ROSETTA STONE

PRESENTED BY ELENA DEL ROSAL

DATASET STRUCTURE

	sentence1	sentence2	score	lang1	lang2
0	Ein Flugzeug hebt gerade ab.	An air plane is taking off.	5.0	de	en
1	Ein Flugzeug hebt gerade ab.	Un aviÃ³n estÃ¡ despegando.	5.0	de	es
2	Ein Flugzeug hebt gerade ab.	Un avion est en train de dÃ©coller.	5.0	de	fr
3	Ein Flugzeug hebt gerade ab.	Un aereo sta decollando.	5.0	de	it
4	Ein Flugzeug hebt gerade ab.	é£□è¡□æ©□ã□□é□¢é□¸ã□□ã□¾ã□□ã□□	5.0	de	ja
...	...	...	...	...	...
949075	é□©å□½å®£å¸□MERSç□«æ□□ç»□æ□□	å□□æ□□é©¬ã□®ä»£è¡¨å□£ã□□é□□å□½æ□¿å°□é□¢ä¿□è□□ã...	0.0	zh	ja
949076	é□©å□½å®£å¸□MERSç□«æ□□ç»□æ□□	Delegatie Noord-Korea Ontmoetingen met Zuid-Ko...	0.0	zh	nl
949077	é□©å□½å®£å¸□MERSç□«æ□□ç»□æ□□	Delegacja Korei PÃ³Å□nocnej spotyka siÄ□ z urz...	0.0	zh	pl
949078	é□©å□½å®£å¸□MERSç□«æ□□ç»□æ□□	DelegaÃ§Ã£o da Coreia do Norte reÃºne-se com f...	0.0	zh	pt
949079	é□©å□½å®£å¸□MERSç□«æ□□ç»□æ□□	Ð□ÐµÐ»ÐµÐ³Ð°Ñ□Ð¸Ñ□ Ð¡ÐµÐ²ÐµÑ□Ð½Ð¾Ð¹ Ð□Ð¾Ñ□ÐµÐ¸...	0.0	zh	ru

LIBRARIES USED

- pandas – For data manipulation and DataFrame operations.
- numpy – For mathematical operations, normalization, and array handling.
- sentence-transformers – To generate sentence embeddings using models like LaBSE.
- torch (PyTorch) – Backend framework used by sentence-transformers.
- scikit-learn (sklearn) – For:
  - Evaluation metrics (accuracy, MSE)
  - Confusion matrix and classification report
  - Optional scaling or normalization
- deep-translator – To automatically translate sentences using GoogleTranslator.
- Parrot – For generating paraphrased versions of English sentences.
- spaCy – To compute semantic similarity using pre-trained GloVe word embeddings.
- matplotlib – For creating scatter plots, bar charts, and data visualizations.
- seaborn (optional) – For visualizing the correlation matrix (e.g., heatmaps).
- time – To pause between translation attempts and handle retries safely.

MAIN PURPOSE:

analyzing the semantic similitude between the sentences

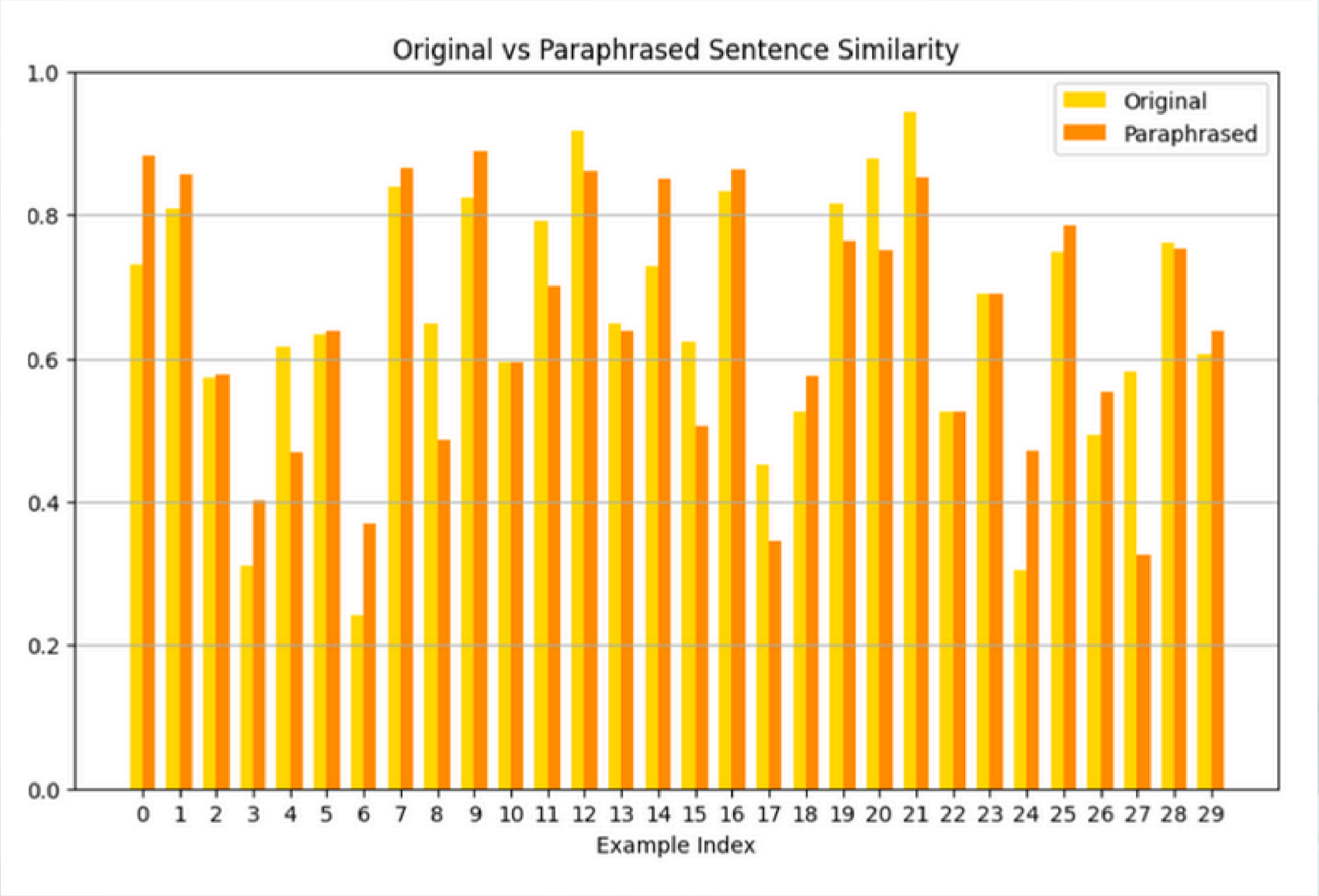
OBJECTIVE:

comparing two different models to see how they perform with the same task and to see their