Content Style Transfer between Shakespearean and CNN News Style: T5 Model Approach

Project Link(All the code and Data): https://github.com/GussLii/Natural-Language_Project

Group members: Xiaoyu Yang, Xiang Li, Lingwen Deng

Abstract

In this project, we tackle the complex task of text style transfer, focusing on transforming between Shakespearean English and modern news language. Employing a multifaceted approach, we experimented with directly prompting pre-trained GPT and T5 models, fine-tuning T5 for style swapping, and deploying dual encoders for distinct content and style extraction. Our methods also included latent vector splitting using T5 in conjunction with a pretrained-adversarial network, aiming to differentiate content and style vectors effectively. Finally, our team successfully extracted the features of two different styles by latent vector splitting and dual encoder methods, but we still need to make efforts to improve the readability of generated sentences.

Introduction

One big challenge within the field of NLP is text style transfer, which involves transforming the stylistic elements of a text while preserving its original content. This project is specifically on bidirectional style transferring between Shakespearean style and modern news style. The Shakespearean style is a form of old English used in dramatic dialogues, known for its intricate vocabulary, metaphor-rich language, and rhythmic variation. Modern news style is typically factual and straightforward which focuses on delivering information efficiently. This work not only engages with the nuances of computational linguistics and natural language processing but also seeks to bridge the gap between the ornate language of Shakespearean literature and the straightforward, factual style of modern news reporting. Working on style transfer between these two styles aims at creating a bridge between old and new literature styles and capturing the complexities of human language. Our approach could further contribute to other literature styles, for example, poetry, speech, etc.

Data

Resource1: https://huggingface.co/datasets/tiny_shakespeare
Resource2:https://huggingface.co/datasets/cnn_dailymail/viewer/1.0.0/train?p=2&row=213

We used the data listed above and they are both from the Hugingface website. Resource1 contains 40,000 lines of Shakespeare from a variety of Shakespeare's plays and resource 2 contains sentences from CNN News data. These data are combined through preprocessing with labels of 0 and 1 indicating the style. We did some basic tokenization techniques to the row data leaving 7887 rows of sentences from each style.

Methodology

T5, or Text-to-Text Transfer Transformer, is a versatile language model developed by Google Research. The key innovation of T5 is its unified framework that treats every natural language processing (NLP) task as a "text-to-text" problem. This means that T5 takes text input and produces new text output, regardless of the nature of the task, whether it's translation, summarization, question answering, or any other NLP task.

1. Prompting

To perform prompting, an LSTM classifier was designed and trained for text classification tasks, which specifically distinguishes between the Shakespearean styles and the news style. It is architected as a sequence processing model with an embedding layer to vectorize input tokens, followed by an LSTM layer to capture sequential patterns, and a linear layer to classify the output into style categories based on the hidden state of the LSTM. This LSTM classifier determines the probability that a sentence belongs to the styles, and the classifier is further used throughout all further approaches for the purpose of evaluation.

The method of prompting involves directly putting sentences into pre-trained models (GPT2, GPT3.5, T5) and prompting the model to perform style transfer to another style without explicit style extraction. Different ways of promoting models experiment and the best way of prompting is presented for each model as follows. There are two approaches, one is directly prompting T5 to transfer the style of a single sentence to a given target style, and the other one is to randomly pair the sentence and prompt pre-trained models to transfer the text to the target sentence style. Prompting also works as a navigation of different models, and ends up with a choice on T5

1.1Prompting Single Sentences

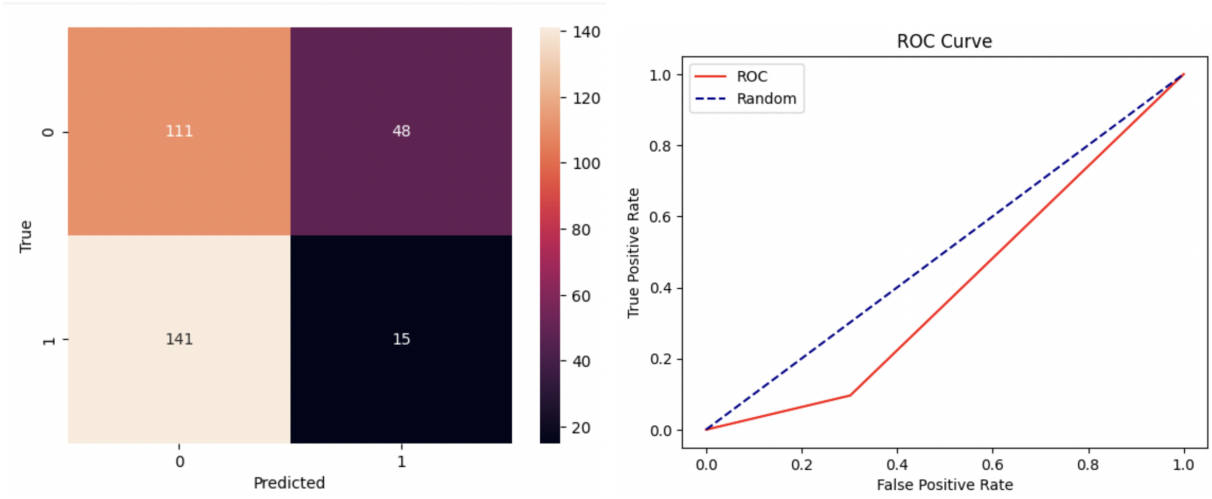
1.1.1 GPT 3.5

To start with, GPT-3.5 is being used for prompting since it is a relatively new LLM. The LSTM classifier is first used to determine the style of the sentence, and a prompt is constructed to either transfer from Shakespearean to news style or vice versa. The prompt is fed into OpenAI's text-Davinci-003 model to generate the transformed sentence.

We also use the GPT model to generate samples as a control group. These are three generated samples by utilizing GP-T3.5, through human feedback, GPT-3.5 shows a good understanding and performance of style while keeping the main content. It transfers "before November 2006 disappearance" to "Ere the disappearance in November 2006" when transferring to a Shakespearean style, which has a more complex syntax and an archaic term for "before". For transferring to the new style, it transfers "ANTONIO: Widow! a pox o' that!" to "Antonio expressed shock", which transfers the Shakespearean dialogue into a modern narrative news style summary, which captures Antonio's emotion and describes in a modern and straightforward way. Through our observations on the results, GPT-3.5 has the capacity to perform this specific style transfer.

```
origional sentence: She won court battle before November 2006 disappearance
generated sentence: Ere the disappearance in November 2006, she had the victory in the courtly fray.
====================
origional sentence: ANTONIO:Widow! a pox o' that! How came that widow in?widow Dido!SEBASTIAN:What if he had said 'widower AEneas' too? G
generated sentence: Antonio expressed shock at the mention of widow Dido. Sebastian jokingly asked what if Antonio had said 'widower Aene
====================
origional sentence: Members of the public allegedly attempted to gas and even spear the animal before she was freed
generated sentence: Verily, 'tis said that members of the populace didst attempt to gas and even lance the creature afore she was freed.
```

The result is further analyzed quantitatively. By classifying the generated sentences and comparing them to their targeted style, the accuracy score of the prediction is 0.4. According to the Receiver Operating Characteristic (ROC) curve and the confusion matrix, the model does not perform as well as a random chance which is different from previous observations on generated sentences. The difference might be due to the LSTM classifier being trained by the data different from GPT-3.5, so that the classifier may interpret the style not in a completely same way. Moreover, though there are differences in wording between sentences, such differences might not be considered enough to make the sentence classified as the new style. GPT-3.5, though powerful, might not be pre-trained for this specific task.

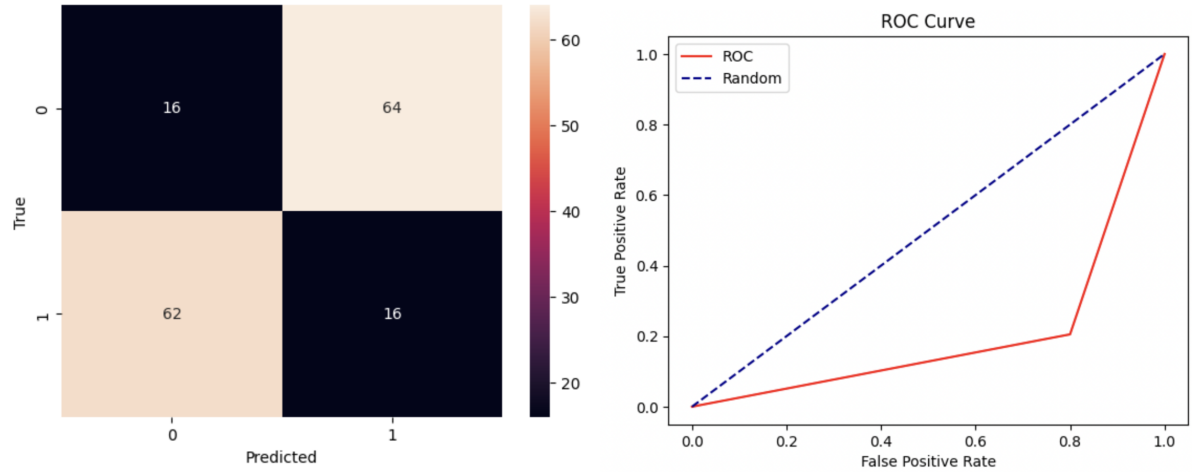


1.1.2 T5: T5-Small and T5-Base

To further explore the differences between models, T5-small model is used for the similar approach of prompting. Different from prompting GPT-3.5, during preprocessing, the prompts are tokenized by a T5-tokenizer and a mask technique is applied for contextual learning. The sample outputs are as follows.

```
transfer to shakespeare style: They say primates developed the ability to make groups recognisable and stop inter
ableerespectingrespecting.
==================
transfer to news style: CAMILLO:This shows a sound affection
BR..BRILLO. a greatBRILLO
==================
transfer to shakespeare style: Anti-Defamation League, Simon Wiesenthal Center say it uses anti-Semitic imagery
pepeare. Anti-Defamation League, Simon Wiesenthalpe
```

