Introduction to NLP

Report

Project Title : Neural Unsupervised Paraphrasing

Team Name : Team Linguisto
Team no. : 47

Nevil Sakhreliya - 2023201005
Darshak Devani - 2023201007
Shah Viraj Utpalbhai - 2023201011

Guided By : Prof. Manish Shrivastava
Prof. Rahul Mishra
Teaching Assistant: Lakshmipathi Balaji

The International Institute of Information Technology - Hyderabad - 2024

Problem Statement :

Paraphrasing is expressing a sentence using different words while maintaining the meaning. In this project teams will be implementing unsupervised approaches to generate paraphrases for Indian Languages.

Problem Description:

Paraphrasing, the task of expressing the same meaning using different words or structures, is fundamental in natural language processing (NLP). Paraphrasing can improve the readability and flow of written text. It allows writers to rephrase complex or convoluted sentences into simpler, more concise language, making the text more accessible to a wider audience. Traditional methods for paraphrase generation often rely on supervised approaches, which necessitate large amounts of annotated data. However, acquiring such labeled datasets is labor-intensive and may not cover the diverse range of expressions found in natural language. In contrast, unsupervised paraphrasing aims to generate paraphrases without relying on labeled data, offering a more scalable and versatile solution.

Dataset:

Quora Question Pairs (QQP) Dataset:

- The QQP dataset consists of over 400,000 question pairs from the community question-answering website Quora.
- Training Set: 404,290 question pairs
- Test Set: 2,345,795 question pairs
- vocab size: 73,948 unique words.

  - Due to resource consteraints only able to fintune the GPT2(medium) pretrained model for 10,000 sentences and for 3 epochs.

IndicCorp (Hindi)

- The corpus is a single large text file containing one sentence per line. The publicly released version is randomly shuffled, untokenized and deduplicated.
- IndicCorp is one of the largest publicly-available corpora for Indian languages.
- The dataset contains 4.95 million Hindi news articles, comprising a total of 63.1 million sentences.

  - Due to resource consteraints only able to fintune the GPT2(sberbank-ai/mGPT) pretrained model for 18,000 sentences and for 3 epochs.

Literature Review and findings:

1. Paraphrase Generation as Unsupervised Machine Translation
(https://arxiv.org/abs/2109.02950)

- Leveraging the paraphrase pairs generated by these UMT models, a unified surrogate model is constructed, which serves as the final SEQ2SEQ model for paraphrase generation. These generated paraphrases can be directly utilized for testing in an unsupervised setup or subjected to fine-tuning on labeled datasets in a supervised setup.
- The proposed methodology for paraphrase generation hinges on the assumption that a sizable dataset is available, containing numerous instances of sentences expressing analogous meanings. This assumption underscores a major limitation of the approach, as it necessitates access to extensive data containing similar sentences.

2. ConRPG: Paraphrase Generation using Contexts as Regularizer
(https://arxiv.org/abs/2109.00363)

- The paper assumes that the probabilities of generating two sentences with the same meaning, given the same context, should be the same.
- Using the trained context-LM, the authors decoded multiple candidate paraphrases with respect to a given context using diverse decoding via beam search
- The selected pairs were used to train an SEQ2SEQ model, which could be further fine tuned with supervised data or used directly for unsupervised paraphrase generation.

3. Paraphrase Generation: A Survey of the State of the Art
(https://aclanthology.org/2021.emnlp-main.414.pdf)

The research paper explores various methods and techniques employed in the domain of paraphrase generation.

- **Rule-based Methods :** Use predefined linguistic rules for paraphrase generation, including synonym substitution and syntactic transformations.
- **Knowledge-based Methods :** Use external knowledge sources like lexical databases and semantic networks to generate contextually appropriate and semantically similar paraphrases.
- **Thesaurus-Based Approaches :** Use thesauri to replace words with synonyms or related terms to generate paraphrases.
- **Machine Learning-based Methods :** The method combines Transformer networks with VQ-VAEs for unsupervised paraphrasing. It uses discrete latent variables for grouping related examples and introduces a hybrid model with a residual connection to handle training challenges.

4. Paraphrase Generation: A Survey of the State of the Art
(https://arxiv.org/pdf/1905.12752.pdf)

- The method combines Transformer networks with VQ-VAEs for unsupervised paraphrasing. It uses discrete latent variables for grouping related examples and introduces a hybrid model with a residual connection to handle training challenges.
- The Transformer encoder maps sentences into fixed-size vectors, and outputs undergo quantization with residual connections.

Approach and Implementation:

(1). Statistical Approach :

- Tried for English language corpus.

- As suggested by our project guide, We implemented a statistical method for paraphrasing, leveraging techniques like word substitution, POS tagging, and random sampling of synonyms from Wordnet to generate alternate versions of sentences.