differences in a practical way.

PROCEDURE

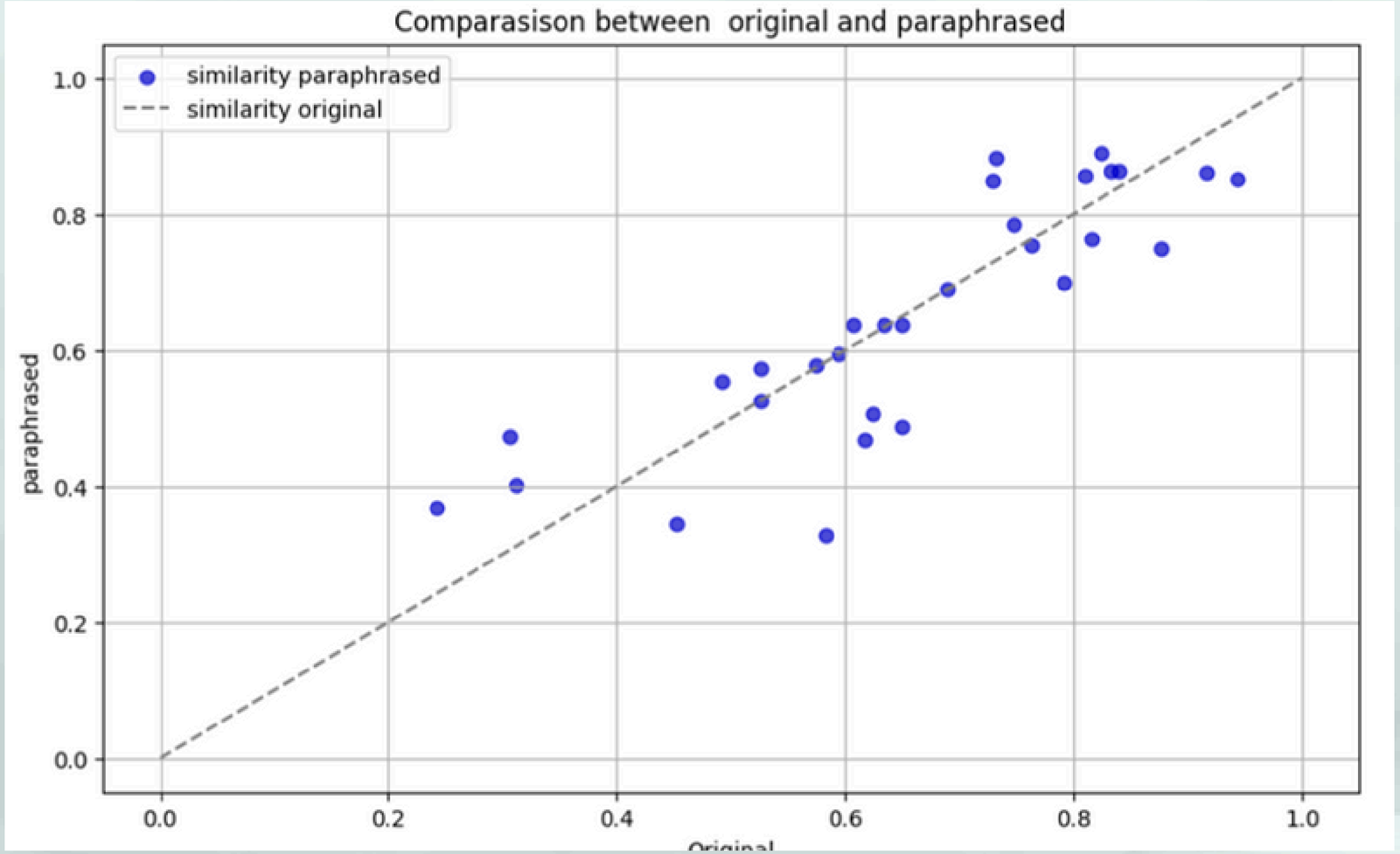
1. Data processing: Stemming
2. Translation
3. Data augmentation: Paraphrasing
4. Transformers model
5. SpaCy model