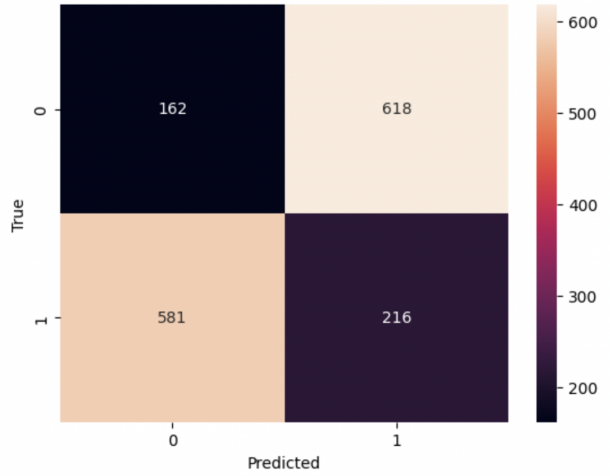
These generated sentences from prompting T5-small are mostly garbled text, meaning less punctuations and not showing significant stylistic changes, which largely retains the structure of original sentences. It has an accuracy score of 0.20, which is expected from the observed sentences. The ROC curves should show that the performance of the model is not as good as the random curve, which suggests a failure on this model, and the confusion matrix shows there is a relatively equal accuracy between transferring in both directions.

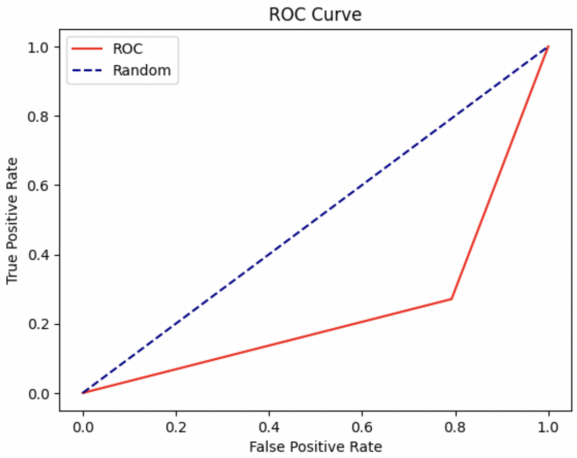


Since T5-small does not have a good performance on prompting for single sentences, the T5-base model is being experimented with since it is larger with more parameters. T5-base has a slight improvement with an accuracy score of 0.24 and a slightly better ROC curve, while the generated texts show a similar

pattern of meaningless or repeating sentences.

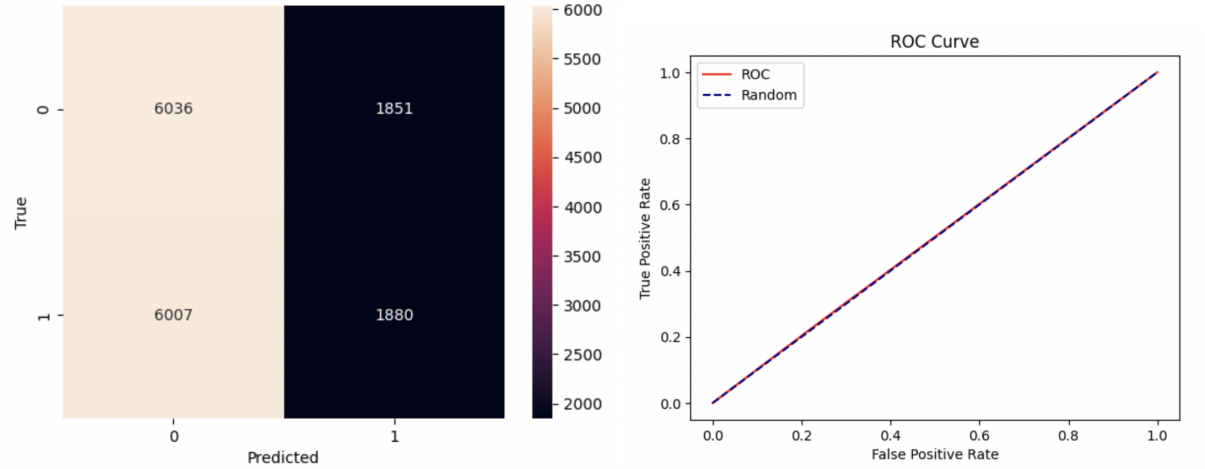
```
transfer to news style: ISABELLA:The better, given me by so holy a man
ableableableableableable,,
transfer to news style: GRUMIO:Am I but three inches? why, thy horn is a foot; andso long am I at the least
a foot, or two end?
transfer to shakespeare style: Met Office says the UK can expect a warm and mostly dry weekend with temperat
weather thisareare theare. Temperatures are expected
```





The preprocessing process is further modified by using T5-base, which mask is not used and just using T5-tokenizer for tokenization. From the observation, it shows that the generated sentences become more coherent. Instead of just repeating the original sentence, the generated started to make little modifications to wording and sentence structure. For example, in the first sentence, when transferring to a Shakespearean style text, the word “accusations” is used instead of “claims”, which is a more complex and less straightforward word. Also, in the last sentence, the part which shows strong emotion in the dialogue is deleted and only the informative part is left, which shows an improvement in the text generated. In this approach of prompting T5-base, there is some successful change in original sentences compared to the previous tests on T5. The accuracy rate is 0.502 and the ROC curve shows a slightly better performance than the random one and has a big improvement compared to the previous approaches. The confusion matrix shows a similar performance on transferring style in both directions. The improvement might be because without masking, the model can have access to the full context of input, and reference to all parts of the sentence while generating, and since the prompt is longer, using a non-mask approach can increase coherence.

```
origional sentence: Claims she suffered catalogue of abuse at hands of Italian former partner
/usr/local/lib/python3.10/dist-packages/transformers/generation/utils.py:1273: UserWarning: Using
  warnings.warn(
transformed sentence: Accusations she suffered catalogue of abuse at hands of Italian former part
========================
origional sentence: Six crew and 158 passengers evacuated from American Airlines flight
transformed sentence: American Airlines flight 158: American Airlines evacuates six crew members
========================
origional sentence: ISABELLA:O just but severe law!I had a brother, then
transformed sentence: : ISABELLA:O a brother, then a sister, then
------------------------
```



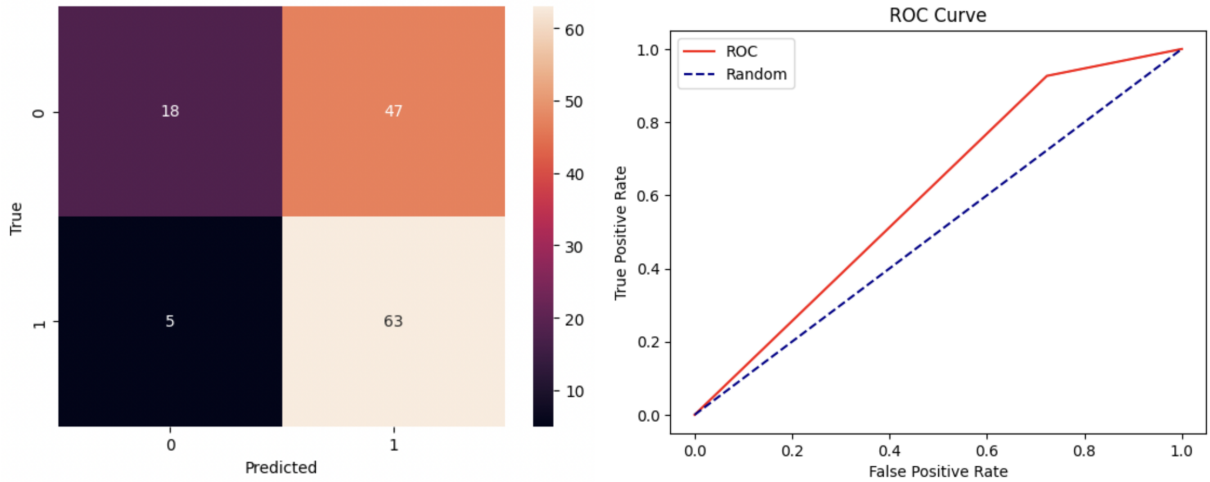
1.2 Prompting Paired Sentences

As prompting T5-base turns out to be the best choice for prompting pre-train models, there is another approach of randomly pairing the sentence with a target sentence selected from the other style and prompting pre-trained models to transfer the text to the target sentence style. The randomly selected sentence can work as a reference for the target text style. During the preprocessing, each input sentence is paired with another randomly selected text and constructs a prompt that is further tokenized and masked. GPT-2 and T5-small are being prompted in this way.

