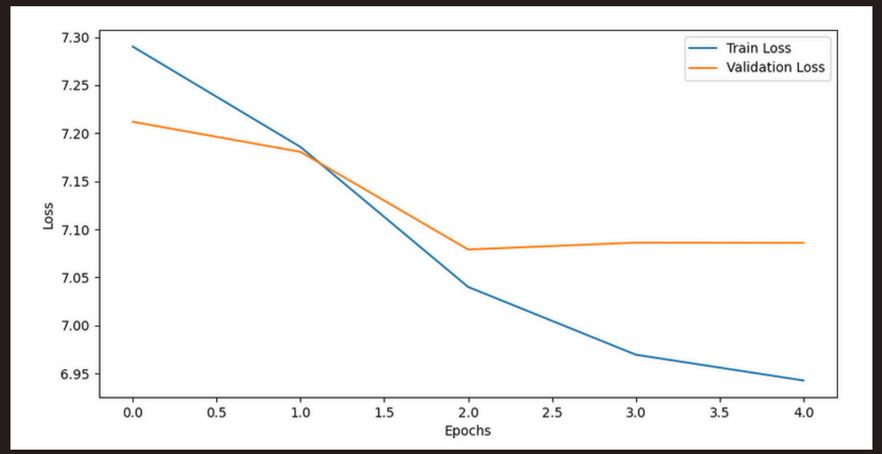
About Model & Training

- Custom-built Seq2Seq model with LSTM layers
- BERT tokenizer for robust text tokenization
- Model trained from scratch for seq2seq task conversion of biased text to unbiased text

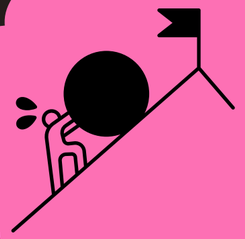
Model Configuration

- Encoder-Decoder Architecture: LSTM
- Gradient Accumulation Steps: 4
- Learning Rate: CosineAnnealingLR
- Optimizer: Adam
- Early Stopping: Yes
- Number of Layers: 4
- Dropout: 0.5

Performance



Challenges



- Requires careful tuning of hyperparameters
- Less sophisticated compared to pre-trained models
- Risk of overfitting due to training from scratch
- Insufficient training might be a factor
- The task complexity might be too high for this model

BART MODEL

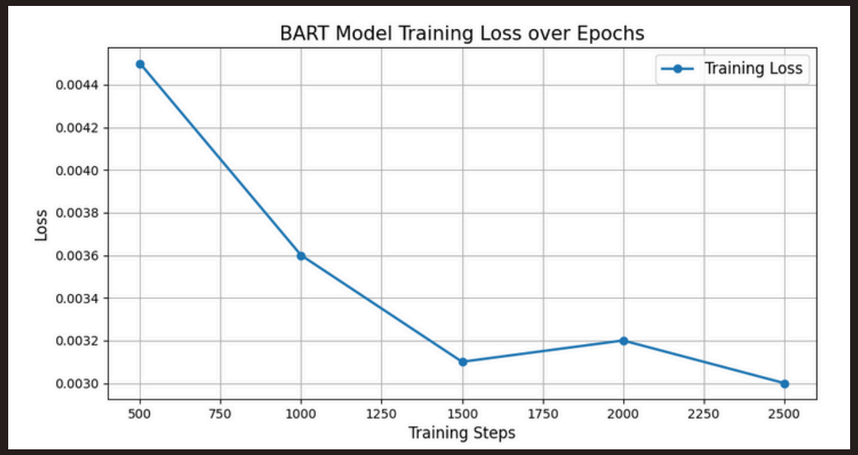
About Model & Training

- 'facebook/bart-base' pre-trained model
- Hugging Face Model Hub (Link to facebook/bart-base)
- Fine-tuned for seq2seq task conversion of biased text to unbiased text

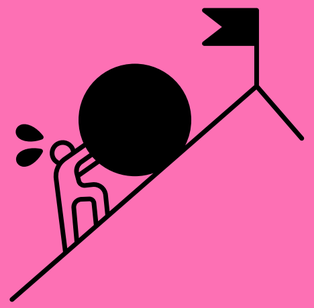
Model Configuration

- Gradient Accumulation Steps: 2
- FP16 Precision: Enabled
- Evaluation Strategy: Steps
- Load Best Model at End: Enabled
- Early Stopping: Yes (Patience: 3)
- Max Source Length: 1024
- Max Target Length: 128
- Learning Rate: 2e-5
- Weight Decay: 0.01

