Semantic Textual Relatedness

CS779A Group Project
Prof. Ashutosh Modi

Group 3
Rajarshi Dutta (200762)
Udvas Basak (201056)
Shivam Pandey (200938)

PROBLEM STATEMENTS

Similarity and relatedness are dissimilar !!!

- Lets take some examples :

SBERT-cosine similarity

S1: He has a heart of gold. S2: His kindness knows no bounds.	0.38
S1: Time and tide wait for no man. S2: "Opportunities can be fleeting if not seized promptly.	0.20

PROBLEM STATEMENTS

TRACK A - SUPERVISED

- Objective: Train systems to predict semantic textual relatedness (STR) in labelled sentence pairs
- Data: Annotated scores (0-1) in available languages, ranked in decreasing order of score.
- Challenge: Distinguish relatedness from similarity
- Use Cases: Sentence representation evaluation, multilingual content recommendation
- Data: Released for select languages (Amharic, English, Marathi, Telugu)
- Evaluation: Spearman rank correlation

PROBLEM STATEMENTS

TRACK B - UNSUPERVISED

- Objective: Develop STR systems without labelled data
- Need: Due to limited annotated data, unsupervised models offer a scalable approach for semantic relatedness in multiple languages
- Data: Need to **create** unigram or bigram relatedness datasets from any language
- Challenge: Build models from scratch
- Use Cases: Cross-lingual search, historical text analysis
- Evaluation: Spearman rank correlation

PROBLEM STATEMENTS

TRACK C - CROSS-LINGUAL

- Objective: Create STR systems for languages lacking target data
- Need: Crucial for languages with limited available training data, like Kinyarwanda
- Data: Choose a source and target language and rely on labelled data from the source to develop models for the target
- Challenge: Train models without target language data
- Use Cases: Enhance machine translation, analyze low-resource languages
- Evaluation: Spearman rank correlation

LITERATURE REVIEWS

1. Semantic textual similarity for modern standard and dialectal Arabic using transfer learning - Mansour Al Sulaiman, Abdullah M. Moussa

Refer to this link: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0272991

LITERATURE REVIEWS

1. Semantic textual similarity for modern standard and dialectal Arabic using transfer learning - Mansour Al Sulaiman, Abdullah M. Moussa

Interleaving English STS data with Arabic BERT models via transfer learning, notably improving model accuracy.

Refer to this link: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0272991

LITERATURE REVIEWS

1. Semantic textual similarity for modern standard and dialectal Arabic using transfer learning - Mansour Al Sulaiman, Abdullah M. Moussa

Interleaving English STS data with Arabic BERT models via transfer learning, notably improving model accuracy.

Using knowledge distillation-based STS models, fine-tuning them with translated data which uses a Siamese architecture for dialects orientation

Refer to this link: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0272991

LITERATURE REVIEWS

1. Semantic textual similarity for modern standard and dialectal Arabic using transfer learning - Mansour Al Sulaiman, Abdullah M. Moussa

Interleaving English STS data with Arabic BERT models via transfer learning, notably improving model accuracy.	Using knowledge distillation-based STS models, fine-tuning them with translated data which uses a Siamese architecture for dialects orientation	Converting ArabicBERT to sBERT and finetuning the trained model on SnLI datasets.

Refer to this link: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0272991

LITERATURE REVIEWS (CONTD..)

2. Google Similarity Index

Refer to this link: https://arxiv.org/pdf/cs/0412098.pdf

LITERATURE REVIEWS (CONTD..)

2. Google Similarity Index

This is a new semantic distance measure. It's based on the number of hits returned by the Google search engine for a given set of words or phrases.

The Normalized Compression distance calculates the distance between the Kolmogorov compressed version of the two strings

Refer to this link: https://arxiv.org/pdf/cs/0412098.pdf

LITERATURE REVIEWS (CONTD..)

2. Google Similarity Index

This is a new semantic distance measure. It's based on the number of hits returned by the Google search engine for a given set of words or phrases.