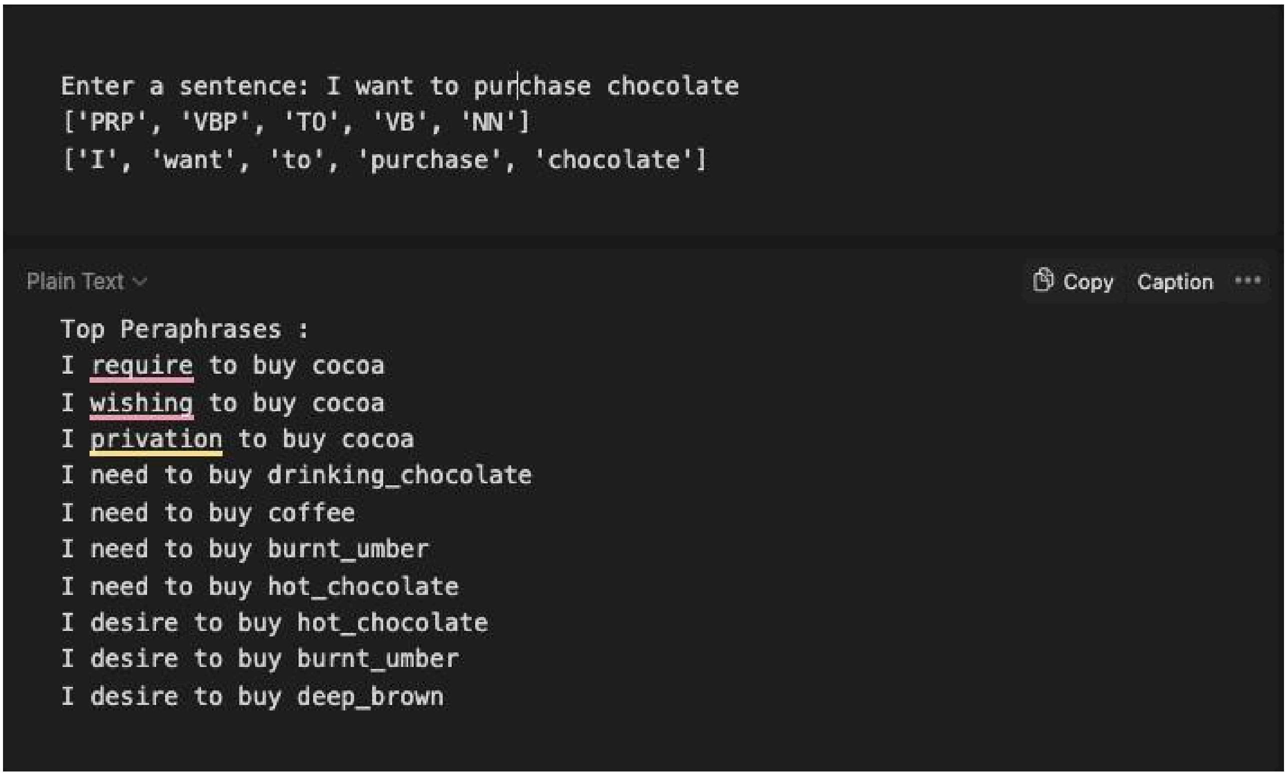
- **POS Tagging and Synonym Replacement:** POS tagging plays a crucial role in identifying the grammatical structure of sentences. We utilized Flair, a powerful NLP library, to perform POS tagging on input sentences. By tagging words with their respective POS labels, we gained insight into the syntactic structure of the text.

- We identified specific POS tags for synonym replacement, including nouns (NN, NNS, NNP, NNPS), verbs (VB, VBD, VBG, VBN, VBP), adjectives (JJ, JJR, JJS), and interjections (UH). These tags cover a wide range of lexical categories, allowing for comprehensive paraphrasing.

- **Generating Paraphrases:**

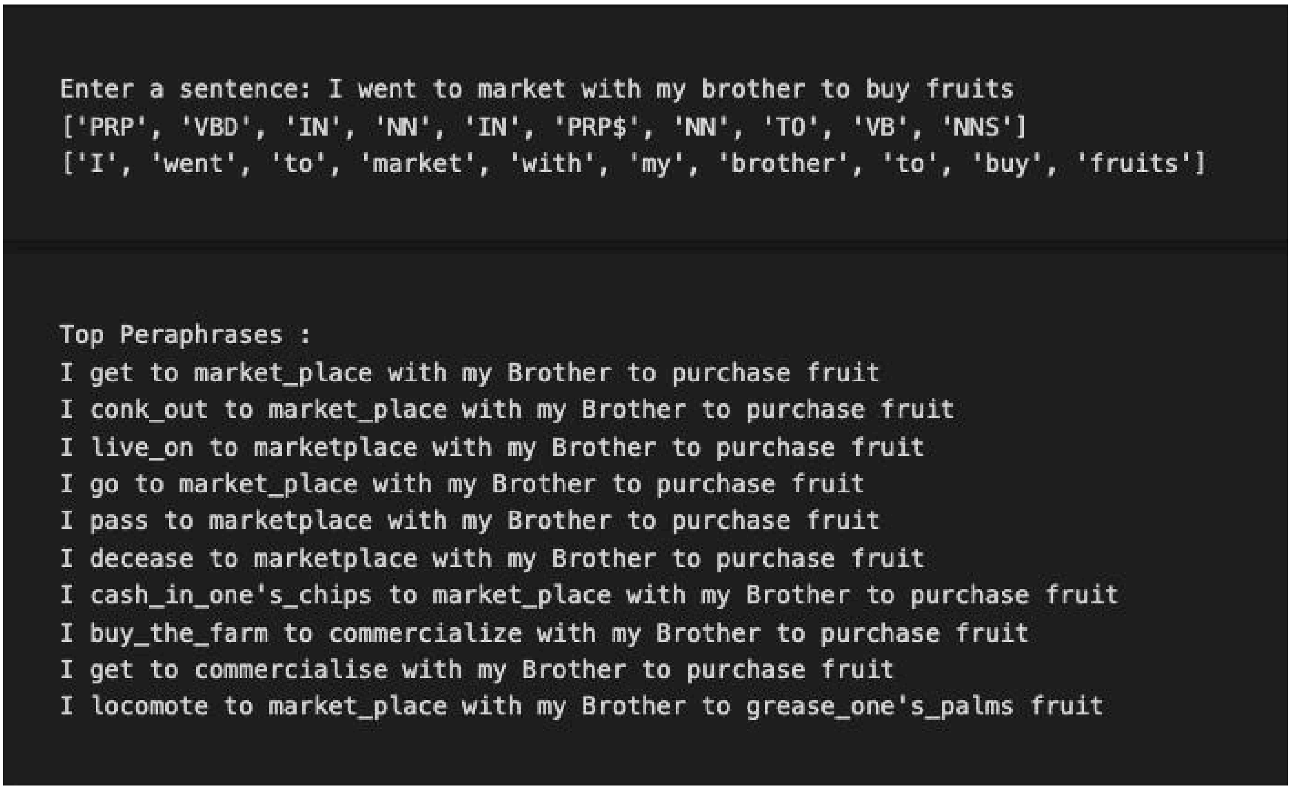
    1. **Synonym Retrieval:** Leveraging WordNet, a lexical database of English, we obtained synonyms for words corresponding to the chosen POS tags. WordNet offers an extensive repository of synonyms, enabling us to explore diverse alternatives for each word.
    2. **Combination of Sentences:** After obtaining synonyms, we systematically replaced words in the original sentence with their synonymous counterparts, generating multiple paraphrased versions. The process involved computing all possible combinations of sentences resulting from these replacements.
    3. **Random Sampling:** To ensure diversity in the paraphrases and reduce compute complexity, we employed random sampling. By selecting 20% of the total possible sentences randomly, we aimed to generate a varied set of paraphrases in each iteration.
    4. **Paraphrasing:** With the selected samples, we proceeded to paraphrase the sentences. Using the synonyms obtained from WordNet, we replaced words in the original sentences, creating rephrased versions.
    5. **Similarity Evaluation:** To evaluate the quality of paraphrases, we calculated similarity scores between the original sentence and its paraphrased versions. Metrics such as BERT score and BLEU score were employed to measure the resemblance between the sentences, providing quantitative insights into the paraphrasing process.
    6. **Conclusion:** our approach leverages Flair for POS tagging and WordNet for synonym retrieval to generate paraphrases of input sentences. By systematically replacing words and employing random sampling, we ensure diversity in the paraphrased outputs. Additionally, similarity scores offer a quantitative assessment of the resemblance between the original and paraphrased sentences, aiding in the evaluation of the paraphrasing process.

