Sarvam Assignment : Cross-Lingual Sentence Embedding Alignment

Below is a detailed report on optimizing autoregressive models, with a focus on Parler TTS for latency and throughput improvement.

Index :

Model Used : bert-base-multilingual-cased; Dataset Used : cfilt/iitb-english-hindi

---

Part 1 :

Here, I used the **BERT-Base Multilingual Cased** model along with the **CFILT IITB English-Hindi parallel dataset**, which contains **1.6 million sentence pairs**.

PREPROCESSING

- Analyzed sentence length distribution for both English and Hindi sentences to guide tokenization and fine-tuning strategies.
- Separated English and Hindi sentences to extract sentence embeddings from the BERT model.

EMBEDDING EXTRACTION

- Used the **[CLS] token output** instead of averaging the last hidden states of all tokens. Both approaches were tested, the **[CLS] token yielded higher cosine similarity**, suggesting it may be a more effective method for this dataset.

EVALUATION & BASELINE METRICS

- Computed **cross-lingual semantic alignment metrics**, including **cosine similarity, retrieval accuracy, and Mean Reciprocal Rank (MRR)** to assess the baseline alignment between English and Hindi embeddings.
- Stored **sentence embeddings and metadata** for both languages and merged them into a single **TSV file** for visualization in TensorFlow's embedding projector.

OBSERVATIONS & VISUALIZATIONS

- **Figures 2 & 3** present **2D and 3D projections** of the sentence embeddings. As observed, **English and Hindi embeddings remain separate**, indicating that the model does not naturally align translations in the embedding space.
- Ideally, **parallel sentences (translations) should be closer together, regardless of language**, highlighting the need for further fine-tuning.

REFER TO THE FOLDER "WITHOUT FINETUNING" FOR TSV FILES AND EXTRACTED EMBEDDINGS. TO VISUALIZE EMBEDDINGS IN 3D, VISIT TensorFlow Projector AND UPLOAD THE TSV FILES CONTAINING VECTORS AND METADATA.

```python
metrics = compute_alignment_metrics(en_embeddings, hi_embeddings)

print("Baseline Analysis Results:")

print(f"\nMean Cosine Similarity: {metrics['mean_similarity']:.4f}")
print(f"Retrieval Accuracy: {metrics['retrieval_accuracy']:.4f}")
print(f"Mean Reciprocal Rank: {metrics['mean_reciprocal_rank']:.4f}")
```

```
Baseline Analysis Results:

Mean Cosine Similarity: 0.7160
Retrieval Accuracy: 0.2585
Mean Reciprocal Rank: 0.3414
```