1.2.1 GPT-2

GPT-2 is prompted on the paired data mentioned above and generates the following sentences. As the sample output shows as follows, the generated sentence does not exhibit this change and merely repeats the task instruction. According to the confusion matrix, transferring to Shakespearean has a lower accuracy than transferring to the news style, though there are relatively fewer texts transferred to the news style within the dataset. This model of GPT-2 has an accuracy score of 0.61, and the ROC curve shows its performance is better than the random one. The GPT-2 shows a good performance quantitatively, though the generated sentence does not show any transformation in style. These generated sentences due to GPT-2 are not pre-trained for such text-style transfer tasks.

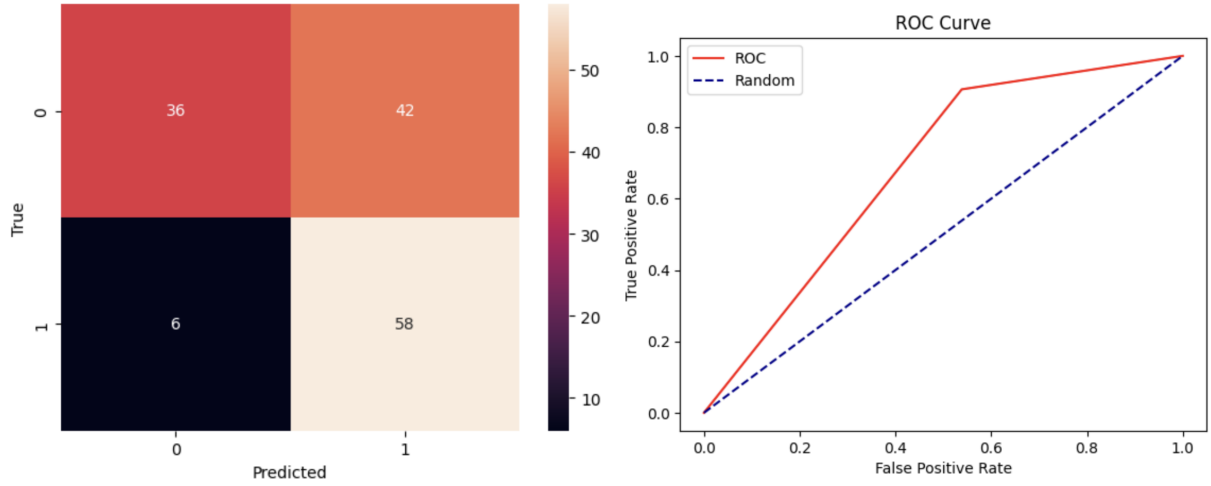
```
Original Sentence: BUCKINGHAM:I go: and towards three or four o'clockLook for the news that the Guildhall affords
Target Sentence style: Clinton appears on Rachael Ray on Friday and talks babies, first kisses, and her professional life
Generated Sentence: transform the written style of BUCKINGHAM:I go: and towards three or four o'clockLook for the news that the Guildhall
==========
Original Sentence: But that still use of grief makes wild grief tame,My tongue should to thy ears not name my boysTill that my nails were
Target Sentence style: Report found family finances have been battered over past decade
Generated Sentence: transform the written style of But that still use of grief makes wild grief tame,My tongue should to thy ears not nam
==========
Original Sentence: The Vtech Kidizoom smartwatch has a built-in motion sensor and games
Target Sentence style: Lord:Go, sirrah, take them to the buttery,And give them friendly welcome every one:Let them want nothing that my h
Generated Sentence: transform the written style of The Vtech Kidizoom smartwatch has a built-in motion sensor and games to the written st
```



1.2.2 T5-small

T5 is prompted the same way as the GPT-2 which uses the randomly paired data and preprocesses by tokenizing and masking. The generated sentences are as follows, which is better than the sentences generated by GPT-2. Though there is not much transformation of style, it does not involve the complete prompt and has some variation, though it is still mostly repeating the original sentences and the target sentence. The quantitative results observed from the ROC curve is slightly better than the GPT-2 performance. The confusion matrix shows a similar pattern that transferring to news style has a better performance than transferring to the Shakespearean style.

```
Original Sentence: Starc says the victory can help reopen England's scars from the Ashes
Target Sentence style: A shepherd's daughter,And what to her adheres, which follows after,Is the argument of Time
Generated Sentence: Starc says the victory can help reopen England's scars from the Ashes to the written style of A shephe
Original Sentence: What is that curt'sy worth? or those doves' eyes,Which can make gods forsworn? I melt, and am notOf st
Target Sentence style: An intern erroneously confirmed the names of the flight crew, the NTSB says
Generated Sentence: I melt, and am notOf stronger earth than others to the written style of An intern erroneously confirme
```

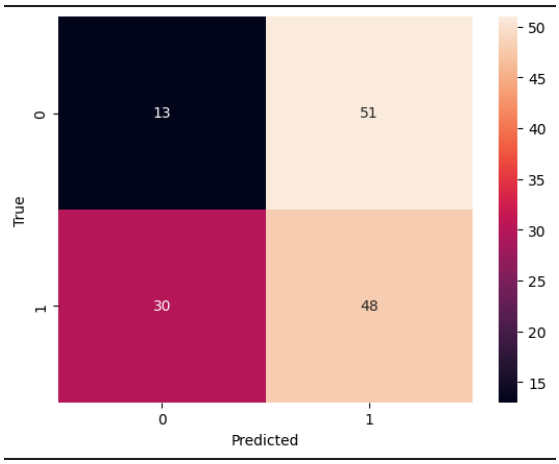


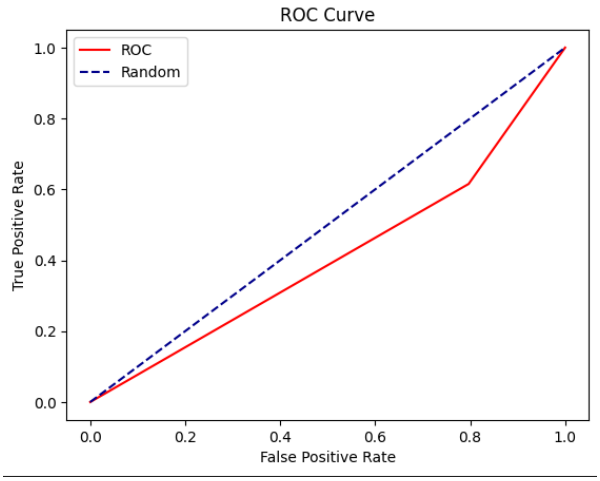
1.2.3 T5-Large

We also tried the T5-Large model with prompting sentence: transform the written style of {source_sentence} to the written style of {target_sentence}. The generated sentence and the evaluation curve are shown in the following and we can see that the result is even worse than the t5-small model.

The classification result is worse than the random variable according to the ROC curve and there are more negative false results.

```
Original Sentence: Ex-Rep
Target Sentence style: Welcome hither,As is the spring to the earth
Generated Sentence: -Rep to the written style of Welcome hither,As is the spring to the earth.....-Rep..... As is the spring to the earth. As is the spring to the earth.
========================================
Original Sentence: I shall tell youA pretty tale: it may be you have heard it;But, since it serves my purpose, I will ventureTo stale 't a little more
Target Sentence style: NEW: The Israel Defense Forces declined to comment on the report
Generated Sentence: . The.. The Israeli Defense Forces declined to comment on the report. The Israeli Defense Forces declined to comment on the report. The Israeli Defense Forces declined to comment.. The Isr
========================================
Original Sentence: Fie, fie, thou shamest thy shape, thy love, thy wit;Which, like a usurer, abound'st in all,And usest none in that true use indeedWhich should bedeck thy shape, thy love, thy wit:Thy noble s
Target Sentence style: 650,000 sufferers in the UK
Generated Sentence: the written style of 650,000 sufferers in the UK transform the written style of transform the written style of transform the written style of 650,000 sufferers in the UK transform the writ
========================================
Original Sentence: LADY ANNE:What, do you tremble? are you all afraid?Alas, I blame you not; for you are mortal,And mortal eyes cannot endure the devil
Target Sentence style: A nuclear test would also fit North Korea's pattern of serial provocations, analysts said
Generated Sentence: of LADY ANNE:What, do you tremble? are you all afraid? to the written style of LADY ANNE:What, do you tremble? are you all afraid? to the written style of LADY ANNE:What, do you tremble? a
========================================
Original Sentence: GLOUCESTER:These eyes could never endure sweet beauty's wreck;You should not blemish it, if I stood by:As all the world is cheered by the sun,So I by that; it is my day, my life
Target Sentence style: military base in Britain
Generated Sentence: the written style of GLOUCESTER:To:To of GLOUCESTER:To transform the written style of GLOUCESTER:To:To transform the written style of GLOUCESTER:To transform the written style of GLOUCESTI
========================================
```