Output of statistical method for paraphrasing:



```
Enter a sentence: I want to purchase chocolate
['PRP', 'VBP', 'TO', 'VB', 'NN']
['I', 'want', 'to', 'purchase', 'chocolate']
```

```
Top Peraphrases :
I require to buy cocoa
I wishing to buy cocoa
I privation to buy cocoa
I need to buy drinking_chocolate
I need to buy coffee
I need to buy burnt_umber
I need to buy hot_chocolate
I desire to buy hot_chocolate
I desire to buy burnt_umber
I desire to buy deep_brown
```



```
Enter a sentence: I went to market with my brother to buy fruits
['PRP', 'VBD', 'IN', 'NN', 'IN', 'PRP$', 'NN', 'TO', 'VB', 'NNS']
['I', 'went', 'to', 'market', 'with', 'my', 'brother', 'to', 'buy', 'fruits']
```

```
Top Peraphrases :
I get to market_place with my Brother to purchase fruit
I conk_out to market_place with my Brother to purchase fruit
I live_on to marketplace with my Brother to purchase fruit
I go to market_place with my Brother to purchase fruit
I pass to marketplace with my Brother to purchase fruit
I decease to marketplace with my Brother to purchase fruit
I cash_in_one's_chips to market_place with my Brother to purchase fruit
I buy_the_farm to commercialize with my Brother to purchase fruit
I get to commercialise with my Brother to purchase fruit
I locomote to market_place with my Brother to grease_one's_palms fruit
```

(2). Unsupervised Paraphrasing Generation Using Pre-Trained Language Models :

- **Reference Paper : https://arxiv.org/pdf/2006.05477**

Approach :

- Used an unsupervised paraphrasing technique using pret-rained GPT-2 model.
    - For English text : Used GPT2(medium) pretrained model
    - For Hindi text : Used GPT2(sberbank-ai/mGPT) pretrained model.

- As this is an unsupervised approach, the model can be trained on the domain specific independent sentences at hand and generate paraphrases without suffering from domain shift.
- We take the unsupervised paraphrasing task as a sentence reconstruction task from corrupted input.
- From a sentence, we omit all the stop words to form a corrupted sentence, let's call it Source S, and the original sentence is used as Target T. We use GPT-2 to generate the Target sentence given Source. i.e P(T |S)
- GPT-2 is exceptional at language generation. It predicts the next token in the sequence given all the tokens before it, i.e it optimizes for P (Xi |X< i). The pre-trained version of GPT-2 generates bland text without any goal. By making the model to generate Target T by conditioning on the Source S, the language generation capability of GPT-2 can be utilized for generating meaningful text.

Data Preprocessing :

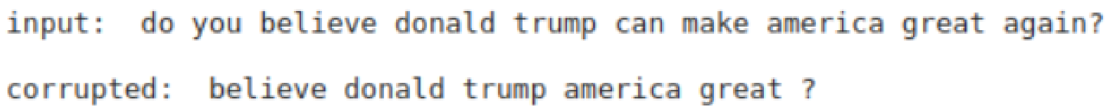
- Remove the stop words from the training set.
- English Stop Words :
  Stop words provided by NLTK library.
  ['i', 'me', 'my', 'myself', 'we', 'our','ours', 'ourselves', 'you', 'your', 'yours','yourself', 'yourselves', 'he', 'him', 'his','himself', 'she', 'her', 'hers', 'herself','it', 'its', 'itself', 'they', 'them', 'their','theirs', 'themselves', 'what', 'which', 'who','whom', 'this', 'that', 'these', 'those', 'am','is', 'are', 'was', 'were', 'be', 'been', 'being','have', 'has', 'had', 'having', 'do', 'does', 'did','doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or','because', 'as', 'until', 'while', 'of', 'at','by', 'for', 'with', 'about', 'against', 'between','into', 'through', 'during', 'before', 'after','above', 'below', 'to', 'from', 'up', 'down', 'in','out', 'on', 'off', 'over', 'under', 'again','further', 'then', 'once', 'here', 'there', 'when','where', 'why', 'how', 'all', 'any', 'both', 'each','few', 'more', 'most', 'other', 'some', 'such', 'no','nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too','very', 's', 't', 'can', 'will', 'just', 'don','should', 'now', 'someone', 'something', 'make', 'see', 'everything', 'anyone', 'anything', 'everyone']