Fig 1 : Cosine Similarity & Accuracy for Embeddings without Fine Tuning

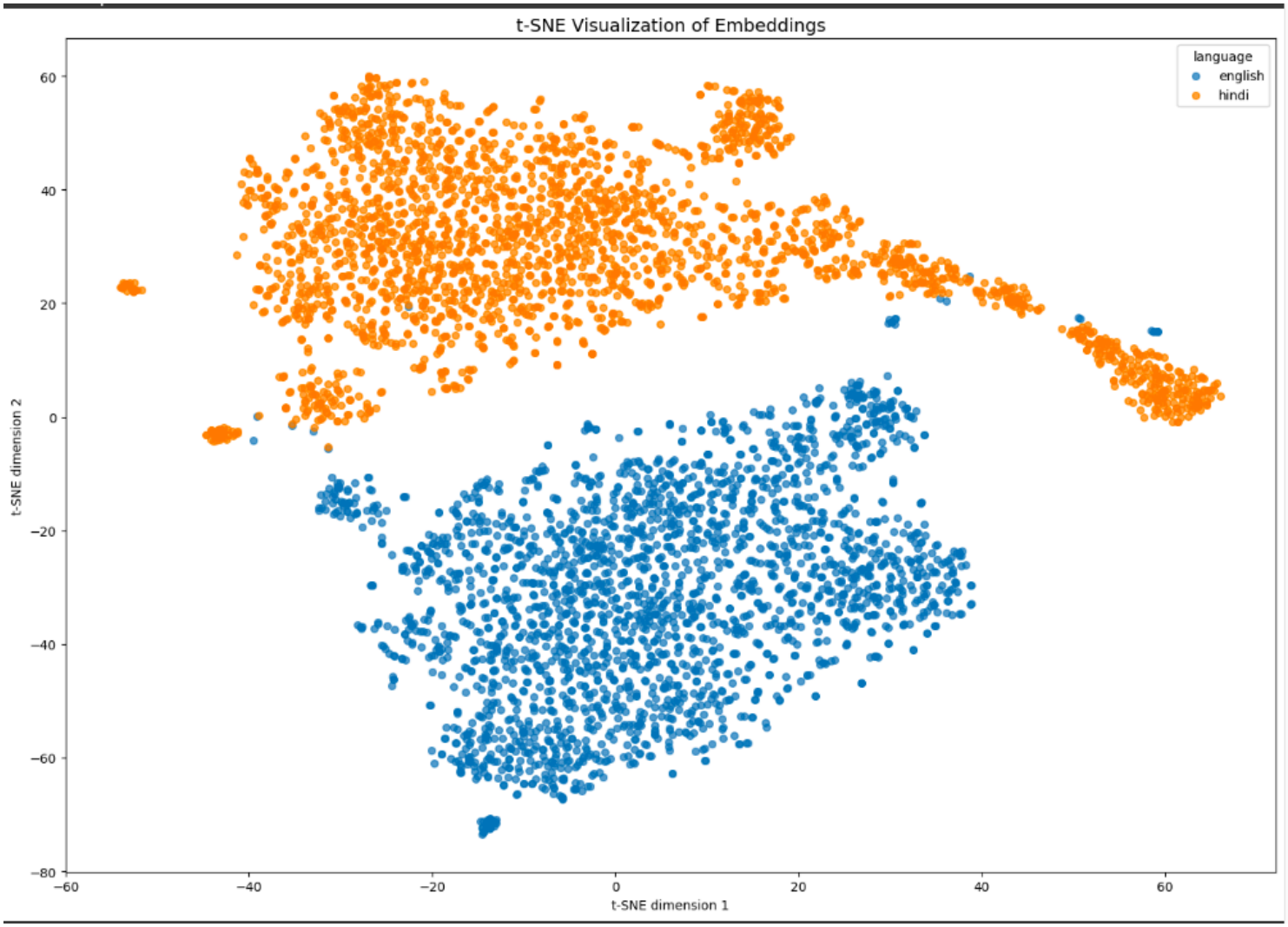

Fig 2 : Embeddings Projected in 2D

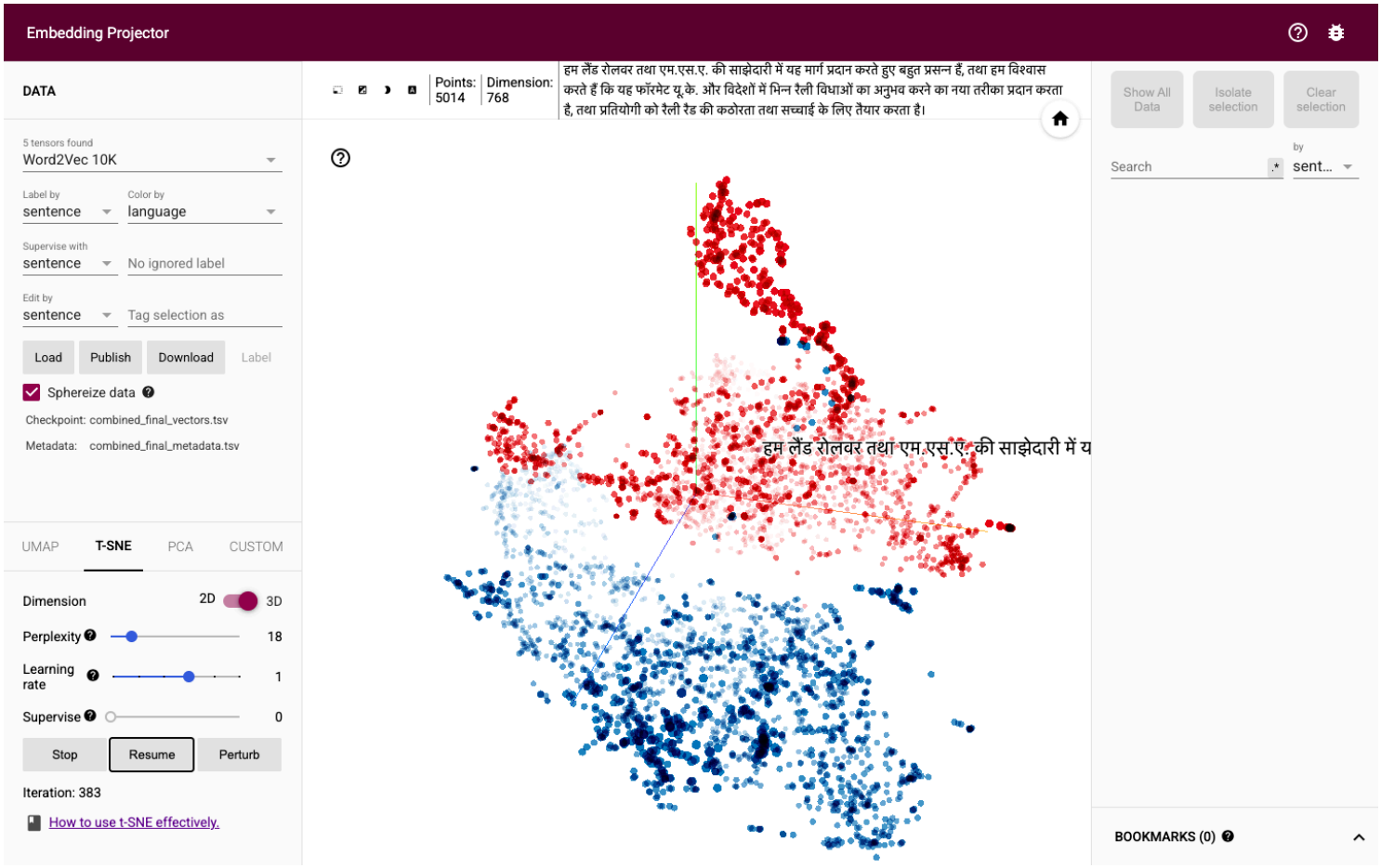

Fig 3 : Embeddings Projected in 3D

Part 2 & 3 :

To align **English and Hindi embeddings**, I referred to the **Sentence-BERT** paper (Reimers & Gurevych, 2019), which achieved **state-of-the-art (SOTA) performance** in fine-tuning BERT for sentence embeddings. Given its widespread adoption, I implemented a similar approach for cross-lingual alignment.

FINE-TUNING APPROACH

- Employed a **classification objective function**: $o=\text{softmax}(W_t(u,v,|u-v|))$ where **u** represents the **English** sentence embedding and **v** represents the **Hindi** sentence embedding. **Refer Fig 4.**
- Added an **additional classification layer** on top of BERT to predict whether a given **English-Hindi sentence pair** is a correct translation (**label: 1**) or not (**label: 0**).
- Generated **negative pairs** by randomly selecting a **Hindi sentence** for a given **English sentence**, ensuring a **balanced dataset** of **equal positive and negative pairs**.

TRAINING DETAILS

- Used **100,000 sentence pairs** (50,000 positive & 50,000 negative) due to GPU constraint.
- **Training setup: 3 epochs, Adam optimizer, Cross-Entropy Loss, Batch Size 32**

RESULTS

- Figure 5 shows the average validation loss of **0.1906** and an overall accuracy of **0.9202**.
- The model achieved a **positive class accuracy of 0.9654** and a **negative class accuracy of 0.8750**.
- The fine-tuned model effectively captured cross-lingual context, bringing English and Hindi sentence embeddings closer together.
- This alignment improved both classification accuracy and embedding quality, reinforcing the effectiveness of fine-tuning.