The Normalized Compression distance calculates the distance between the Kolmogorov compressed version of the two strings

$$NGD(x,y) = \frac{\max\{\log(f(x)), \log(f(y))\} - f(x,y)}{\log(N) - \min\{\log(f(x)), \log(f(y))\}}$$

$$NCD = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}$$

Refer to this link: https://arxiv.org/pdf/cs/0412098.pdf

DATA – SUPERVISED TRACK

Annotating a new dataset is challenging due to the vague definition of relatedness

Using larger standard similarity datasets is impractical because of the distinctions between similarity and relatedness

	Training set	Dev Set
English	250	5500
Amharic	496	95
Marathi	1200	600
Telugu	1170	260

DATA – UNSUPERVISED

Dataset Generation

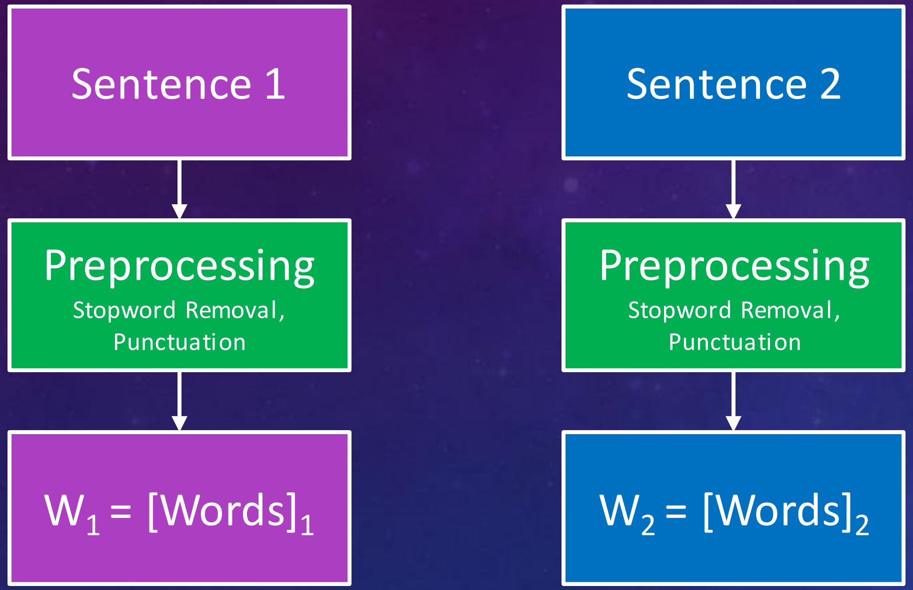
1. Select Relevant Corpus
2. Extract Bigrams
3. Count Bigram frequencies
4. Perform Negative Sampling

Model Training

1. Train Word Embeddings
2. Bigram Embeddings
3. Utilize cosine similarity
4. Apply Hierarchical Clustering
5. Set a Similarity Thresholding