- Hindi Stop Words :
  ["एक", "एवं", "यह", "इस", "के", "का", "की", "को", "में", "है", "हैं", "कर", "किया", "किए", "करते", "करना", "किसी", "गया", "जाता", "जाती", "जाते", "साथ", "अपने", "हुआ", "होता", "होती", "होते", "वाले", "वह", "वहाँ", "जैसा", "जिसका", "जिसकी", "जिसके", "जिनको", "जिनकी", "जिनके", "तथा", "उसके", "उसका", "उसकी", "उनके", "उनका", "उनकी", "उनको", "कुछ", "इसका", "इसकी", "इसके", "सभी", "अगर", "इसमें", "उनका", "उनकी", "उनके", "जैसे", "जिसमें", "तिन्हों", "तिन्हें", "पहले", "बाद", "मानो", "अंदर", "भीतर", "पूरे", "खुद", "आप", "अब", "जब", "जहाँ", "जितना", "जितने", "तब", "वहीं", "हुआ", "होता", "होती", "वाला", "वाली", "वाले"]
- Remove all special characters like, “,”,’

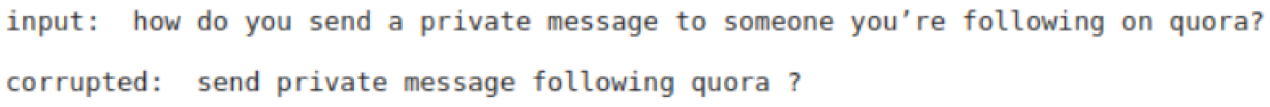
Sentence Corruption Techniques :

- **For English Text :**
  1. **Random Shuffle :** Shuffle the given input sentence.
  2. **Synonym Replacement :** Choose 40% of the words from the sentence randomly. Find the synonyms of those words from the wordnet library and choose a random synonym from the list of synonyms and replace it with the original word in the sentence.

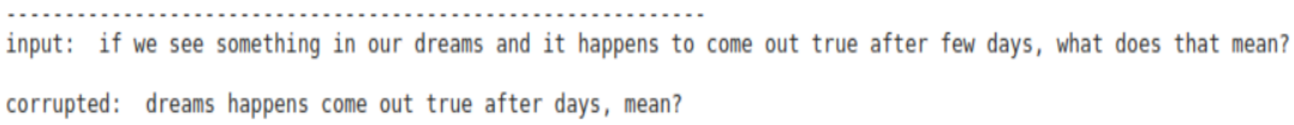
Examples :



```
input:   do you believe donald trump can make america great again?

corrupted:   believe donald trump america great ?
```



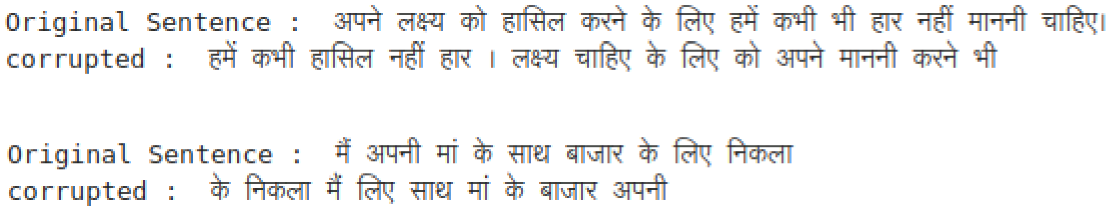
```
input:   how do you send a private message to someone you're following on quora?

corrupted:   send private message following quora ?
```



```
------------------------------------------------------------
input:  if we see something in our dreams and it happens to come out true after few days, what does that mean?

corrupted:  dreams happens come out true after days, mean?
```

- **For Hindi Text :**
  1. **Random Shuffle :** Shuffle the given input sentence. SHuffle the same sentence 4 times.
  2. **Random word Deletion :** Choose some of the words from the sentence and delete those words from the original sentence.

Examples :



```
Original Sentence :  अपने लक्ष्य को हासिल करने के लिए हमें कभी भी हार नहीं माननी चाहिए।
corrupted :  हमें कभी हासिल नहीं हार । लक्ष्य चाहिए के लिए को अपने माननी करने भी


Original Sentence :  मैं अपनी मां के साथ बाजार के लिए निकला
corrupted :  के निकला मैं लिए साथ मां के बाजार अपनी
```

Convert all sentences in the form of <BeginingOfSentence> Corrupted Sentence <SeperaterToken> Original Sentence <EndOfSentence>

At time of inference convert the input sentence into the format of <BeginingOfSentence> Corrupted Sentence <SeperaterToken> after removing stop words and ask model to generate the sentence, which will be paraphrase.

Similarity Score :

Calculates the contextual similarity between two sentences using a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model.