Performance



Challenges



- Long Training Time
- Resource Intensive
- Overfitting

BERT MODEL

About Model & Training

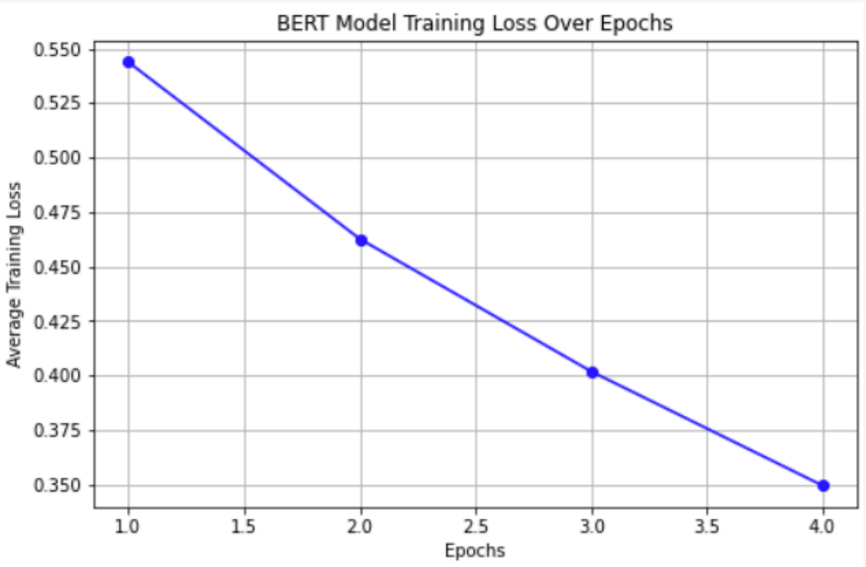
- 'bert-base-uncased' pre-trained model
- Hugging Face Model Hub (Link to bert-base-uncased)
- Fine-tuned for binary classification of biased/neutral texts

Model Configuration

- Inputs: Tokenized text data
- Outputs: Binary classification (biased/neutral)
- Used softmax function on the output layer to obtain probabilities for each class.
- Leveraged BERT's pre-trained word embeddings

Used to evaluate style and content across the **source**, **target**, and **BART-predicted** texts

Performance & Challenges

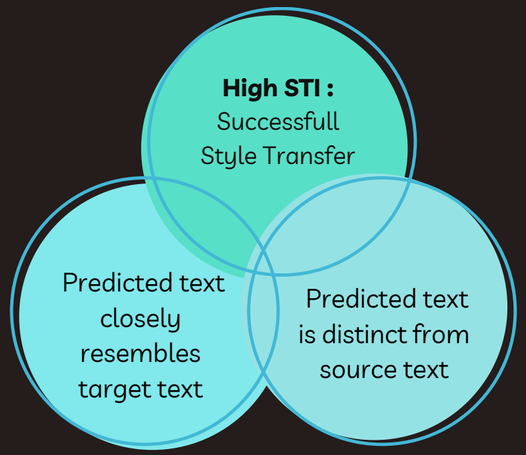


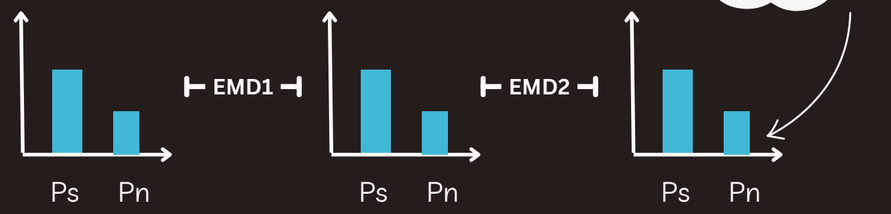
- Consistent Performance with test accuracy of 74.77%
- Balancing model complexity with content preservation

EVALUATION METRICS

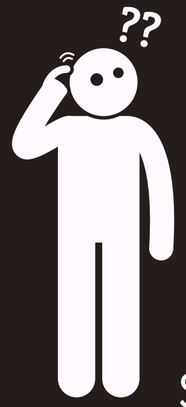
Style Transfer Intensity (STI):

measure style changes.





STI = EMD_SP + 1/(1+ EMD_ST)



Content Preservation Score (CPS):

assess content integrity.