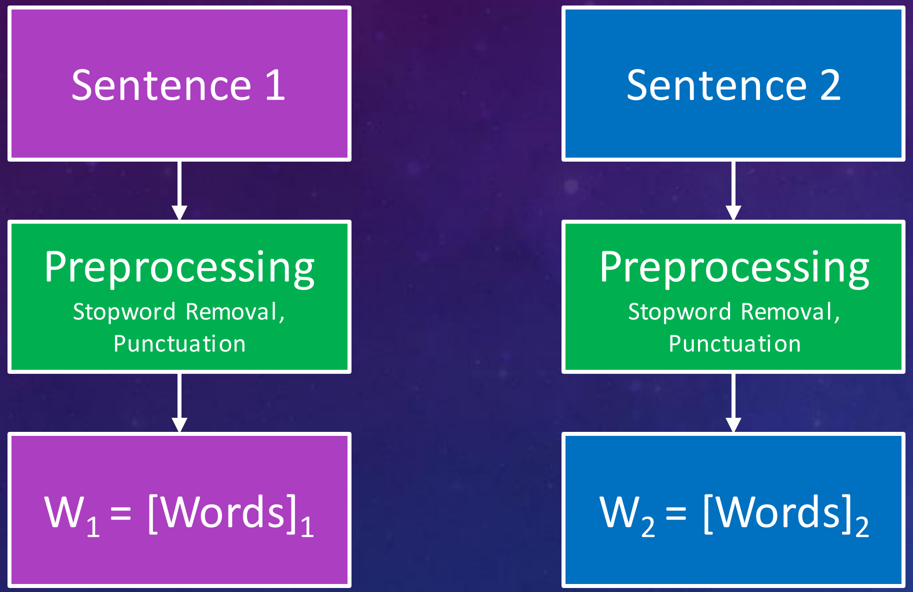
APPROACH TO TASK 1

- Without Transformers



APPROACH TO TASK 1

- Without Transformers

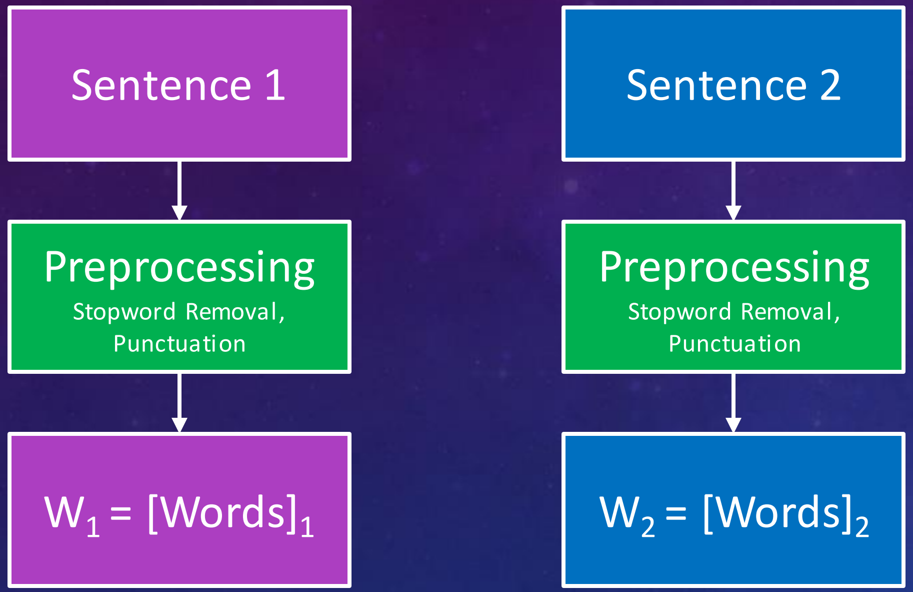


$$\text{Dice Coeff(DSC)} = \frac{2|W_1 \cap W_2|}{|W_1| + |W_2|}$$

$$\text{Jaccard Coeff}\ (J(W_1, W_2)) = \frac{|W_1 \cap W_2|}{|W_1 \cup W_2|}$$

APPROACH TO TASK 1

❖ Without Transformers

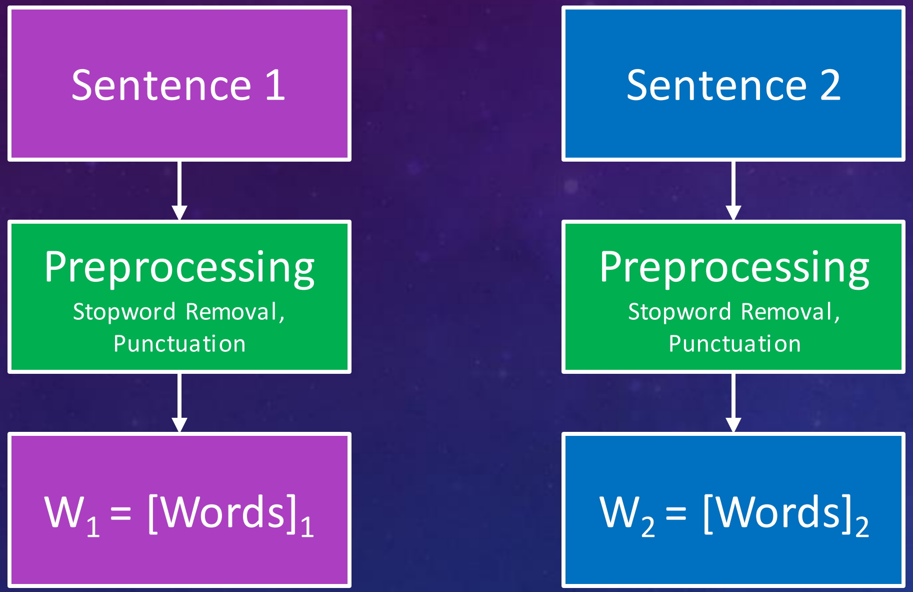


Metrics involving Large Corpus

- ❖ For two terms, find the number of documents that have T_1 and have T_2, and have them together.
- ❖ Using this(and similar measures), we can calculate some similarity metrics
- ❖ Normalized Google Distance
- ❖ Revision Info
  - ❖ and lots moreWikipedia.