2. Fine-tune T5 + Prompt

Through the previous exploration of models through prompting, T5 turns out to be the best one to use for further approaches, due to its accessibility and its overall performance. Under the consideration of the size issue for fine-tuning, T5-small is chosen instead of T5-base for a more efficient fine-tuning process. Since there are nuanced differences between different styles and T5 was not pre-trained on the task for text style transfer, T5 was fine-tuned to capture the stylistic nuances and learn about these two text styles. This approach does not involve feature extractions, it prompts and fine-tunes T5 about different styles. Similar to prompting pre-trained models, T5 is pre-trained on single sentences and randomly selected paired sentences.

2.1 Fine-Tuning Single Sentences

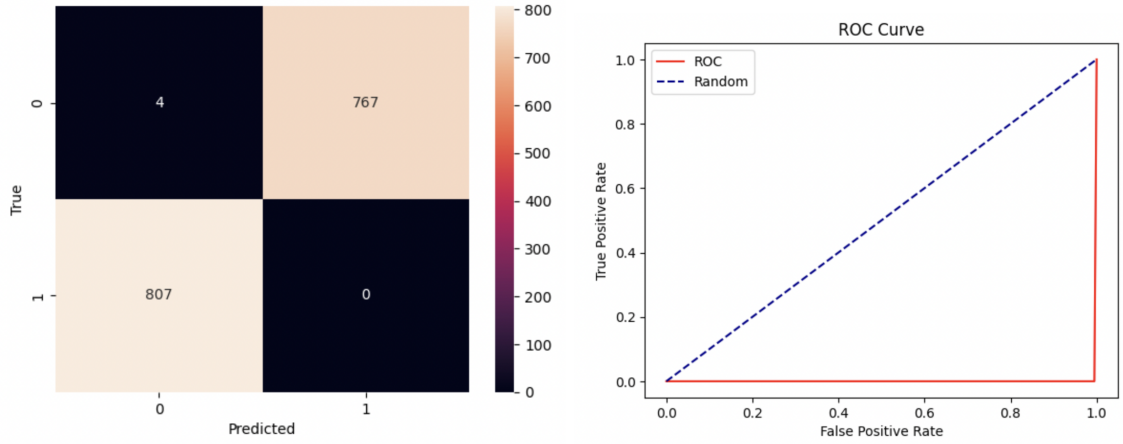
Three approaches of fine-tuning single sentences are taken, which are fine-tuning T5 on the prompt of a single sentence with its corresponding style and swap style for style transfer, fine-tuning T5 by letting the sentence be the encoder input while letting the style be the decoder input, and fine-tuning T5 by customizing the loss function.

2.1.1 Fine-Tuning Sentence with Style and Style Swap

The first simple approach to fine-tuning single sentences is by constructing a prompt of the sentence itself and the style it belongs to, and set the target text as itself, in this way to link the sentences and the text style together and enable T5 to learn about these two styles. When generating the text of a new style, construct the prompt of the sentence and its new style, indicating the target style of this generation task.

Since T5 learned the two sentence styles from the fine-tuning process, it would be able to generate new sentences in target style.
However, after implementing this method and observing the generated sentences, it shows that there is almost no difference between the generated sentences and the original one. Also, the ROC curve shows this failure as well. These observations indicate that the T5 model failed to learn about and have the ability to identify these two models at all. The failure of this model firstly is because of the capacity of T5, since T5 is not trained for any text style transfer task, therefore, it is hard for T5 to identify style through this way of prompting. Also, there is not much difference in the way of prompting for different styles, therefore, it also makes it difficult for that prompting in fine-tuning to be effective. The insufficient data in fine-tuning also increases the difficulty.

```
AUFIDIUS:I have not deserved it
generated sentence: AUFIDIUS:I have not deserved it
==========
LARTIUS:So, let the ports be guarded: keep your duties,As I have set them down
generated sentence: LARTIUS:So, let the ports be guarded: keep your duties,As I have set them down
==========
frustrated by heavy tax burdens and rising costs
generated sentence: frustrated by heavy tax burdens and rising costs
```



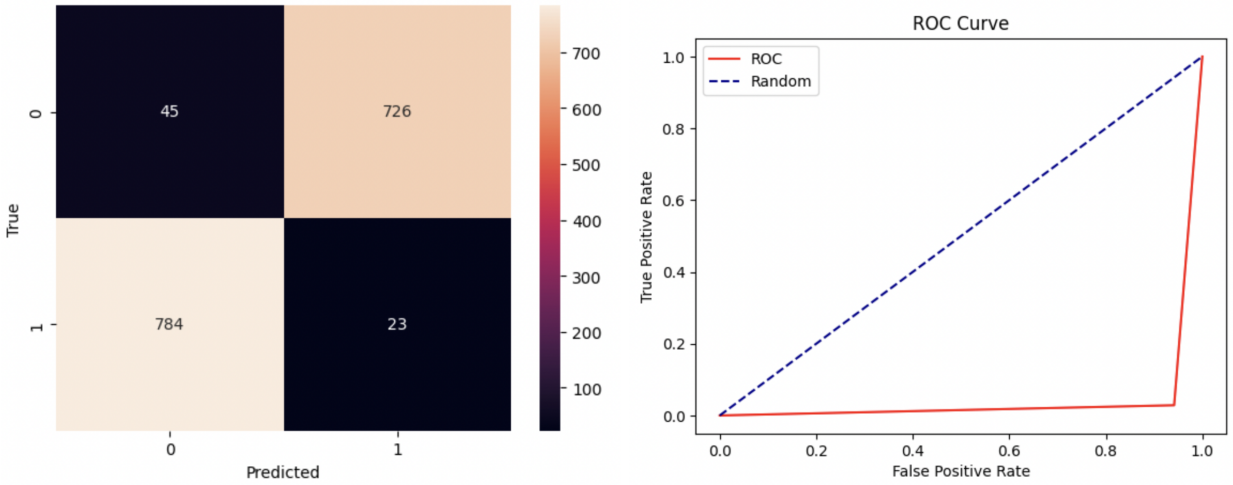
2.1.2 Sentence as Encoder Input and Style as Decoder Input

Due to the failure of fine-tuning T5 with the sentence and corresponding style, in order to enhance the separation between the sentence and style when prompting and make style-specific generation, the decoder and encoder to T5 are fine-tuned. Similar to fine-tuning T5 with style, in order to make T5 know the target style, the style is necessary as a part of the input to make T5 learn about the differences and details of each detail. In the new fine-tuned model, the sentence that needs to be transferred in style works as the encoder input while the target style in the decoder output. Similarly, the data is preprocessed through the T5 tokenizer, and masking is not used due to experiments. By fine-tuning the encoder, the ability to comprehend the linguistic and stylistic nuance is enhanced, and more focus can be spent on the sentence itself instead of having a misleading style together with the sentence. The design of making sentence style as decoder input has emphasized the difference between styles especially from the generation step, and fine-tuning decoder makes it better possible to generate text with the target style's characteristics.

The generated text of this approach is relatively coherent and has some modification of the original text instead of presenting the completely same text. However, the generated texts are still mostly copying the original text though it might present the words in a different structure, and the sentences still seem somehow meaningless and little repeating. However, it shows some good signs, for example, when translating to news style, the format to dialogue is removed. The confusion matrix shows this approach

has a slightly better performance on transferring Shakespearean style to news style. Also according to the ROC curve, there is so little effective and successful style transfer performed within the test data.

```
------------
What foreign liberty-seekers want from us is moral validation of their cause, he writes
generated_text:  from us is moral validation of their cause, he writes he writes he wants from us is moral
============
A Scotsman on holiday captured the strange encounter from his deck
generated_text:  on holiday captured the strange encounter from his deck. He captured the strange encounter
============
DUCHESS OF YORK:I will be mild and gentle in my speech
generated_text: :I will be mild and gentle in my speech, especially in my speech.
```



The failure of this model would firstly be due to the model capacity, as T5-small might not have enough complexity to handle the complexity of this specific style transfer task, especially under this insufficient amount of data. Also, T5 itself is not pre-trained for any text-style transfer task, which makes fine-tuning it for this specific task more difficult. There is a reduced loss when training and the model could fit the data better if there is time for training in more epochs, which might increase the performance. Moreover, though the model shows some signs of performance style transfer in the right direction and modification of original text compared to the previous approach, these changes are not enough to make generated sentences a completely new type.