EMBEDDING PROJECTION & VISUALIZATION

- Saved fine-tuned embeddings along with metadata and merged them into a TSV file for visualization in TensorFlow's Embedding Projector.
- Figure 6 shows the 2D t-SNE embedding after fine-tuning Cross-Lingual BERT, where English and Hindi sentences are no longer separate.
- The embeddings are clustered together, indicating improved alignment across languages.
- Figure 7 presents cosine similarity and MRR metrics, showing an increase in all three metrics.
- Cosine similarity between parallel sentence embeddings is 0.8977, and MRR is 0.5202, validating the effectiveness of fine-tuning.
- Figure 8 shows the t-SNE projection using combined_ft_metadata.tsv and combined_ft_vectors.tfv, confirming that English and Hindi embeddings are closer together after fine-tuning.

REFER TO THE FOLDER "WITH FINETUNING" FOR TSV FILES AND EXTRACTED EMBEDDINGS. TO VISUALIZE EMBEDDINGS IN 3D, VISIT TensorFlow Projector AND UPLOAD THE TSV FILES CONTAINING VECTORS AND METADATA.

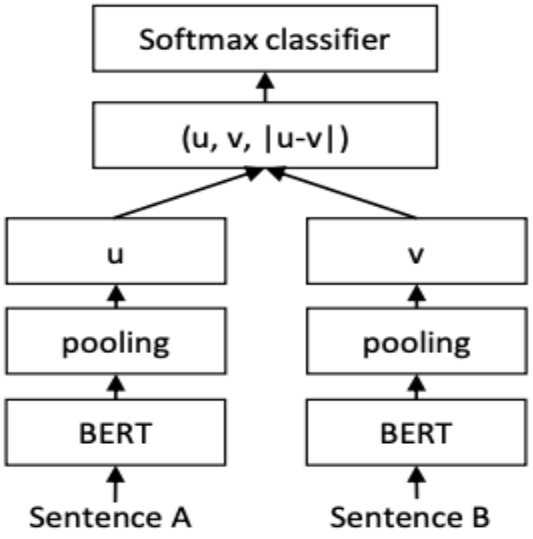

Figure 1: SBERT architecture with classification objective function, e.g., for fine-tuning on SNLI dataset. The two BERT networks have tied weights (siamese network structure).

Fig 4 : My Approach for Fine-Tuning BERT

```
Epoch 1/3
Training: 100%|██████████| 3125/3125 [1:05:02<00:00,  1.25s/it, loss=0.1619, acc=0.9089]
Training Loss: 0.2290, Accuracy: 0.9089
Evaluating: 100%|██████████| 33/33 [00:13<00:00,  2.37it/s, loss=0.4091, acc=0.9115]
Validation Metrics:
Loss: 0.1985
Accuracy: 0.9115
Positive Pair Accuracy: 0.9846
Negative Pair Accuracy: 0.8385
Saved new best model with validation accuracy: 0.9115

Epoch 2/3
Training: 100%|██████████| 3125/3125 [1:05:01<00:00,  1.25s/it, loss=0.0600, acc=0.9509]
Training Loss: 0.1327, Accuracy: 0.9509
Evaluating: 100%|██████████| 33/33 [00:14<00:00,  2.35it/s, loss=0.3644, acc=0.9202]
Validation Metrics:
Loss: 0.1906
Accuracy: 0.9202
Positive Pair Accuracy: 0.9654
Negative Pair Accuracy: 0.8750
Saved new best model with validation accuracy: 0.9202
```