APPROACH TO TASK 1

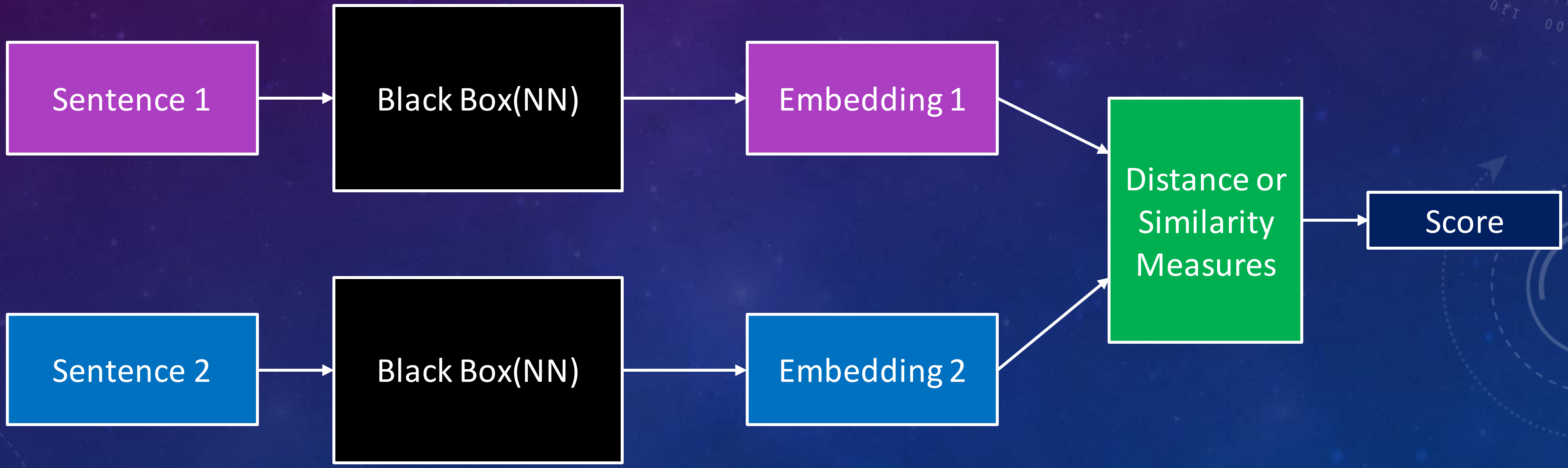
❖ Without Transformers



Using WordNet, we can also check for similar words, and then, broaden the metrics.