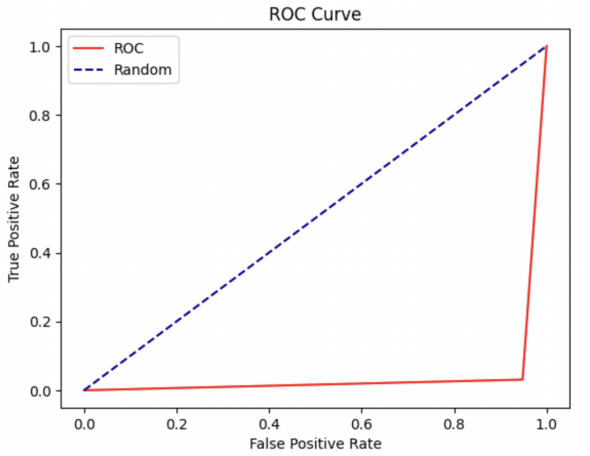
2.1.3 Customizing Loss Function

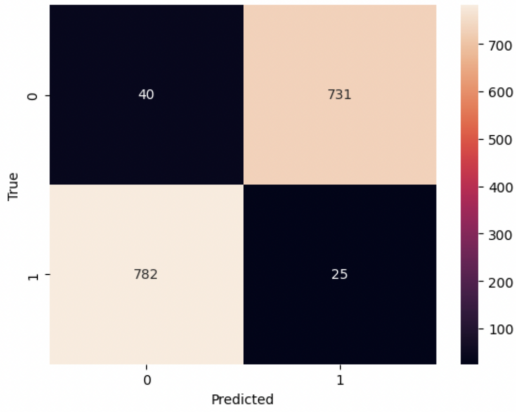
After observing the previous performance of the models, it is a common result that the generated texts do not perform much style transfer and just repeat the original text in some order. Therefore, we decided to customize the loss function by utilizing the LSTM classifier we built before and also involves the cosine similarity. This approach is built on the previous approach of fine-tuning the encoder and decoder by adding the custom loss function instead of the original loss. Specifically, the customized loss function involves three parts, the original loss, the style discrimination loss, and a content preservation metric. The style discrimination loss is provided by a discriminator model, which specifically is our LSTM classifier, which could encourage the model to generate text style to be closer to the target style. A Content Preservation Metric is implemented through embedding representations of the original and generated texts, followed by computing their cosine similarity. This metric is an indicator of content retention, which ensures the content remains unchanged. The overall loss function is a weighted sum of these three losses, specifically total_loss=loss+0.5×discrimination_loss+0.5×content_loss.

The generated sample is quite similar to the generated ones from the model without customized loss, which shows the pattern of mostly repeating the original sentence but in some different order, and it would expand the sentence through repeating. The ROC curve and the confusion matrix are mostly the same, which shows there is not much effect of customizing his loss. One possibility is that this form of output is mostly due to the encoder and decoder input design, so customizing the loss in this way does not

provide enough impact to be observed from the generated sentences. Another possibility is due to the weight of three different losses, which would determine whether it would value the style more or it would value the content more. Therefore, further experiments on the weights can be tried that might improve the performance, and other weights can also be considered to be involved in this customized total loss.

```
------------
BAPTISTA:KATHARINA:No shame but mine: I must, forsooth, be forcedTo give my hand opposed against my heartUnto a mad-brain rudesby full of spleer
generated_text: I must, forsooth, be forced to give my hand against my heartUnto a mad-brain rudesby full of spleen;Who woo'd in haste and means
============
Servant:What, think you then the king shall be deposed?Gardener:Depress'd he is already, and deposed'Tis doubt he will be: letters came last nig
generated_text: , think you then the king shall be deposed?Gardener:Depress'd he is already, and deposed'Tis doubt he will be: letters came lasi
============
GRUMIO:Ay, but the mustard is too hot a little
generated_text: y, but the mustard is too hot a little, but the mustard is too hot a little bit. GRUMIO:It's too hot the mustard is too hot a li
------------
```



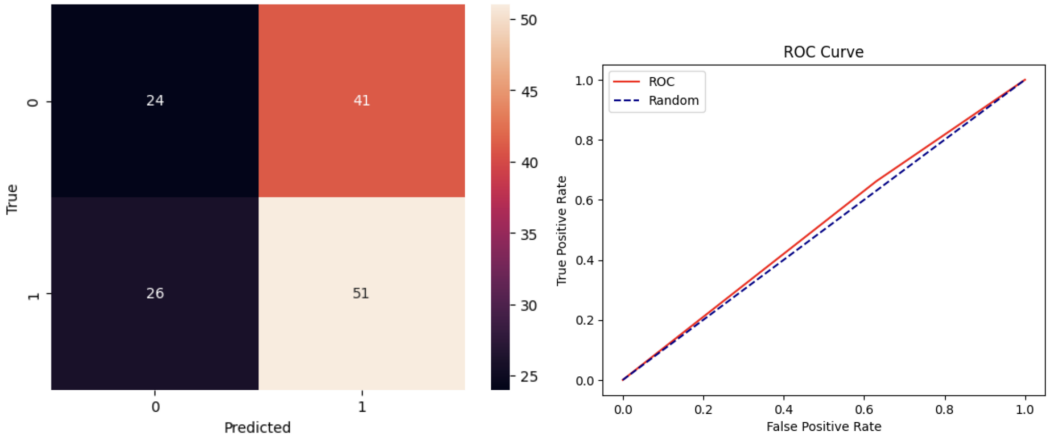


2.2 Fine-Tuning on Pair Sentences

2.2.1 Direct Fine-Tuning on Randomly Paired Sentences

Similar to the prompt part, each training data is paired with a randomly selected sentence from the other style as the target sentence style. Such data is used for fine-tuning, and it turns out the model performs slightly better than the random variable according to the ROC curve. This shows this model is relatively successful, though it is surprising that it does not perform as well as the same way of prompting. There is a change of structure of the original sentence and target sentence in generated sentences while it is mostly a combination of the content in the original and target one. The generated sentence results look similar to the prompt one, showing it is hard to just use the style of the target sentence but not have its content. The decrease in performance might be due to insufficient performance which is not enough to train T5 to accept prompting and extract style and content in such a way.

```
Original Sentence: SICINIUS:One thus descended,That hath beside well in his person wroughtTo be set high in place, we did commendTo your remembrances: but you have found,Scaling h
Target Sentence style: PM predicts the next election will be re-run of '92 when Major beat Kinnock
Generated Sentence: SICINIUS:One thus descended,That hath beside well in his person wroughtTo be set high in place, and revokeYour sudden approbation to the written style of PM pr
Original Sentence: O cursed wretch,That knew'st this was the prince,and wouldst adventureTo mingle faith with him! Undone! undone!If I might die within this hour, I have livedTo d
Target Sentence style: But with Vincent Kompany out injured, now is the time for Mangala to shine
Generated Sentence: is the time for Mangala to shine in the written style of O cursed wretch,That knew'st this was the prince,and wouldst adventureTo mingle faith with him! Undone
```



2.2.2 Fine-Tuning with Two Separate Models

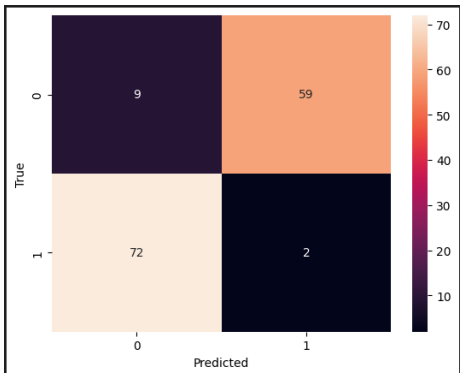
According to what we have before, the generated sentences seem to just combine with two different sentences together as a new output. So we decided to try fine-tuning with two different models and use these different models to generate sentences based on the sentence style in the test set.

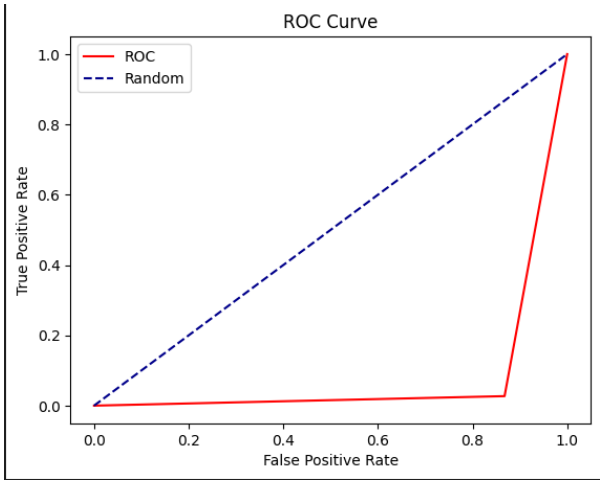
The following picture shows the layout of our data input. We split this data set into two sub-data. The first sub-dataset contains all the rows with transfer_syle=1 which means transfer style CNN to Sheakspear and vice versa.