Discraded

Tajmahal is a very beautiful place → **Tajmahal is a beautiful place**

VERY	0.7
Beautiful	0.6

Subjectivity Score

SentenceBert

[Source Embedding] [Predicted Embedding]

CPS = Cosine(Source, Predicted)

EVALUATION RESULTS

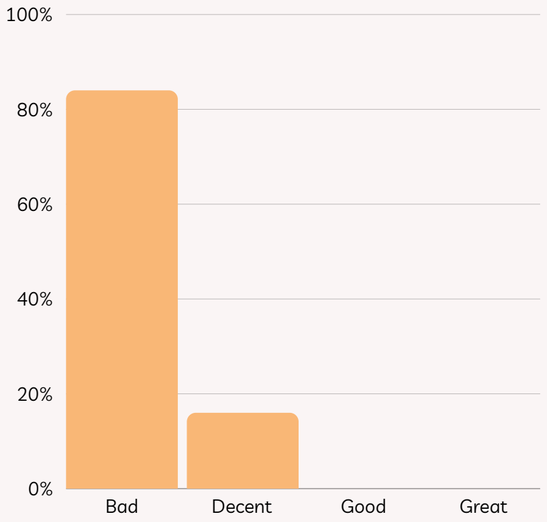
STI

source text : muzaffarabad is the capital of the pakistani territory of pakistan **occupied** kashmir.

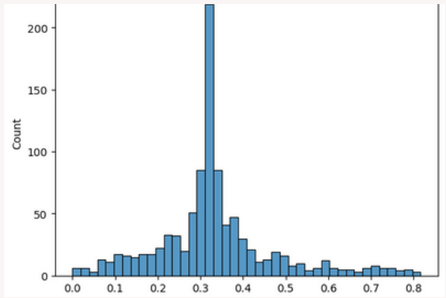
Target Text : muzaffarabad is the capital of the pakistani territory of pakistan **administered** kashmir.

Predicted Text : muzaffarabad is the capital of the pakistani territory of pakistan **administered** kashmir.

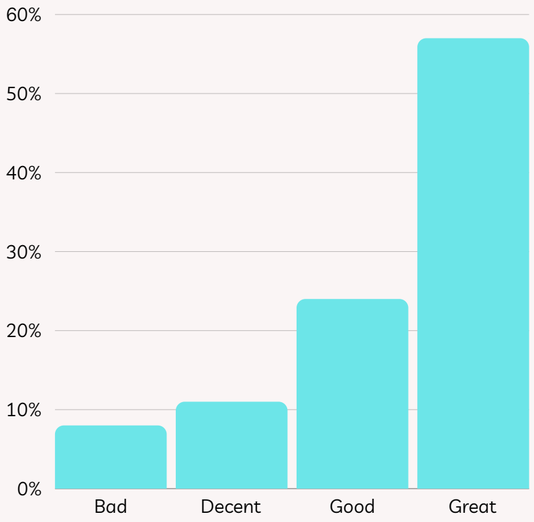
STI : 0.8166432930063067



Baseline



BART STI Distribution (train)

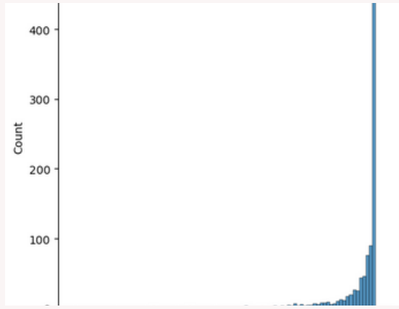


BART

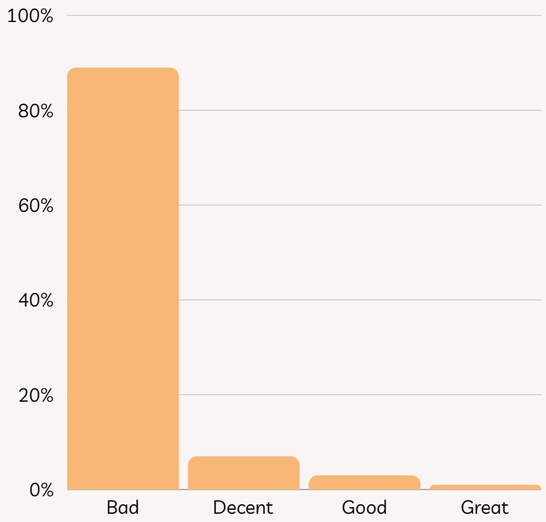
CPS

CPS : 0.99

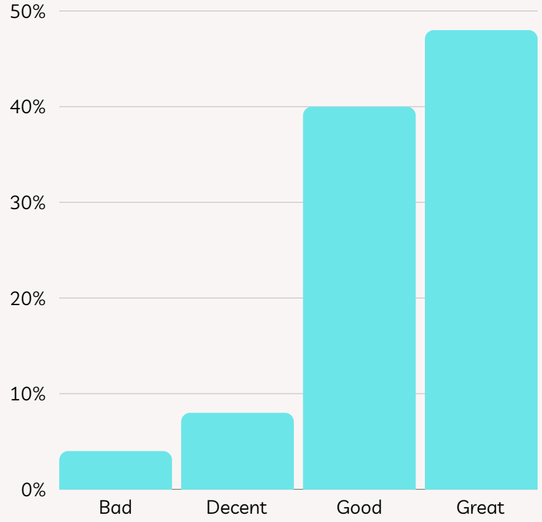
[CLS] pakistani territory of pakistan occupied kashmir . [SEP]