Fig 5 : Results while Fine Tuning BERT

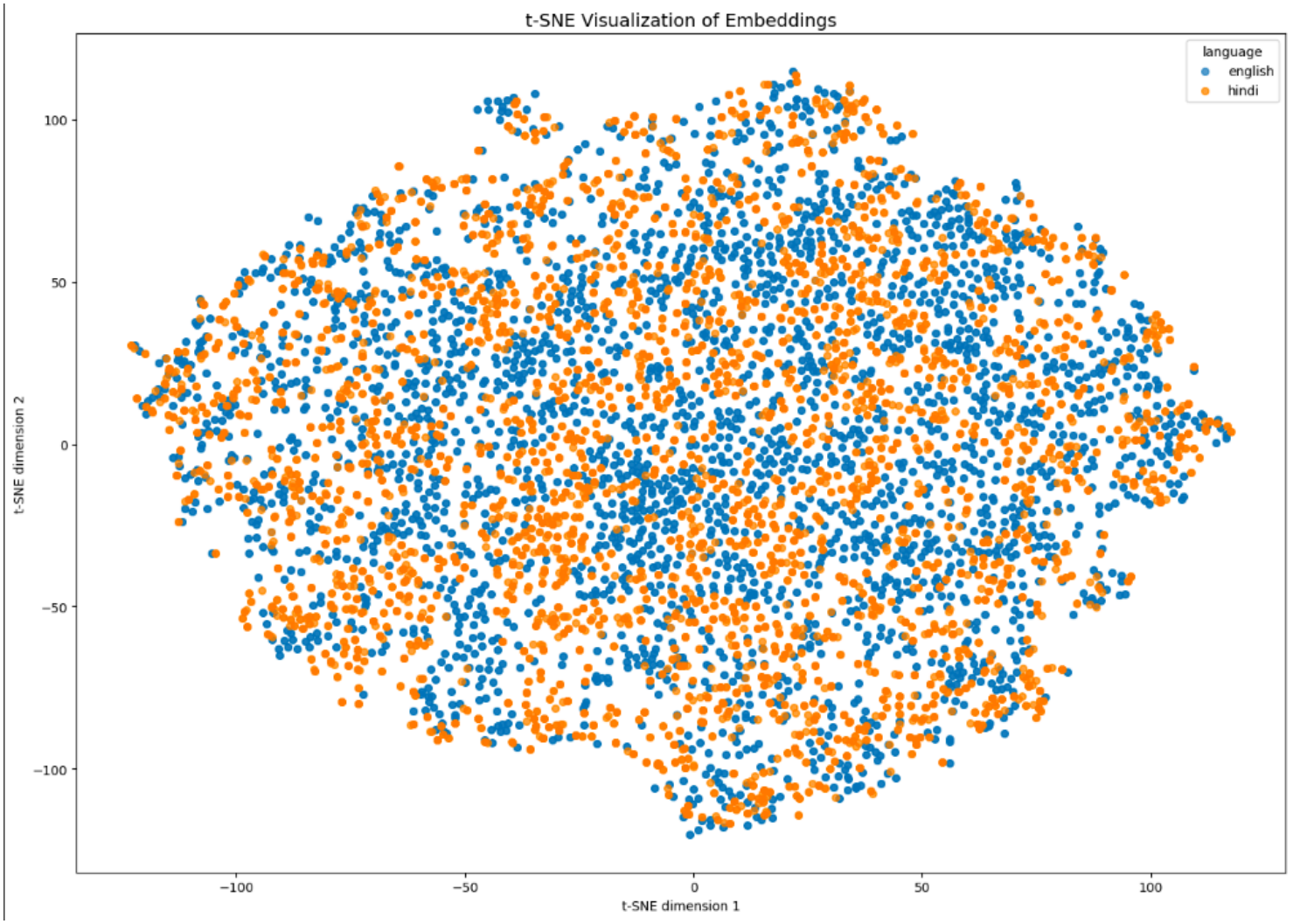

Fig 6 : Embeddings Projected using t-SNE using 2d after Fine Tuning

```
Baseline Analysis Results:

Mean Cosine Similarity: 0.8977
Retrieval Accuracy: 0.4164
Mean Reciprocal Rank: 0.5202
```