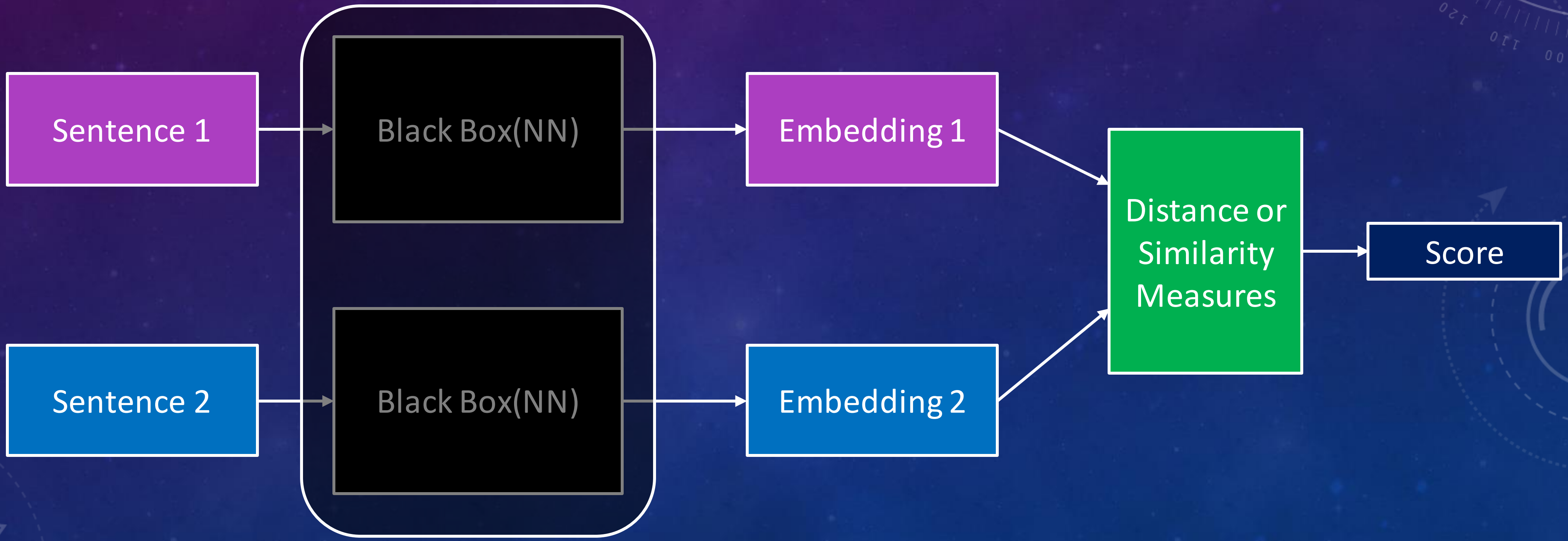
APPROACH TO TASK 1

❖ With Transformers



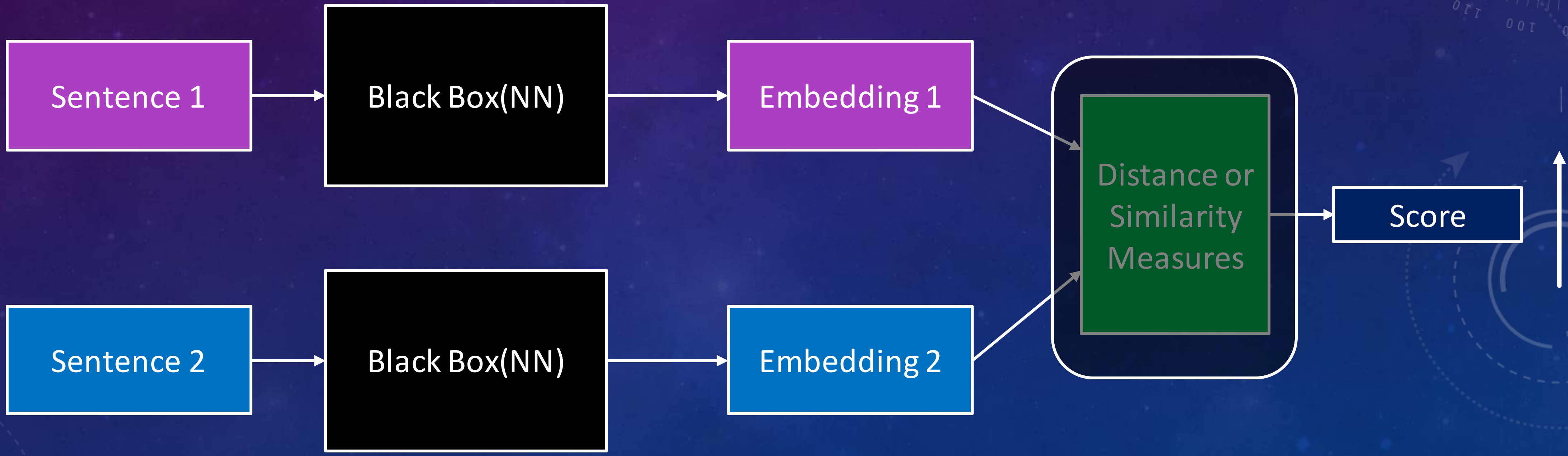
APPROACH TO TASK 1

❖ With Transformers



APPROACH TO TASK 1

❖ With Transformers



APPROACH TO TASK 1

❖ With Transformers

Black Box(NN)

- sBERT
- Universal Sentence Encoder

APPROACH TO TASK 1

❖ With Transformers

Black Box(NN)