BART CPS Distribution (train)



Baseline



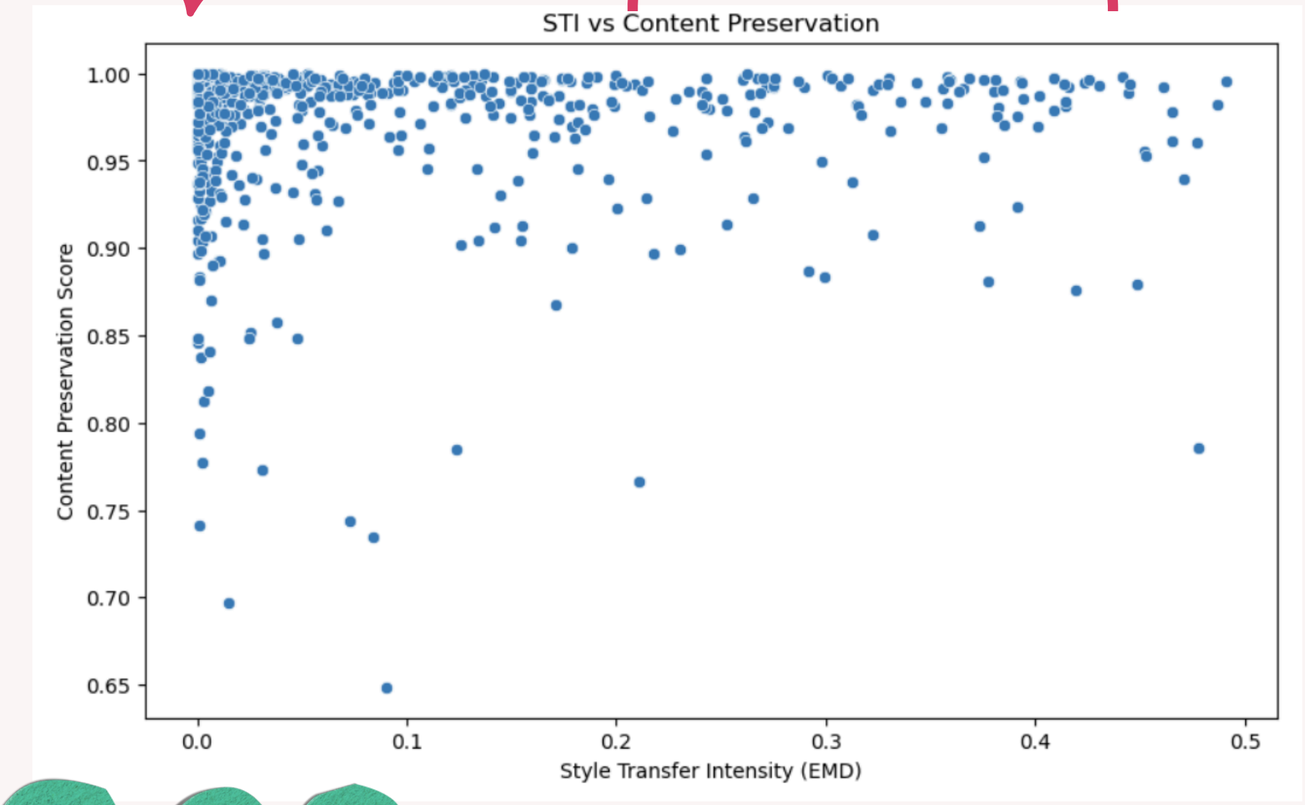
BART

EVALUATION ANALYSIS

Conservative style changes with good content preservation

Balance b/w style alteration & content integrity

STI: average values showing <u>moderate style shift.</u>



CPS: high on average (~0.981) showing <u>effective content maintenance.</u>

Correlation

EMD Source-Predicted ⟷ CPS

Weak -ve

Slight content loss with style shift

EMD Target-Predicted ⟷ CPS

Slight +ve

Better CP as style aligns with target

CONCLUSION & FUTURE SCOPE

Conclusion

- Fine-tuned BART model effectively transformed biased text to neutral.
- BART outperforms LSTM seq2seq model
- BERT and SentenceBERT used for CPS and STI evaluation metrics provide valuable quantitative insights
- Text style transfer for bias mitigation is complex but achievable

Future Scope

- Use full dataset instead of one-word edits for richer language understanding.
- Explore other evaluation metrics and methodologies for comprehensive performance analysis.
- Experiment with diverse pre-trained models or architectures for improved performance.

THANK YOU!