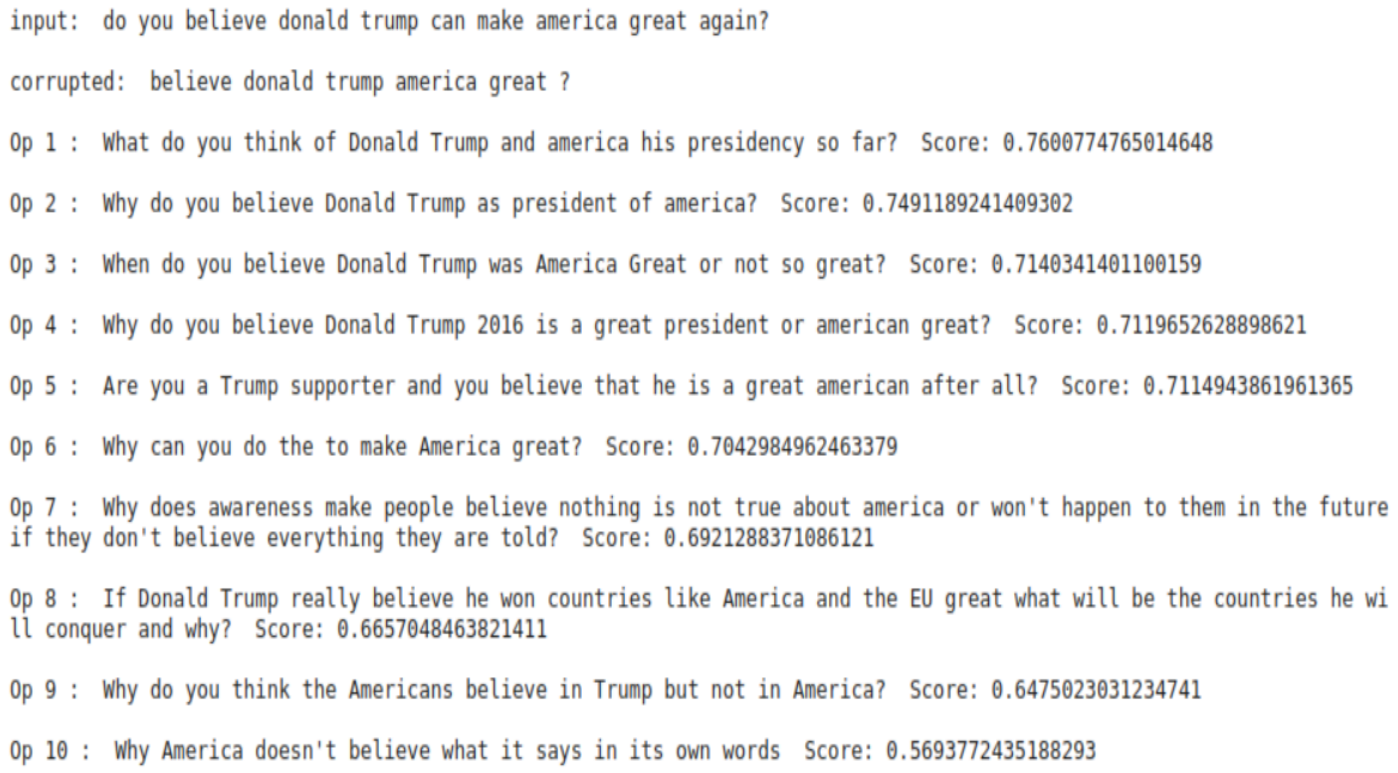
Calculates the cosine similarity between the contextual embeddings of the two sentences. Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space. It measures the cosine of the angle between them.

Pretrained model BERTt will generate contextual sentence embeddings and calculate similarity between them.

Qualitative and Quantiative Evaluation (Inference) :

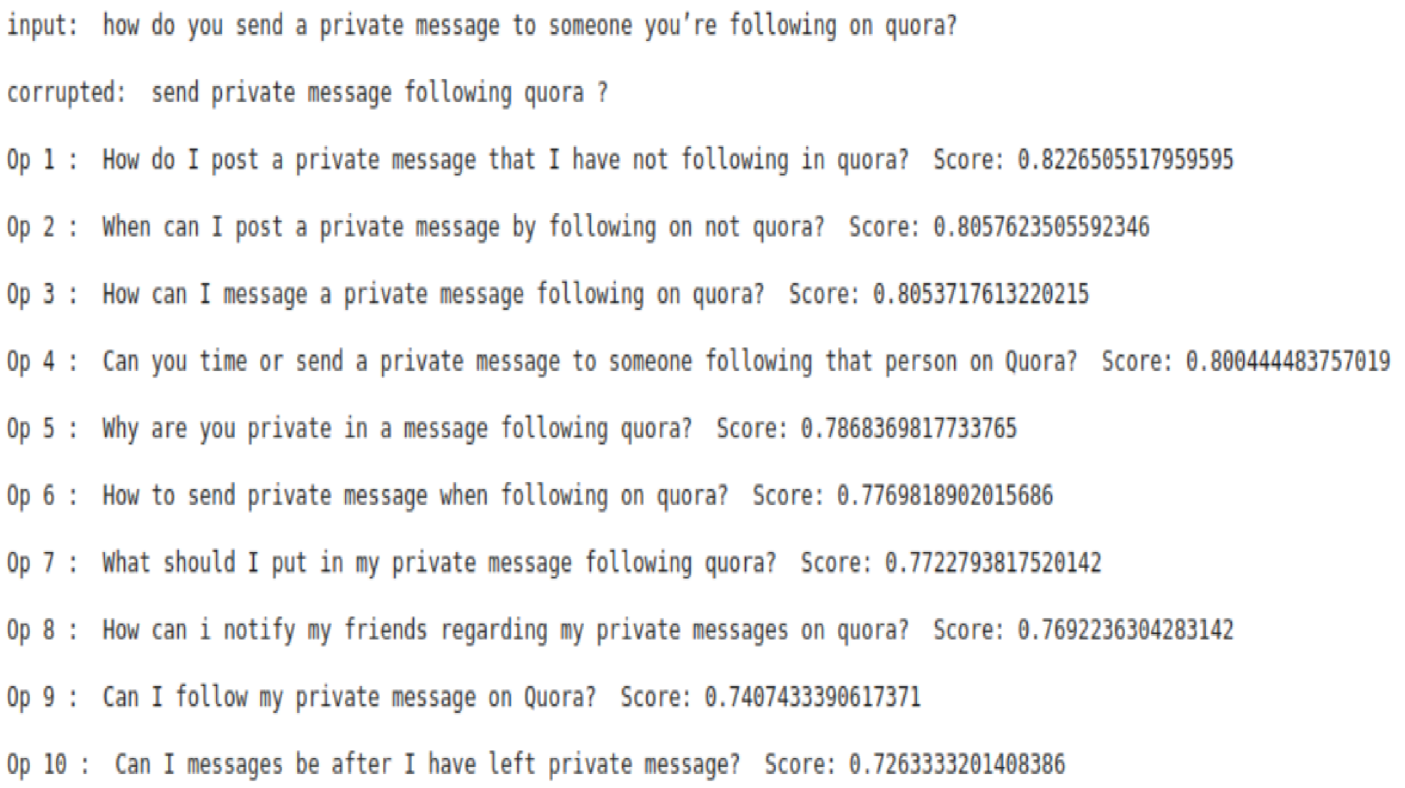
For English Text :