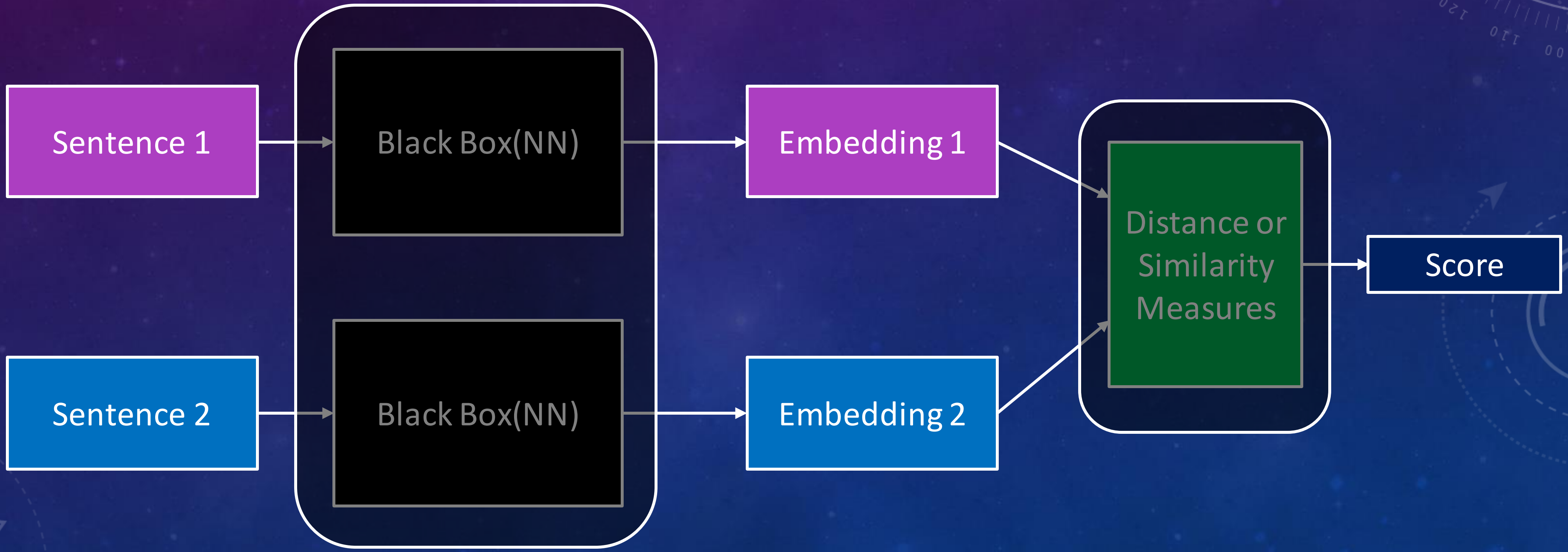
Distance or Similarity Measures

- sBERT
- Universal Sentence Encoder

- Cosine Similarity
- Euclidean Distance
- Manhattan Distance
- Mahalanobis Distance
- COSMIC(COmbining Various Similarity MeasurEs for Cosine similarity)

APPROACH TO TASK 1

- ❖ With Transformers
  - ❖ End-to-end (Siamese Architecture)