We train two models Model1 is trained on sub-data 1 while Model2 is trained on sub-data2. The final result is shown following. The result is worse compared to the baseline. The last picture shows the generated sentence. The interesting thing is that the generated sentences do not just simply combine two different style sentences together but rather disorganize and then reorganize. Also when the two sentences are too different from each other, distance for example. The two-model method will not generate new sentences. We believe that may be a potential reason to influence the roc curve.

	sentence	label	input_text	target_text	transform_style
10934	The 35-year-old denies one charge of pervertin...	0	The 35-year-old denies one charge of pervertin...	GLOUCESTER:Your beauty was the cause of that e...	1
15658	GRUMIO:There	1	GRUMIO:There	Thousands watch activities of Korean woman, du...	0
13357	LUCIO:She it is	1	LUCIO:She it is	The music is played through whistle openings o...	0
4927	Here comes a gentleman that haply knows more	1	Here comes a gentleman that haply knows more	Russian fishermen net a huge Steller sea lion ...	0
6203	DUKE VINCENTIO:Why, you are nothing then: neit...	1	DUKE VINCENTIO:Why, you are nothing then: neit...	Major search was launched, with 100 locals joi...	0
...	...	...	...	...	...
794	PAULINA:Had our prince,Jewel of children, seen...	1	PAULINA:Had our prince,Jewel of children, seen...	The group has 600 paid youth apprenticeship pl...	0
15178	Title of the new western-themed episode is "Ha...	0	Title of the new western-themed episode is "Ha...	First Murderer:So do not I: go, coward as thou...	1
12178	Cupp: At press conference, Hillary Clinton ins...	0	Cupp: At press conference, Hillary Clinton ins...	The benefit thereof is always grantedTo those ...	1
9644	GLOUCESTER:What! threat you me with telling of...	1	GLOUCESTER:What! threat you me with telling of...	Oxford, Manchester, Cardiff and Sheffield amon...	0
6176	Common Core, adopted by 44 states, is a set of...	0	Common Core, adopted by 44 states, is a set of...	Here comes your father: never make denial:I mu...	1

14054 rows × 5 columns






```
Original Sentence: Had an 11-hour operation in which a titanium rod was fitted to her spine
Target Sentence style: TYBALT:I am for you
Generated Sentence:
========================================
Original Sentence: Kennedy family has never confirmed or denied the story
Target Sentence style: 'QUEEN ELIZABETH:Thus have you breathed your curse against yourself
Generated Sentence: Kennedy family has never confirmed or denied the story to the written style of 'QUEEN ELIZABETH:Thus have you breathed your curse against yourself'
========================================
Original Sentence: MONTAGUE:Who set this ancient quarrel new abroach?Speak, nephew, were you by when it began?BENVOLIO:Here were the servants of your adversary,And yours, close fighting ere I did approach:I
Target Sentence style: The match will be Van Gaal's first competitive match in charge of the Reds
Generated Sentence: In the instant camethe fiery Tybalt, with his sword prepared,Which, as he breathed defiance to my ears,Which, as he breathed defiance to my ears,Which, as he breathed defiance to my ears,
========================================
Original Sentence: First Lord:The higher powers forbid!PAULINA:I say she's dead; I'll swear't
Target Sentence style: Marine Le Pen is leader of France's far-right Front National party
Generated Sentence: The higher powers forbid!PULINA:I say she's dead; I'll swear't to the written style of Marine Le Pen is leader of France's far-right Front National party.I say she's dead; I'll swear't to
========================================
Original Sentence: ROMEO:Then plainly know my heart's dear love is setOn the fair daughter of rich Capulet:As mine on hers, so hers is set on mine;And all combined, save what thou must combineBy holy marriag
Target Sentence style: Facebook is considering offering more data or creating an internal think tank
Generated Sentence: I pray,That thou consent to marry us to-day to the written style of ROMEO:Then plainly know my heart's dear love is setOn the fair daughter of rich Capulet:As mine on hers, so hers is set
```

More Advanced Methods

After experimenting with T5-small, T5-base, and GPT-3.5 models, our team explored content and style extraction, guided by the works of Di Jin et al. on deep learning for text style transfer. We applied Latent Representation Splitting and two-encoder approaches to our task. Our findings indicate that Latent Representation Splitting was ineffective in separating content and style, as well as in creating style-transformed sentences. However, the two-encoder method did manage to isolate content and style vectors but also struggled to produce style-transformed sentences. We will delve into each method's neural network architecture, loss functions, training strategies, and performance metrics. Both methods utilized the T5-small model for training. Data preprocessing involved creating training sets with tokenized sentences from CNN and Shakespeare's works, style labels, and attention masks, all fed into an LSTM classifier for style identification.

3. Latent Representation Splitting

In the Latent Representation Splitting approach, we attempted to separate the latent representations obtained from the T5-small model into two distinct components: content and style. We hope that a given piece of text could be decomposed into its semantic content and stylistic elements, which could then be manipulated independently.

3.1 Training Method

For our style transfer task, we used a modified T5-small model architecture with an encoder and decoder. The encoder processes input sentences to form a latent space representation of [batch size, maximum sentence length, 512], with 512 being the latent dimension for each token. Contrary to our initial assumption that style vectors within a sentence would be similar, our approach divided the 512-dimension vector into two parts: 384 for content and 128 for style. We then swapped the 128-dimension style vectors between sentences. The decoder, also derived from T5-small, was fed a merged latent vector containing

the content from one sentence and the style from another, intending to generate text that combines the content's meaning with the new style. The graph illustrates how we manipulate the latent vector's hidden state and use an empty decoder input to start sentence generation.

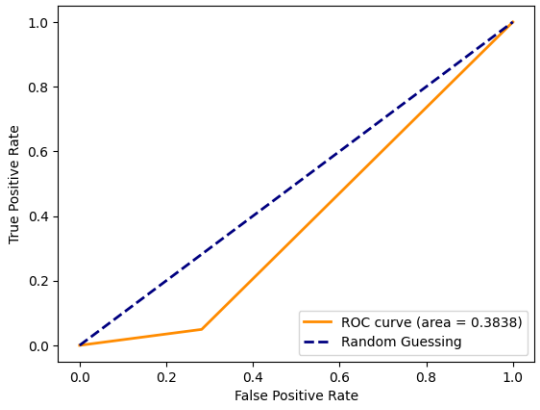
3.2 Loss Evaluation

We use two components to create a dual-objective loss function. The first component was a reconstruction loss, which ensured that the model could accurately reconstruct the original sentence from its latent representation. The second component was a style-transfer loss, computed using a pre-train LSTM classifier to distinguish between different styles, which encouraged the network to generate sentences that were stylistically aligned with the target style sentences. We use this loss function to ensure that the splitting of the latent vector did not hinder the model's ability to reproduce the original text accurately and progressively tune the model to generate sentences that retained the original content and adopted the desired style.

The following picture shows an example of transferring content.

```
original text1:  Chinese demand government clamps down on terrorism
transfer text1:  government clamps down on terrorism
original text2:  First Senator:There's no remedy;Unless, by not so doing, our good cityCleave in the midst, and perish
transfer text2:  :'s no;, by not so doing, our good cityCleave in the midst, and perish
```

In our examination of the model's outputs, we observed that while the generated sentences remained coherent and largely preserved the original information, the transformation in writing style was not as pronounced as anticipated. We believe this suggested that while the sentences were readable and the core content was retained, the model struggled to capture and replicate the deeper stylistic nuances of the target genres. Based on the roc curve of this method, we found that the model failed to achieve the desired style transformation. Moreover, the curve's proximity to the line of random guessing underscored the model's limited capability in style manipulation (as shown in the graph below).



4. Two Encoder Methods for Content and Style Extraction

Moving on from the Latent Representation Splitting method, our team explored an alternative approach to address previously encountered limitations. This approach, referred to as the "Two Encoder Method," involves using separate encoders for content and style, allowing for a more distinct and controlled extraction of these elements.

The Two Encoder Method is based on an architecture involving two separate encoders – one for content and the other for style – and a single decoder. This design choice was driven by our aim to isolate content and style features in the text more effectively.