IS PARAPHRASING AS IMPORTANT AS WE THINK?

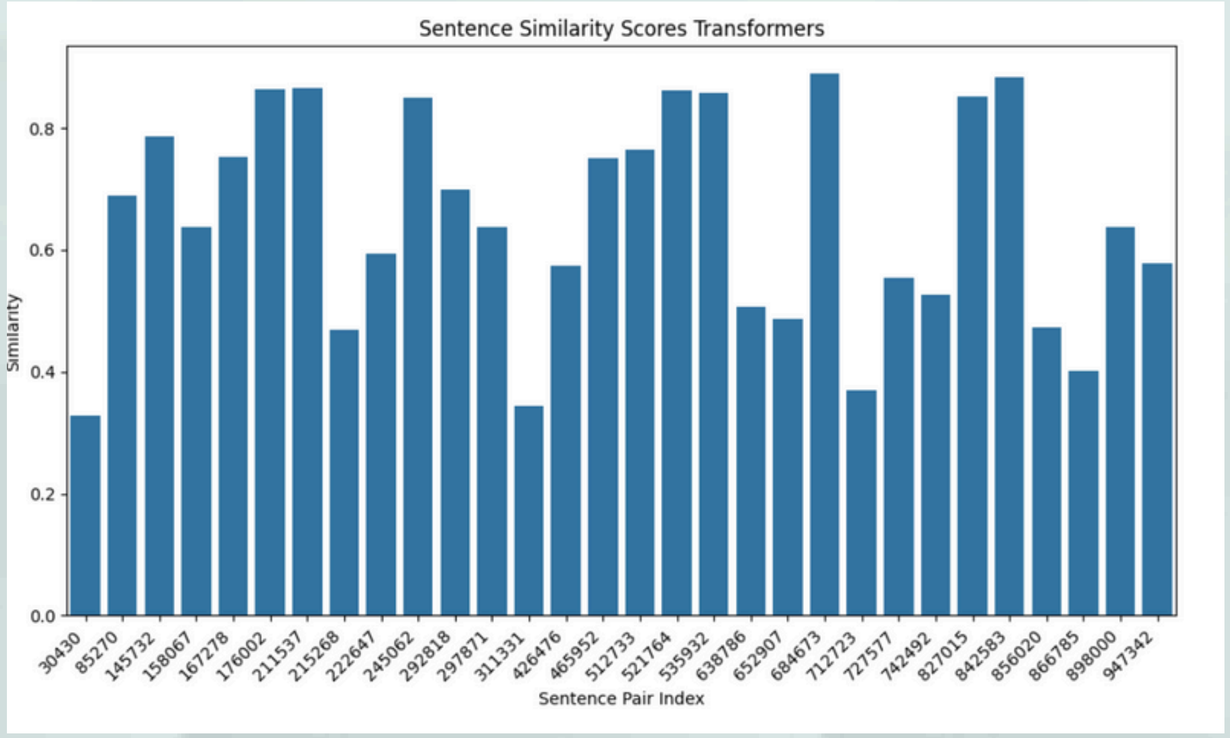


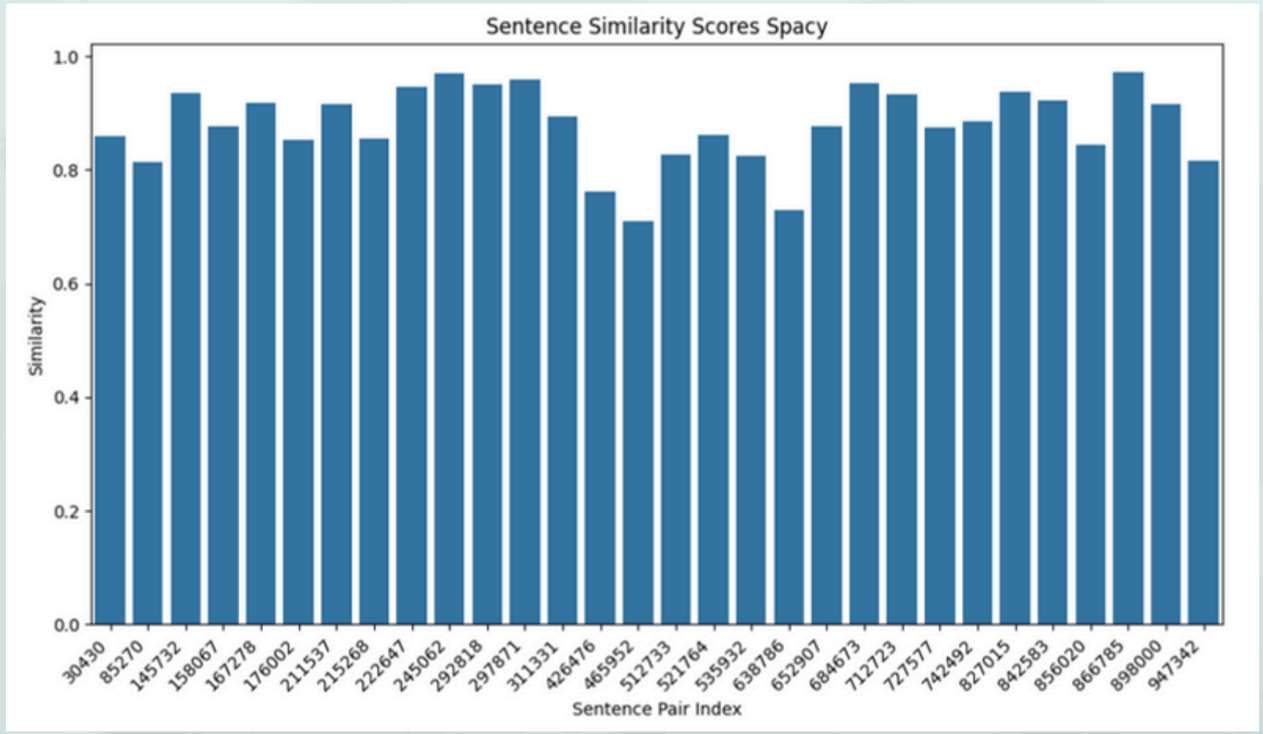
SCATTER PLOT

compares original similarity with the paraphrased sentences similarities.