Fig 7 : Baseline Analysis Results after Fine Tuning

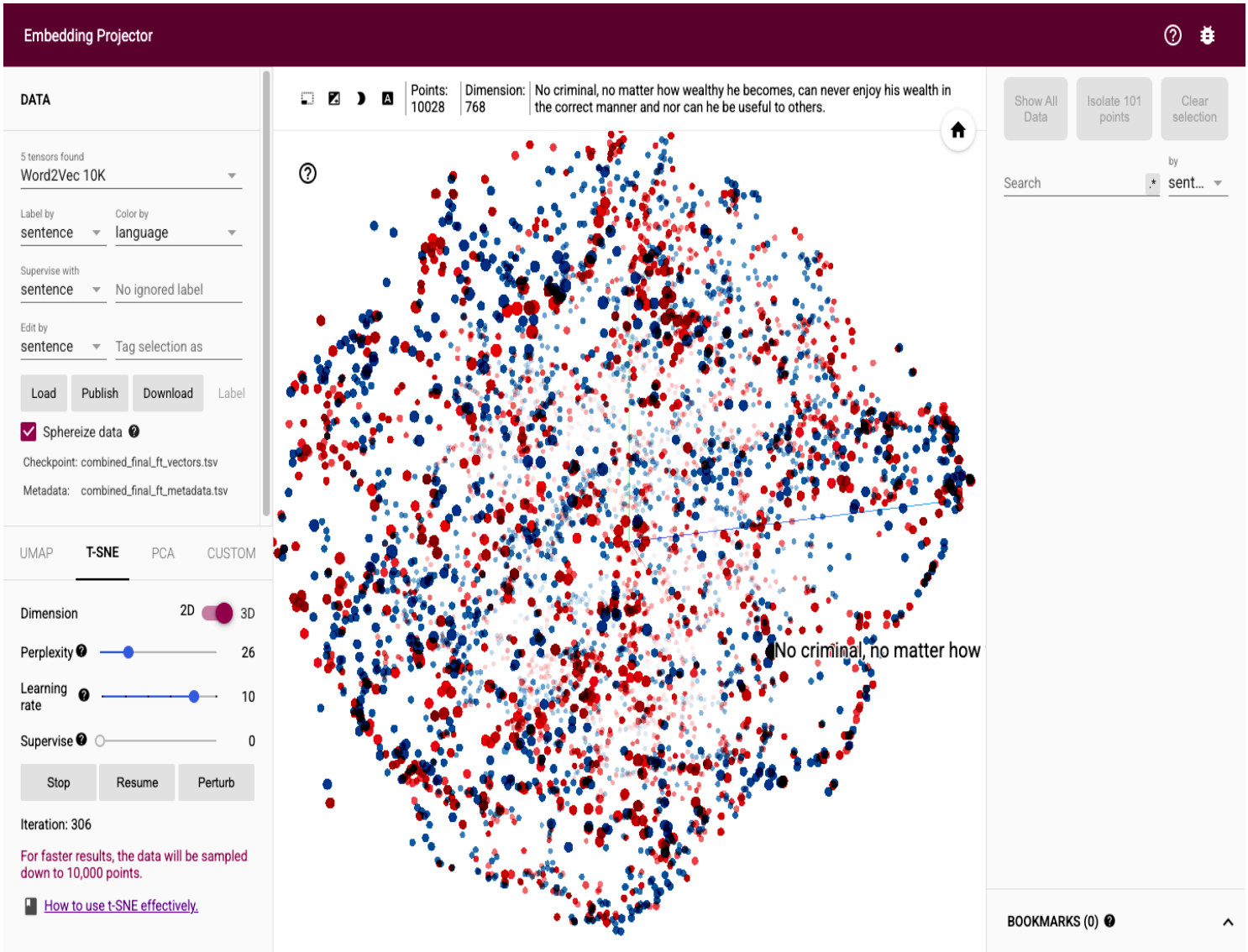

Fig 8 : Embeddings projected using t-SNE in 3d after Fine Tuning

VALIDATION OF RESULTS

- To confirm the effectiveness of fine-tuning, we analyze whether English and corresponding Hindi sentence embeddings are closer in the projected space.
- The embeddings are clustered together, indicating similarity between sentences across languages.
- Figures 9, 10, and 11 provide visual evidence, highlighting instances where Hindi sentence embeddings align closely with their English counterparts.
- Red arrows in each figure illustrate the improved alignment, demonstrating that the model successfully captures cross-lingual semantic relationships.