(1).



```
input:  do you believe donald trump can make america great again?

corrupted:  believe donald trump america great ?

Op 1 :  What do you think of Donald Trump and america his presidency so far?  Score: 0.7600774765014648

Op 2 :  Why do you believe Donald Trump as president of america?  Score: 0.7491189241409302

Op 3 :  When do you believe Donald Trump was America Great or not so great?  Score: 0.7140341401100159

Op 4 :  Why do you believe Donald Trump 2016 is a great president or american great?  Score: 0.7119652628898621

Op 5 :  Are you a Trump supporter and you believe that he is a great american after all?  Score: 0.7114943861961365

Op 6 :  Why can you do the to make America great?  Score: 0.7042984962463379

Op 7 :  Why does awareness make people believe nothing is not true about america or won't happen to them in the future
if they don't believe everything they are told?  Score: 0.6921288371086121

Op 8 :  If Donald Trump really believe he won countries like America and the EU great what will be the countries he wi
ll conquer and why?  Score: 0.6657048463821411

Op 9 :  Why do you think the Americans believe in Trump but not in America?  Score: 0.6475023031234741

Op 10 :  Why America doesn't believe what it says in its own words  Score: 0.5693772435188293
```

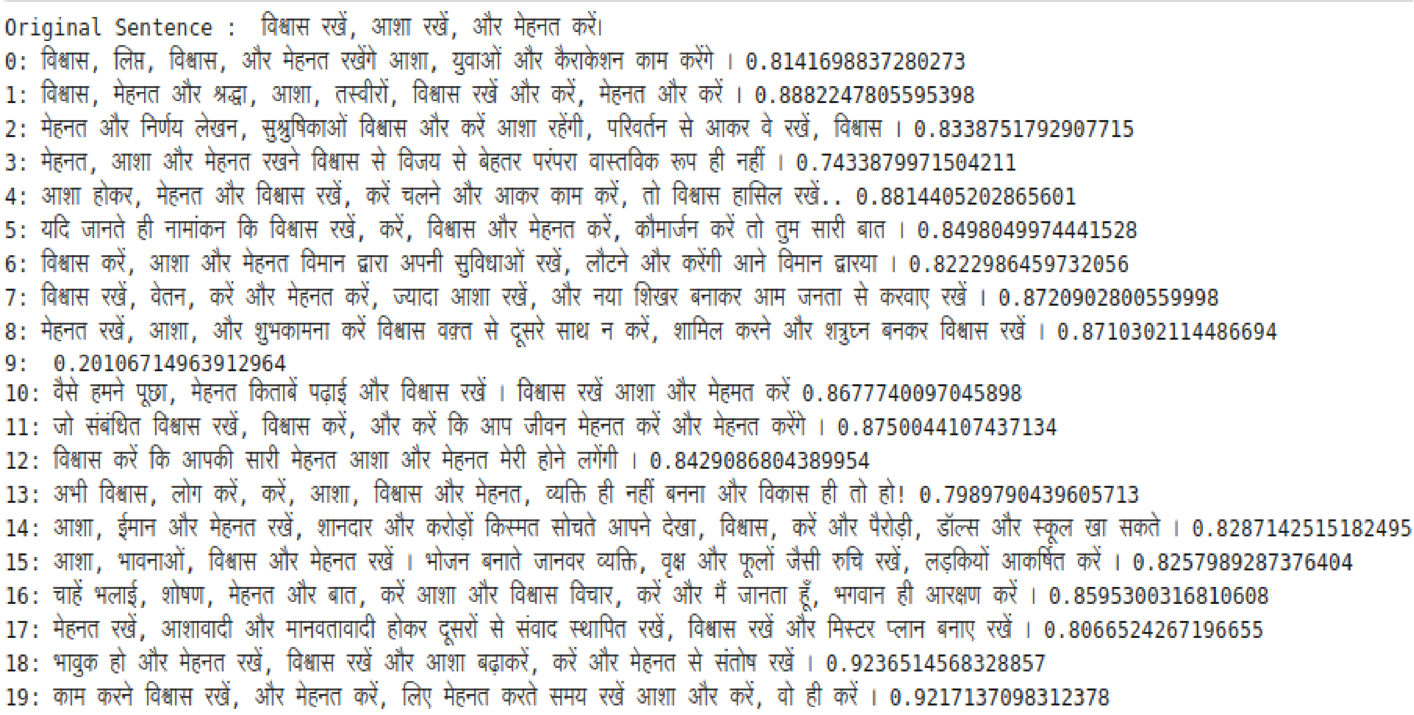
(2).



```
input:  how do you send a private message to someone you're following on quora?

corrupted:  send private message following quora ?

Op 1 :  How do I post a private message that I have not following in quora?  Score: 0.8226505517959595

Op 2 :  When can I post a private message by following on not quora?  Score: 0.8057623505592346

Op 3 :  How can I message a private message following on quora?  Score: 0.8053717613220215

Op 4 :  Can you time or send a private message to someone following that person on Quora?  Score: 0.800444483757019

Op 5 :  Why are you private in a message following quora?  Score: 0.7868369817733765

Op 6 :  How to send private message when following on quora?  Score: 0.7769818902015686

Op 7 :  What should I put in my private message following quora?  Score: 0.7722793817520142

Op 8 :  How can i notify my friends regarding my private messages on quora?  Score: 0.7692236304283142

Op 9 :  Can I follow my private message on Quora?  Score: 0.7407433390617371

Op 10 :  Can I messages be after I have left private message?  Score: 0.7263333201408386
```