The following text show the pseudo-code for this encoder:

Class T5Encoder

Initialize:

Load the pretrained T5-small model as the encoder

Forward(input_ids, attention_mask):

Pass input_ids and attention_mask through the T5-small encoder

Return the encoder outputs

Class T5Decoder

Initialize:

Load the pretrained T5-small model as the decoder

Forward(encoder_output, decoder_input_ids):

Pass encoder_output and decoder_input_ids through the T5-small decoder

Return the logits from the decoder outputs

Class StyleTransformModelEncoder

Initialize:

Create an instance of T5Encoder for content called content_encoder

Create an instance of T5Encoder for style called style_encoder

Forward(input_ids1, attention_mask1, input_ids2, attention_mask2):

Get content and style outputs from content_encoder using input_ids1, attention_mask1

Get content and style outputs from style_encoder using input_ids2, attention_mask2

Return content and style outputs for both sets of inputs

Class StyleTransformModelDecoder

Initialize:

Create an instance of T5 Decoder called decoder

Forward(original_encoder_output1, original_encoder_output2, transfer_encoder_output1, transfer_encoder_output2, decoder_input_ids1, decoder_input_ids2):

Generate original sentence 1 using original_encoder_output1 and decoder_input_ids1

Generate original sentence 2 using original_encoder_output2 and decoder_input_ids2

Generate style transferred sentence 1 using transfer_encoder_output1 and decoder_input_ids1

Generate style transferred sentence 2 using transfer_encoder_output2 and decoder_input_ids2

Return all four sentences

4.1 Training Strategy and Loss Function

We first try the method we did in the previous section which randomly picks a sentence in another style and makes it the target sentence then does the encoder based on the original sentence and the target sentences. However, the generated sentences are either all empty or randomly select a few words from the original sentence and repeat them. So we decided to use the decoder to reconstruct the original sentences Our training strategy involved training both encoders and the decoder simultaneously, using a multi-faceted loss function:

- Reconstruction Loss: Similar to the previous method, we employed a reconstruction loss to ensure that the decoder could accurately reconstruct the original sentence by using cross-entropy or CosineEmbeddingLoss
- Style Transfer Loss: We maintained the use of a pre-trained LSTM classifier to quantify the style transfer success. This loss function is a cross-entropy function to encourage the model to generate sentences that align with the target style.
- KL divergence Loss: During the training process, we realize we need our hidden state following some known distribution; otherwise, resampling the hidden state can't generate a stable outcome. Without KL loss, the decoder will fail to generate any sentence (literally noting).
- Sentence similar score: we add this loss function during the improvement step.
- Style Similarity and Difference Losses: To strengthen the distinct roles of each encoder, we introduced two losses by using Cosine similarity (GPT-generated loss function):
    - The style similarity loss encourages the style encoder to produce similar representations for different tokens within the same sentence, enhancing the coherence of style encoding. The formula is defined as the following where h_ij represents the hidden vector of the jth token and ith sentence of the batch.

$$Loss = \frac{1}{batch\_size} \sum_{i=1}^{batch\_size} \left(\frac{1}{N_i} \sum_{j=1}^{max\_sentence\_length} \sum_{k=j+1}^{max\_sentence\_length} \left| 1 - \cos(\vec{h}_{ij}, \vec{h}_{ik}) \right| \right)$$

    - The content difference loss encourages diversity in the content representations of different sentences, ensuring that the content encoder captures the unique semantic features of each sentence. The formula is defined as the following where S_i is the average vector for the ith sentences.

$$Loss = \frac{1}{N} \sum_{i=1}^{batch\_size} \sum_{k=i+1}^{batch\_size} \left| \frac{\vec{s}_i \cdot \vec{s}_k}{\|\vec{s}_i\| \|\vec{s}_k\|} \right|$$

Also, the latent vectors after re-sampling are beginning to be similar., as shown in the following figure. So, our team decided to add a loss function to penalize the similarity.

4.2 The Process of Style Transfer

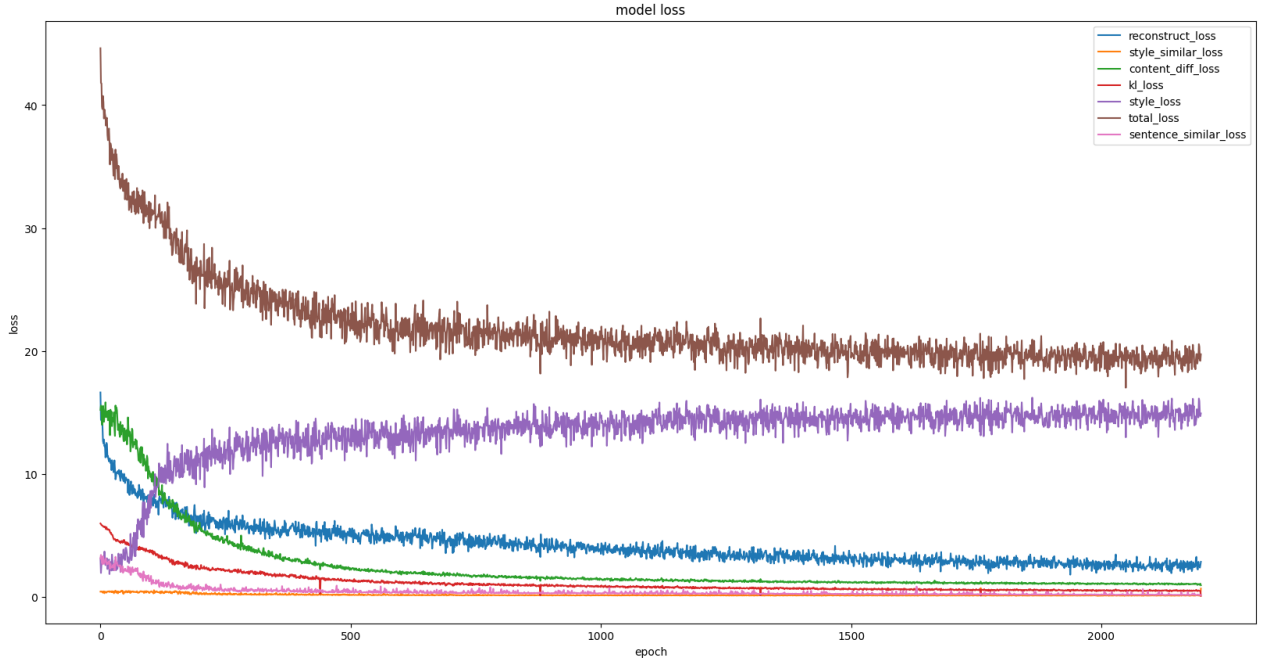
The style transfer process in our Two Encoder Method unfolds through a series of carefully orchestrated steps:

1. Input Processing: Both input sentences undergo parallel processing by the content and style encoders. The content encoder extracts the sentence's fundamental meaning and semantic elements, while the style encoder focuses on capturing the stylistic nuances.

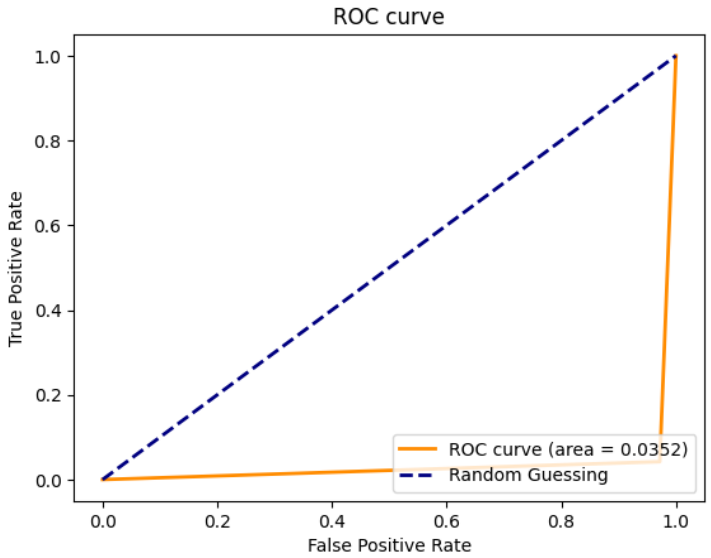
2. Latent Vector Reparameterization: We employ a reparameterization technique to blend content and style features. This involves using the content vector as the mean and the style vector as the log-variance. By doing this, we generate a new latent vector that samples from a distribution representing both content and style attributes. For example:
    a. suppose we have sentence1 and sentence2, the encoders will generate content_vector1, style_vector1 for sentence1, and content_vector2, style_vector2 for sentence2.
    b. For reconstructing, take sentence1 as an example. We use Gaussian sampling content_vector1 (mean) and style_vector1(log var) and get resampling_vector1. The resampling_vector1 will pass to the decoder and get the new sentence. This sentence should be the same as the original input sentence.
    c. For style transferring, take sentence1 as an example. We will use content_vector1 (mean) and style_vector2(log var) to sample sentence1_style2_vector. The sentence1_style2_vector will pass to the decoder and get the new style transfer sentence. This sentence should be classified as label 2 from the LSTM classifier.
3. Latent Vector Combination and Decoding: The new latent vector, a fusion of content and style, is then fed into the decoder. The decoder, trained to interpret and reconstruct textual data, takes this combined vector and generates a sentence that embodies the original content but with the newly adopted style.

4.3 Performance Evaluation

The following graph shows the final training process of this model. The total loss decreased as the number of training epochs increased. However, a significant trade-off was observed in our training objectives: the model sacrificed style loss to facilitate reductions in other losses, particularly the reconstruction loss.