TRANSFORMERS VS SPACY

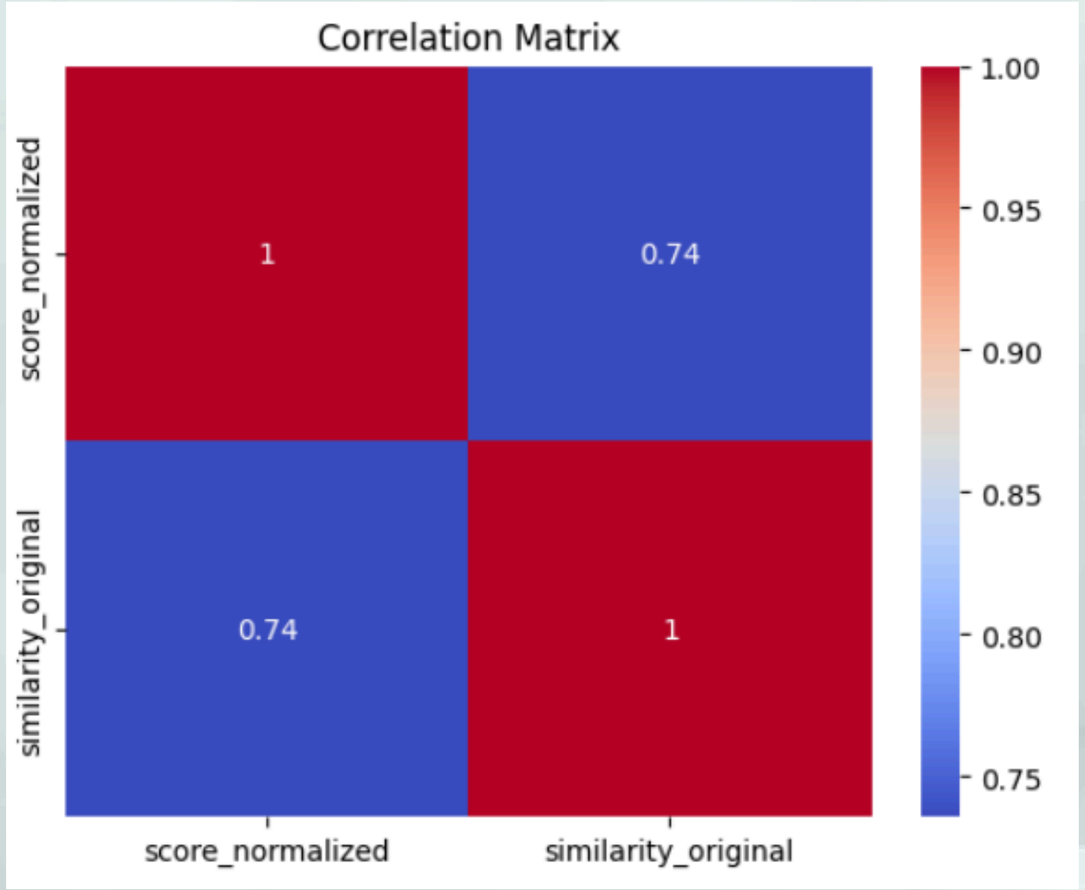


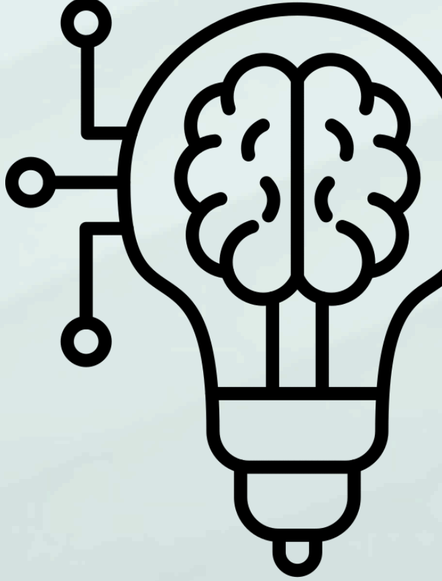


As we can see, SpaCy similarities are higher, which makes sence because it doestn take into account the context, this is due to its reliance on word-level embeddings, which can overestimate similarity based on vocabulary overlap. Transformer-based models, such as LaBSE, typically provide lower but more accurate scores by capturing the full contextual meaning of sentences.

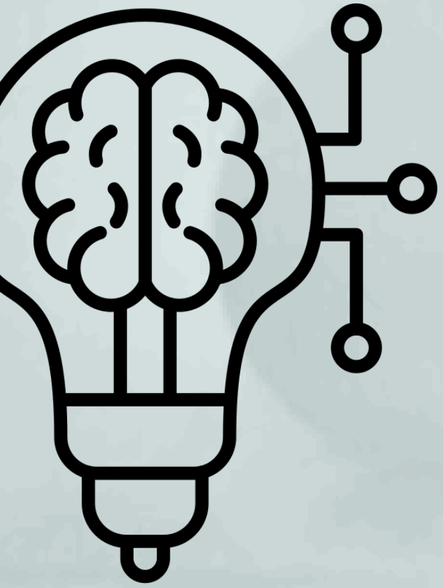
PERFORMANCE OF MY TRANSFORMER MODEL

As we see, performance is really good. This correlation matrix is indicating a strong positive relationship. This suggests that the transformer-based model effectively captures semantic equivalence in line with the reference scores provided in the dataset.





THANK YOU!



Group 21