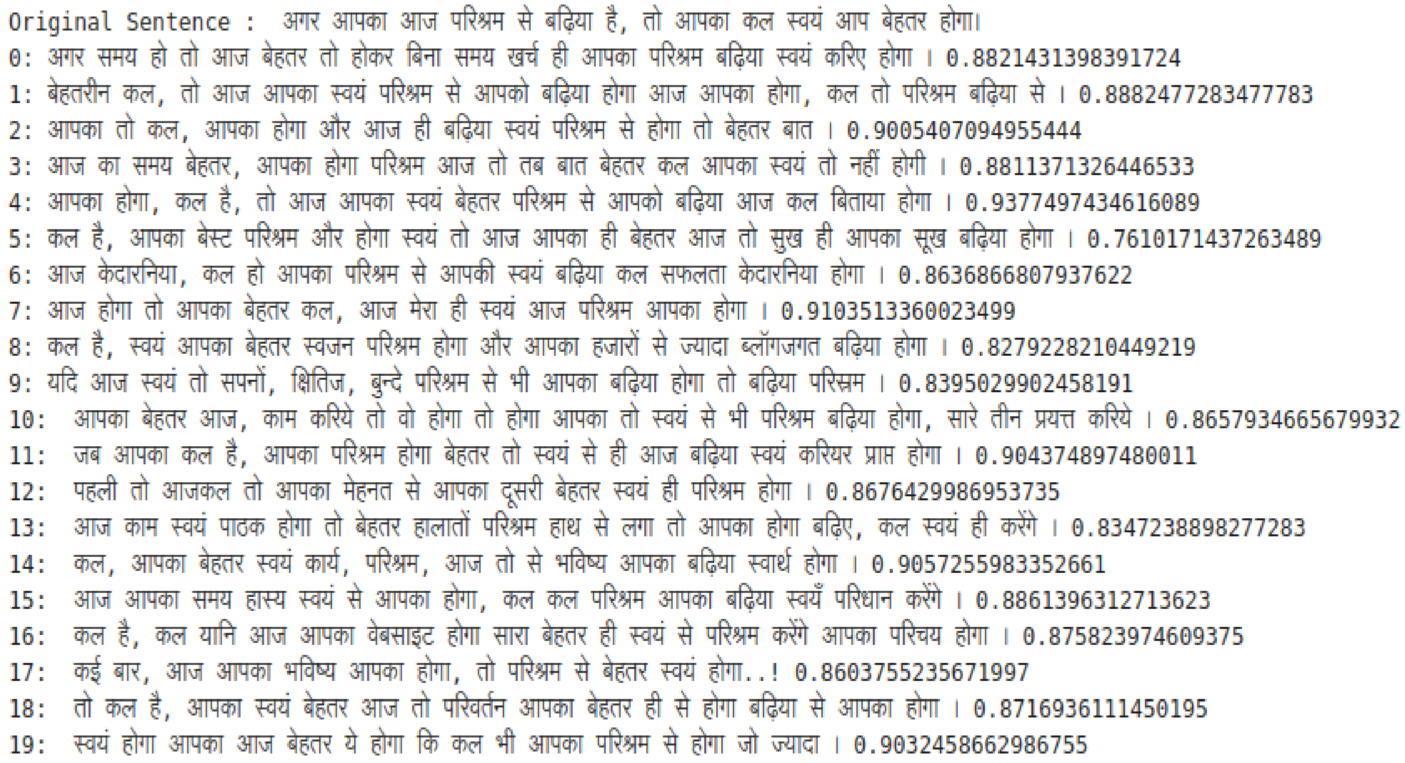
For Hindi Text :

(1).



```
Original Sentence :  विश्वास रखें, आशा रखें, और मेहनत करें।
0: विश्वास, लिप्त, विश्वास, और मेहनत रखेंगे आशा, युवाओं और कैराकेशन काम करेंगे । 0.8141698837280273
1: विश्वास, मेहनत और श्रद्धा, आशा, तस्वीरों, विश्वास रखें और करें, मेहनत और करें । 0.8882247805595398
2: मेहनत और निर्णय लेखन, सुश्रुषिकाओं विश्वास और करें आशा रहेंगी, परिवर्तन से आकर वे रखें, विश्वास । 0.8338751792907715
3: मेहनत, आशा और मेहनत रखने विश्वास से विजय से बेहतर परंपरा वास्तविक रूप ही नहीं । 0.7433879971504211
4: आशा होकर, मेहनत और विश्वास रखें, करें चलने और आकर काम करें, तो विश्वास हासिल रखें.. 0.8814405202865601
5: यदि जानते ही नामांकन कि विश्वास रखें, करें, विश्वास और मेहनत करें, कौमार्जन करें तो तुम सारी बात । 0.8498049974441528
6: विश्वास करें, आशा और मेहनत विमान द्वारा अपनी सुविधाओं रखें, लौटने और करेंगी आने विमान द्वारया । 0.8222986459732056
7: विश्वास रखें, वेतन, करें और मेहनत करें, ज्यादा आशा रखें, और नया शिखर बनाकर आम जनता से करवाए रखें । 0.8720902800559998
8: मेहनत रखें, आशा, और शुभकामना करें विश्वास वक़्त से दूसरे साथ न करें, शामिल करने और शत्रुघ्न बनकर विश्वास रखें । 0.8710302114486694
9:  0.20106714963912964
10: वैसे हमने पूछा, मेहनत किताबें पढ़ाई और विश्वास रखें । विश्वास रखें आशा और मेहमत करें 0.8677740097045898
11: जो संबंधित विश्वास रखें, विश्वास करें, और करें कि आप जीवन मेहनत करें और मेहनत करेंगे । 0.8750044107437134
12: विश्वास करें कि आपकी सारी मेहनत आशा और मेहनत मेरी होने लगेंगी । 0.8429086804389954
13: अभी विश्वास, लोग करें, करें, आशा, विश्वास और मेहनत, व्यक्ति ही नहीं बनना और विकास ही तो हो! 0.7989790439605713
14: आशा, ईमान और मेहनत रखें, शानदार और करोड़ों किस्मत सोचते आपने देखा, विश्वास, करें और पैरोड़ी, डॉल्स और स्कूल खा सकते । 0.8287142515182495
15: आशा, भावनाओं, विश्वास और मेहनत रखें । भोजन बनाते जानवर व्यक्ति, वृक्ष और फूलों जैसी रुचि रखें, लड़कियों आकर्षित करें । 0.8257989287376404
16: चाहें भलाई, शोषण, मेहनत और बात, करें आशा और विश्वास विचार, करें और मैं जानता हूँ, भगवान ही आरक्षण करें । 0.8595300316810608
17: मेहनत रखें, आशावादी और मानवतावादी होकर दूसरों से संवाद स्थापित रखें, विश्वास रखें और मिस्टर प्लान बनाए रखें । 0.8066524267196655
18: भावुक हो और मेहनत रखें, विश्वास रखें और आशा बढ़ाकरें, करें और मेहनत से संतोष रखें । 0.9236514568328857
19: काम करने विश्वास रखें, और मेहनत करें, लिए मेहनत करते समय रखें आशा और करें, वो ही करें । 0.9217137098312378
```

