**A.**

**1. Resources used: Hugging Face.**

1. I used Python Virtual Environment to install all my dependencies and to run the scripts. See first steps in the installation page:<https://huggingface.co/docs/transformers/main/en/installation>
2. Pretrained Models from Hugging Face: <https://huggingface.co/models>
3. LNP / Text2TextGeneration pretrained model:<https://huggingface.co/models?pipeline_tag=text2text-generation&sort=downloads&search=figurative-nlp>
4. I worked with these 2 models:

**First model :** figurative-nlp/t5-figurative-paraphrase.This model can convert the figurative/metaphorical expression to the literal expression.

<https://huggingface.co/figurative-nlp/t5-figurative-paraphrase>

**Second model:** figurative-nlp/sarcasm-hyperbole-humor-paraphrase.

There is no description about this pretrained model, I just tried to see what

We get, even if we don’t have the description of the model. In any case, it

Was created to make a conversion from figurative to literal text:

<https://huggingface.co/figurative-nlp/sarcasm-hyperbole-humor-paraphrase>

**2.** Only for the metaphor I tested only one model, for the others, I tested both.

**3.** I tried to test the models on a sample of 1000 inputs, but it takes a lot of time. Thus, to save time only took a sample of 100 inputs, the first 100 inputs.

4. All the csv files are from AI4GoodLab M3 project / Datasets(downloaded from the following link on Sanday, June 4th) : <https://drive.google.com/drive/folders/1EJRCl9YTmYmpXVkDgCHQQ5XdBK-myTBd>

**5.**  As I said during today's TA session, I just gave the model inputs, and I wanted it to give me an output, and to randomly see visually what the models transform.

**B. Testing\_NLP\_Pretrained\_Models\_Text2Text Generation**

There is the GitHub link that helps to install the BLEURT, how to use BLEURT, how to interpret the result of BLEURT, etc…

a. https://arxiv.org/pdf/2004.04696.pdf

b. <https://github.com/google-research/bleurt>

c. https://ai.googleblog.com/2020/05/evaluating-natural-language-generation.html

d. <https://arxiv.org/abs/2004.04696>

e. <https://huggingface.co/spaces/evaluate-metric/bleurt>

The BLEURT score typically ranges between 0 and 1, where a higher score indicates a higher similarity or quality between the candidate translation and the reference translations. A BLEURT score of 1 signifies that the candidate translation is almost identical to the reference translations, while a score of 0 indicates a complete mismatch or dissimilarity.

It's important to note that the exact calculation and interpretation of the BLEURT score may vary depending on the specific implementation and version of BLEURT used. Different variations or versions of BLEURT may have different scaling or normalization techniques, but the general principle remains the same—to quantify the similarity between the candidate and reference translations.

To calculate the BLEURT score, the candidate translation and reference translations are compared based on various linguistic and contextual features. These features are used to compute a similarity score, which is then converted into the final BLEURT score. The specific methodology and features used by BLEURT may vary based on the implementation and training data.

Overall, the BLEURT score provides a quantitative measure of how well a machine-generated translation matches the quality of human reference translations. It serves as a useful tool for evaluating and comparing different translation models or systems.

**Conclusion:** Based on the information that the Bleurt score

1. **For SARCASM:**

**a.For figurative-nlp/t5-figurative-paraphrase model:**

For figurative-nlp/t5-figurative-paraphrase model,the average BLEURT score for SARCASM is 0.16986634746193885

Suggestion: Not a good model for the SARCASM

**b.For figurative-nlp/sarcasm-hyperbole-humor-paraphrase model:**

For figurative-nlp/sarcasm-hyperbole-humor-paraphrase model,the average BLEURT score for SARCASME is 0.258478859513998

Suggestion: Not a good model for the SARCASM

1. **For METONYMIES:**

**a.For figurative-nlp/t5-figurative-paraphrase model:**

For figurative-nlp/t5-figurative-paraphrase model,the average BLEURT score for METONYMIES is 0.48975935354828

Suggestion: Is not an ideal model but we can use it for METONYMIES

**b.For figurative-nlp/sarcasm-hyperbole-humor-paraphrase model:**

For figurative-nlp/sarcasm-hyperbole-humor-paraphrase model,the average BLEURT score for METONYMIES is 0.6349983891844749

Suggestion:Is not an ideal model bu we can use it for METONYMIES

1. **For METAPHOR:**

**a.For figurative-nlp/t5-figurative-paraphrase model:**

For figurative-nlp/sarcasm-hyperbole-humor-paraphrase model,the average BLEURT score for METAPHOR is 0.6576590889692306

Suggestion:Is not an ideal model bu we can use it for METAPHOR

**b.**I didn't try the figurative-nlp/t5-figurative-paraphrase because (print("figurative-nlp/t5-figurative-paraphrase, out of a hundred sentences, only ",counter\_int, " were modified by the transformer")#21 were modified by the transformer), please see the code.

1. **For IDIOMS:**

**a.For figurative-nlp/t5-figurative-paraphrase model:**

For figurative-nlp/t5-figurative-paraphrase model,the average BLEURT score for IDIOMS is -0.05678766831755638

Suggestion: The worst model for the IDIOMS, never use

**b.For figurative-nlp/sarcasm-hyperbole-humor-paraphrase model:**

For figurative-nlp/sarcasm-hyperbole-humor-paraphrase model,the average BLEURT score for IDIOMS is 0.8051432381570339

Suggestion: Best model for IDIOMS, use it

1. **For HYPERBOLA:**

**a.For figurative-nlp/t5-figurative-paraphrase model:**

For figurative-nlp/t5-figurative-paraphrase model,the average BLEURT score for HYPERBOLE is 0.6017162628471

Suggestion: Good to use it for HYPERBOLA

**b.For figurative-nlp/sarcasm-hyperbole-humor-paraphrase model:**

For figurative-nlp/sarcasm-hyperbole-humor-paraphrase model,the average BLEURT score for HYPERBOLE is 0.7668591256439

Suggestion: Best model for HYPERBOLA, use it