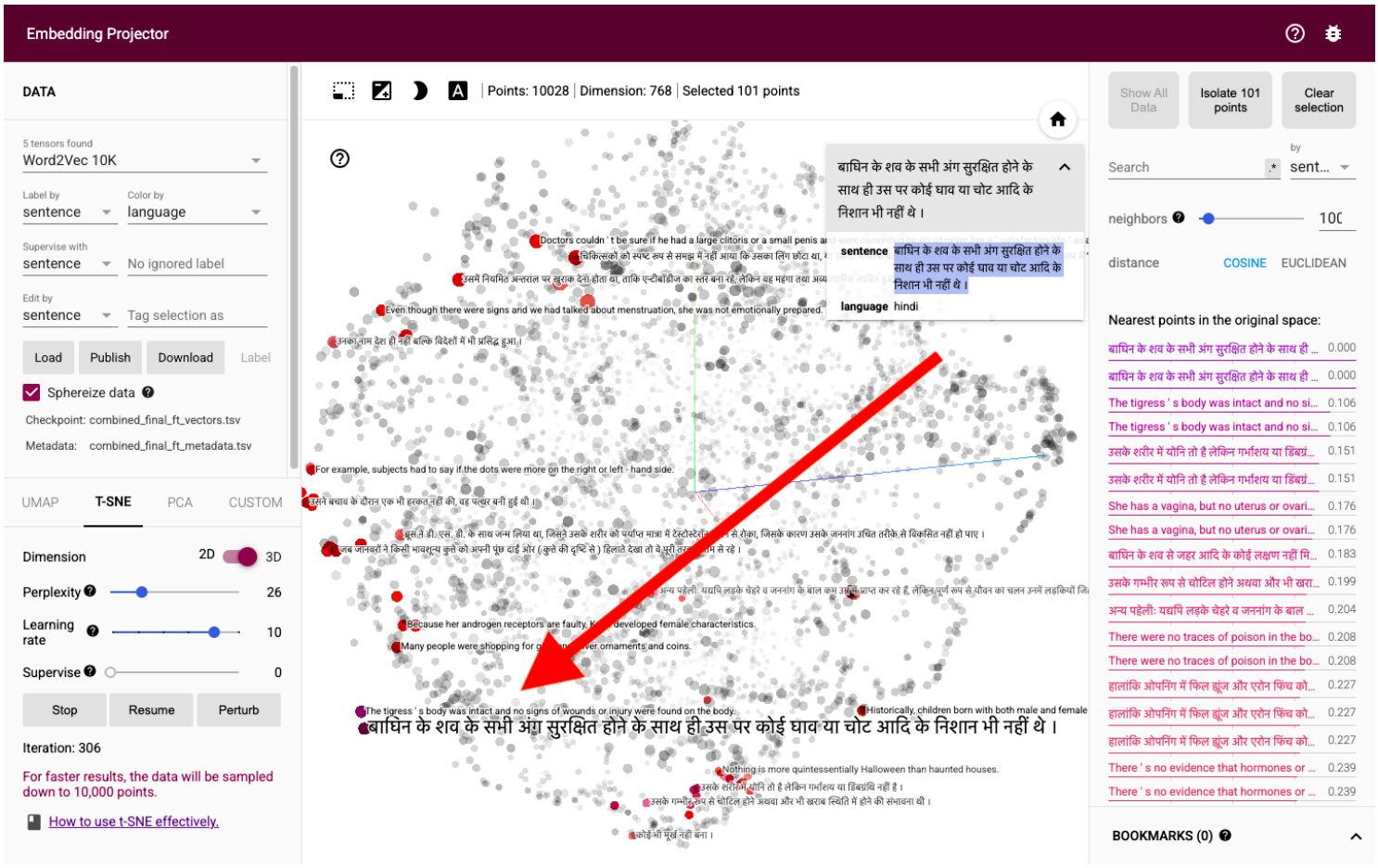

Fig 9 : (Ex. 1) Embeddings Projected using t-SNE using 3d after Fine Tuning

TO BE CONTINUED…

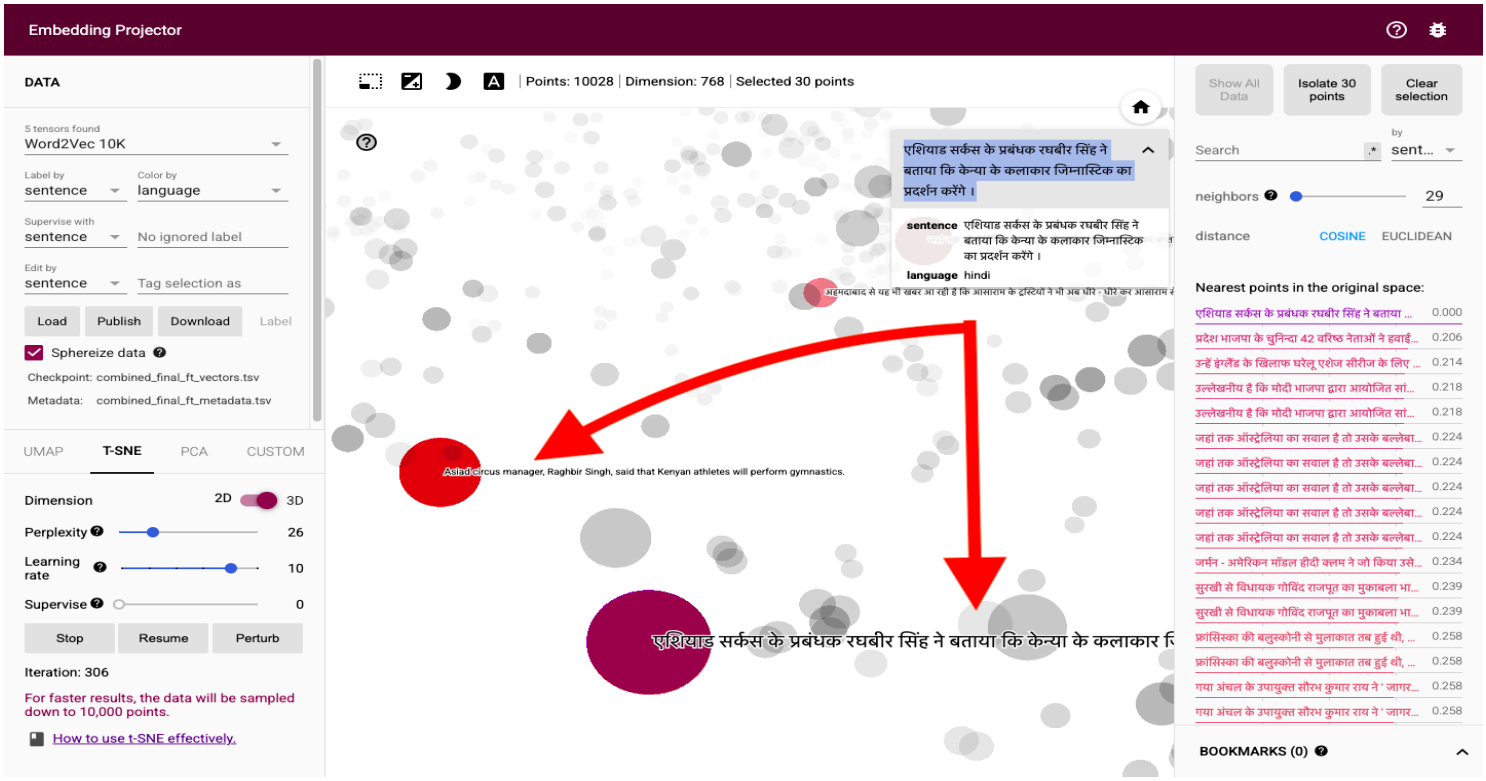

Fig 10 : (Ex. 2) Embeddings Projected using t-SNE using 3d after Fine Tuning

Fig 11 : (Ex. 3) Embeddings Projected using t-SNE using 3d after Fine Tuning

CONCLUSION

- Fine-tuning BERT with a classification objective function using an additional classification layer effectively improves cross-lingual embedding alignment.
- Significant improvements are observed in accuracy and cosine similarity, demonstrating the effectiveness of the approach.
- Embedding visualizations show that semantically similar sentences across languages are brought closer in the vector space.
- The fine-tuning process helps BERT generate more language-agnostic sentence embeddings, enhancing its multilingual representation capabilities.

FUTURE WORK

- With additional GPU resources, further improvements in loss and accuracy can be achieved by incorporating triplet loss and curating a dataset optimized for a regression objective.
- There is potential to leverage the full dataset and maximize GPU compute to enhance model performance further.
- Exploring Language-agnostic BERT Sentence Embedding (LaBSE) could be a promising direction for improving cross-lingual embedding alignment.

I really enjoyed working on this project. Thanks for the opportunity! Reach out at vivekvkashyap10@gmail.com for any questions. Happy to help.

Thanks.