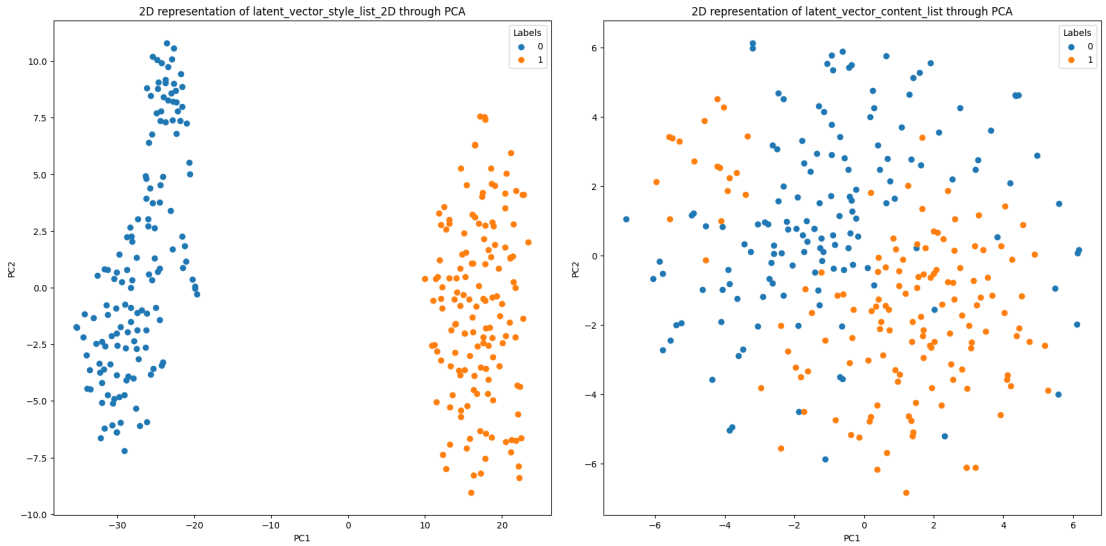

```
original text2:  Some experts recommend natural sources like zinc to boost performance
target text:  GLOUCESTER:Well thought upon; I have it here about me
recon text2:  Some experts recommend natural natural like zinc zinc boost boost performance performance
transfer text2:  Some experts recommend natural natural like zinc zinc boost boost performance performance
```



We believe there are two main arguments for the trade-off. Firstly, it suggests that the model struggles with transferring sentence style. As indicated by the ROC graph, the performance is even less effective than the vector splitting method. Most of the generated sentences are still classified as their original style. The graph also reveals that the reconstructed sentences and style-transferred sentences are nearly identical. The reasons behind this will be explored in the second implication.

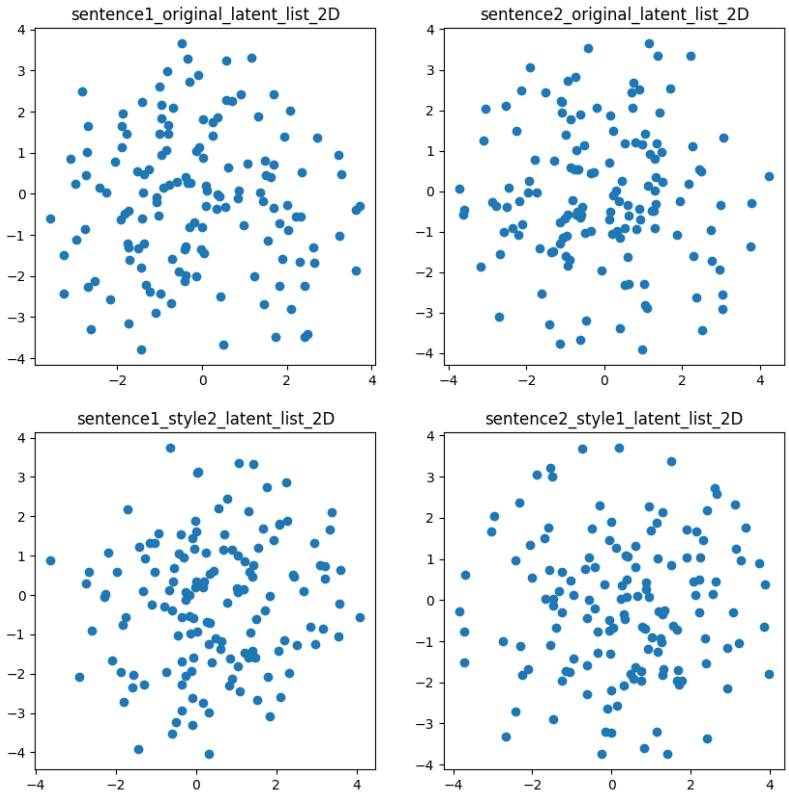
Secondly, the model meets other training objectives. As the subsequent graph shows, the style vectors of two labels are distinctly separated by the style encoder. This means that style vectors from the same labels tend to cluster together, forming two distinct clusters with a clear boundary.

Additionally, the content encoder effectively extracts content vectors for each sentence. While these content vectors are distributed sparsely in the latent space due to content variation, they still exhibit a boundary separating them into two clusters corresponding to two styles (Drama vs. News). This observation aligns with the findings in the subsequent graph.



The model's focus on sentence reconstruction over style transfer results in an unexpected outcome: swapping style vectors and sampling new latent vectors has little effect on the 2D representation. The post-sampling latent vectors, fed into the decoder, are nearly identical in 2D representation, showing that they retain much of the original information. This reflects a similar issue encountered in vector splitting, where the model struggles to effectively alter style while preserving content.

Nevertheless, these vectors are likely still distinguishable in higher dimensions given their significant variance. This suggests that the decoder prioritizes information in the latent vector that regenerates the original sentence.



Limitation

One limitation is the T5 model itself. Though it is the relatively most suitable model, it is not trained on any text style transfer task, which makes it difficult for both prompting and further fine-tuning, especially under the condition of insufficient data.

Another significant limitation is our data. Due to the total number of Shakespeare's play lines and loss of data through preprocessing, our data is insufficient for fine-tuning tasks. Also, lacking parallel data makes the process more challenging, since it makes us not able to directly observe the loss by comparing the generated sentence and the target sentence. CNN news sentences seem more complex than Sheakspear's sentences. We believe this is because CNN news was written by many different authors so the style of the text is more varied compared to Shakespeare's plays which are all written by the same person.

Our fine-tuning method makes the model performance even worse. We believe this is because, during the training process of transfer learning, the model's capacity will be reduced. So it will be worth it for us to find an appreciated training method.

There are also some limitations in the Latent Representation Splitting method. Our Latent Representation Splitting approach for text style transfer faced limitations due to the overlapping content and style representations in latent vectors, as revealed by 2D plots. This overlap indicated that the model failed to effectively separate style from content, with both aspects remaining entangled. Additionally, the training process, dominated by reconstruction loss, prioritized content preservation over style differentiation, leading to redundant information in the latent vectors. This imbalance in training objectives hindered style transfer, highlighting the need for a more nuanced approach that balances content integrity and stylistic variation, an issue we addressed in the Two-encoder part of our project.

Project Extension

Due to the limitation of time and computing resources (We spend an entire night training every fine-tuned model). Here are some extensions we would like to do on this project.
Since the fine-tuning with two separate models seems to have an improvement by not just combining two sentences together. We would like to try this idea on the two encoders method. We can use cross-validation to solve the insufficient data problem.
We also would also like to do some extension work on the two encoders method. First, we can increase the training duration. Currently, the decrease in total loss exceeds the increase in style loss. With extended training, a rise in style loss might eventually contribute to an overall increase in total loss, potentially altering the results.
Second, reevaluate the method for sampling latent vectors. Currently, this is done through a Gaussian distribution and penalized by KL divergence. If the style vector's influence as a variance factor is too subtle to affect the sampling outcome, exploring alternative distributions might be beneficial. Also, we can try the sampling on some other distribution such as the chi-square method, and perform a more sophisticated variable selection method like Ridge regression to tune the decoder.
Third, implement dynamic weighting for each loss function. We currently use hard-coded weights to balance each loss function. Introducing dynamic weighting that increases emphasis on losses as they rise could prove advantageous.

Contribution

Xiaoyu Yang: Primary data processing, Prompt T5, Latent Representation Splitting, Two Encoder Method for Content and Style Extraction, report construction.

Xiang Li: Prompt T5, Fine-Tune + Prompt T5, Two Encoder Method for Content and Style Extraction, report construction and proofreading.

Lingwen Deng: Prompt T5, GPT2, GPT3.5, Fine-Tune+ Prompt T5 in single sentence approach, report construction.

Reference Work

Resource 1: https://text-style-transfer.fastforwardlabs.com/

Resource 2: https://ar5iv.labs.arxiv.org/html/1808.04071

Resource 3: https://arxiv.org/abs/2011.00416

Resource 4: https://arxiv.org/pdf/1910.10683v4.pdf