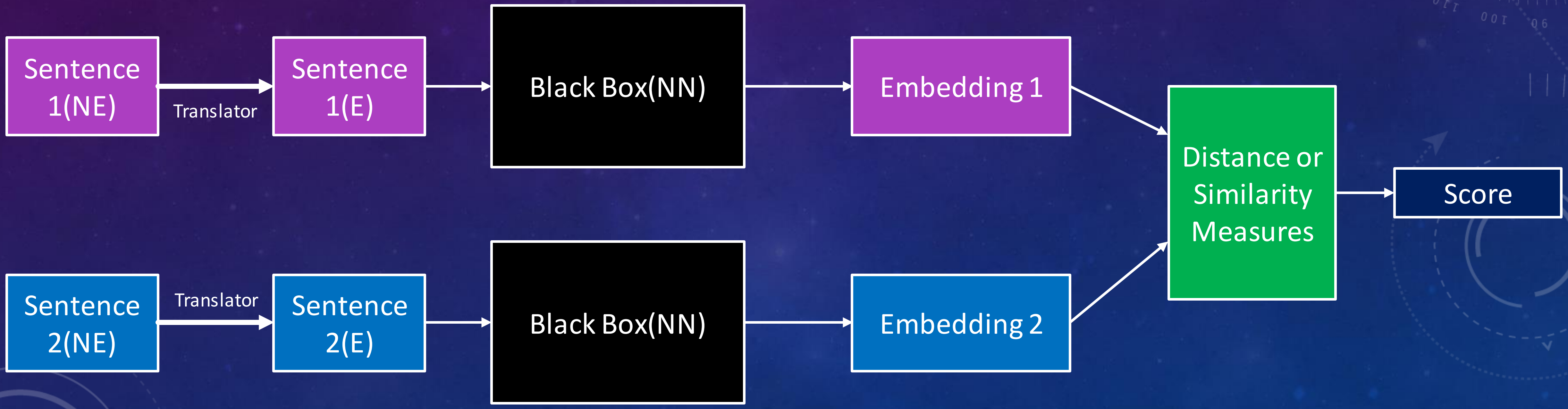


APPROACH TO TASK 1

- ❖ With Transformers
  - ❖ End-to-end (Siamese Architecture)

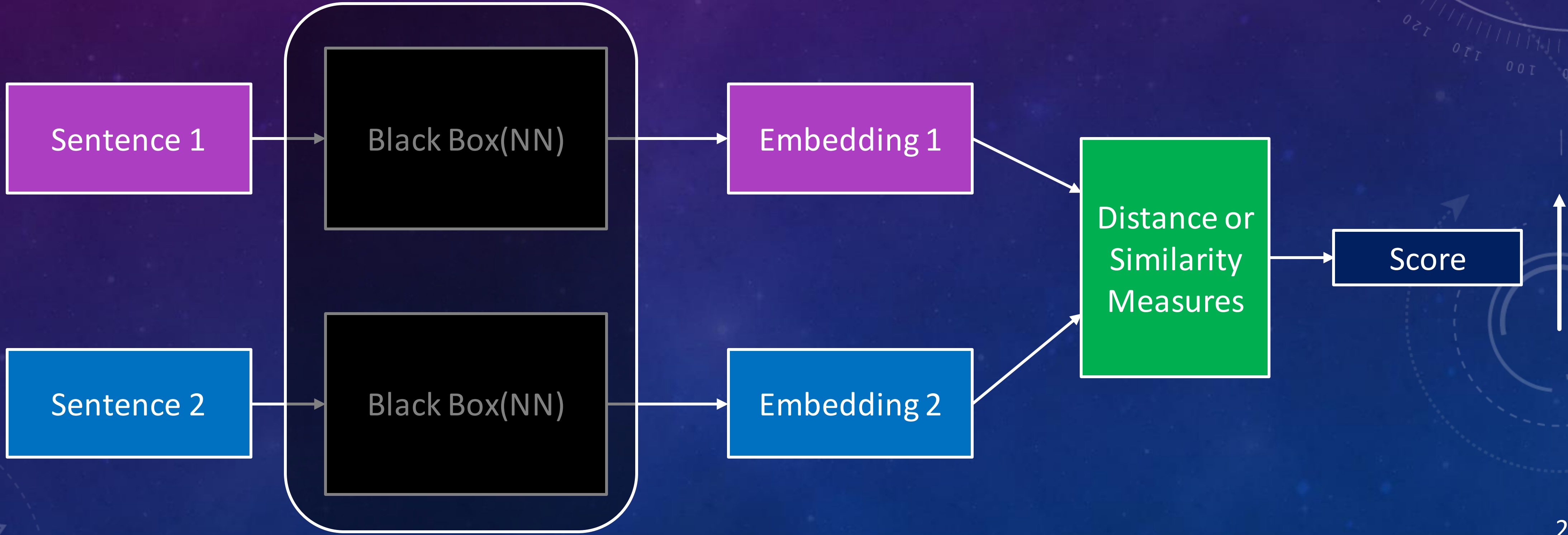
APPROACH TO TASK 1

- ❖ With Transformers
  - ❖ To handle other languages



APPROACH TO TASK 1

- ❖ With Transformers
  - ❖ To handle other languages



Train with other data in other languages

TASK 1

❖ Some results (All done on English 5500 train set)

Scoring Technique	Preprocessing	Spearman Rank Correlation Coefficient
Dice Coefficient	**NA**	**0.58**
Dice Coefficient	Stopwords Removed	0.56
Dice Coefficient	Duplicate words removed	0.58
Dice Coefficient	TF-IDF Weighting	0.36
Jaccard Coefficient	All cases	0.57-0.58
Cosine Similarity	Training using distilbert-base-nli-mean-tokens	0.81