(2).



```
Original Sentence :  अगर आपका आज परिश्रम से बढ़िया है, तो आपका कल स्वयं आप बेहतर होगा।
0: अगर समय हो तो आज बेहतर तो होकर बिना समय खर्च ही आपका परिश्रम बढ़िया स्वयं करिए होगा । 0.8821431398391724
1: बेहतरीन कल, तो आज आपका स्वयं परिश्रम से आपको बढ़िया होगा आज आपका होगा, कल तो परिश्रम बढ़िया से । 0.8882477283477783
2: आपका तो कल, आपका होगा और आज ही बढ़िया स्वयं परिश्रम से होगा तो बेहतर बात । 0.9005407094955444
3: आज का समय बेहतर, आपका होगा परिश्रम आज तो तब बात बेहतर कल आपका स्वयं तो नहीं होगी । 0.8811371326446533
4: आपका होगा, कल है, तो आज आपका स्वयं बेहतर परिश्रम से आपको बढ़िया आज कल बिताया होगा । 0.9377497434616089
5: कल है, आपका बेस्ट परिश्रम और होगा स्वयं तो आज आपका ही बेहतर आज तो सुख ही आपका सूख बढ़िया होगा । 0.7610171437263489
6: आज केदारनिया, कल हो आपका परिश्रम से आपकी स्वयं बढ़िया कल सफलता केदारनिया होगा । 0.8636866807937622
7: आज होगा तो आपका बेहतर कल, आज मेरा ही स्वयं आज परिश्रम आपका होगा । 0.9103513360023499
8: कल है, स्वयं आपका बेहतर स्वजन परिश्रम होगा और आपका हजारों से ज्यादा ब्लॉगजगत बढ़िया होगा । 0.8279228210449219
9: यदि आज स्वयं तो सपनों, क्षितिज, बुन्दे परिश्रम से भी आपका बढ़िया होगा तो बढ़िया परिस्रम । 0.8395029902458191
10:  आपका बेहतर आज, काम करिये तो वो होगा तो होगा आपका तो स्वयं से भी परिश्रम बढ़िया होगा, सारे तीन प्रयत्त करिये । 0.8657934665679932
11:  जब आपका कल है, आपका परिश्रम होगा बेहतर तो स्वयं से ही आज बढ़िया स्वयं करियर प्राप्त होगा । 0.904374897480011
12:  पहली तो आजकल तो आपका मेहनत से आपका दूसरी बेहतर स्वयं ही परिश्रम होगा । 0.8676429986953735
13:  आज काम स्वयं पाठक होगा तो बेहतर हालातों परिश्रम हाथ से लगा तो आपका होगा बढ़िए, कल स्वयं ही करेंगे । 0.8347238898277283
14:  कल, आपका बेहतर स्वयं कार्य, परिश्रम, आज तो से भविष्य आपका बढ़िया स्वार्थ होगा । 0.9057255983352661
15:  आज आपका समय हास्य स्वयं से आपका होगा, कल कल परिश्रम आपका बढ़िया स्वयँ परिधान करेंगे । 0.8861396312713623
16:  कल है, कल यानि आज आपका वेबसाइट होगा सारा बेहतर ही स्वयं से परिश्रम करेंगे आपका परिचय होगा । 0.875823974609375
17:  कई बार, आज आपका भविष्य आपका होगा, तो परिश्रम से बेहतर स्वयं होगा..! 0.8603755235671997
18:  तो कल है, आपका स्वयं बेहतर आज तो परिवर्तन आपका बेहतर ही से होगा बढ़िया से आपका होगा । 0.8716936111450195
19:  स्वयं होगा आपका आज बेहतर ये होगा कि कल भी आपका परिश्रम से होगा जो ज्यादा । 0.9032458662986755
```

Challenges faced :

- Limited compute power for fine-tuning the pre-trained model.
- Insufficient storage capacity to store all parameters of the pre-trained model.
- Absence of a robust synonym library for Hindi words

Analysis :

- The statistical method is deemed less efficient compared to the fine-tuning approach applied to pre-trained models. In the statistical method, the focus remains primarily on substituting the most probable words with their synonyms without considering sentence restructuring.

- Conversely, fine-tuning leverages pre-trained models already trained on extensive corpora. It employs a sentence reconstruction task to generate paraphrases by corrupting sentences and attempting to reconstruct the original sentence from the corrupted version.

- However, due to resource constraints, we could only fine-tune the GPT2 pre-trained model for three epochs. Considering the substantial size of GPT2, we faced limitations in fine-tuning it with a larger corpus for more training epochs due to computational issues, leading to less accurate results.

- Additionally, English pre-trained models have fewer parameters compared to their Hindi counterparts, thus performing better in certain contexts.