TASK 1

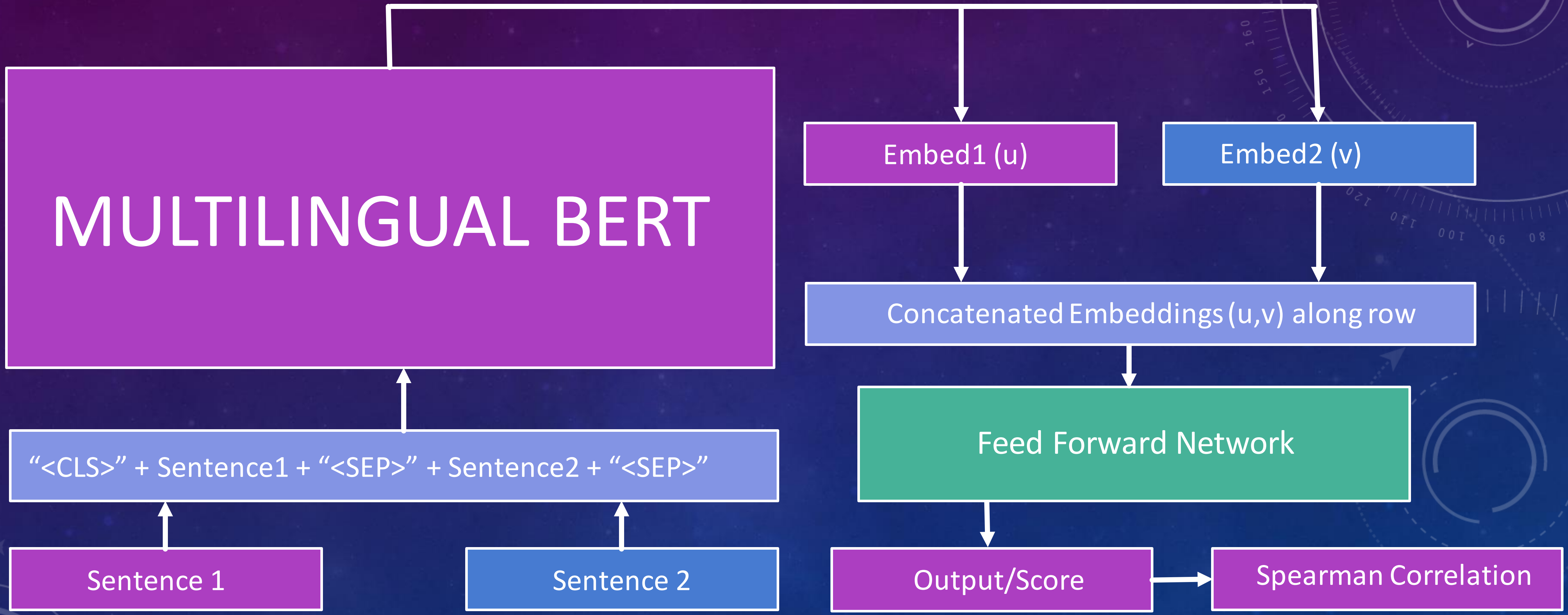
❖Some results (All done on English 5500 train set)

Scoring Technique	Preprocessing	Spearman Rank Correlation Coefficient
Dice Coefficient	**NA**	**0.58**
Cosine Similarity	**Pre-trained all-mpnet-base-v2**	**0.82-0.83**
All metrics	Pre-trained all-MiniLM-L6-v2	0.8 – 0.82
Cosine Similarity	Training using distilbert-base-nli-mean-tokens *	0.81

After all these tests, the target would be to produce an architecture that will make the Spearman Rank Correlation Coefficient more than 0.83.

*Trained on eng_train, and validated on eng_dev

MODEL ARCHITECTURE



PROPOSED TIMELINE

WEEK 1
Pilot Dataset Review, Basic Preprocessing

WEEK 2
Thorough Literature Review

WEEK 3
Evaluation Metric Decisions, Buffer for LitRev

WEEK 4
Initial Models for Baselining

WEEK 5-6
Experimentation and Model Refinement

WEEK 7
MultiLingual Extensions, Transfer Learning

WEEK 8
Buffer for Finishing on Models

WEEK 9
Comparative Analysis of Datasets

WEEK 10
Final Model Training and Evaluation

WEEK 11
Documentation and Presentation

THANK YOU!