ECE 1786

Creative Application for Natural Language Processing

Final Report – IELTS Composition Marker

Word Count: 1799

Ziqin Shang 1002976416

Chuntung Chu 1003442893

Introduction

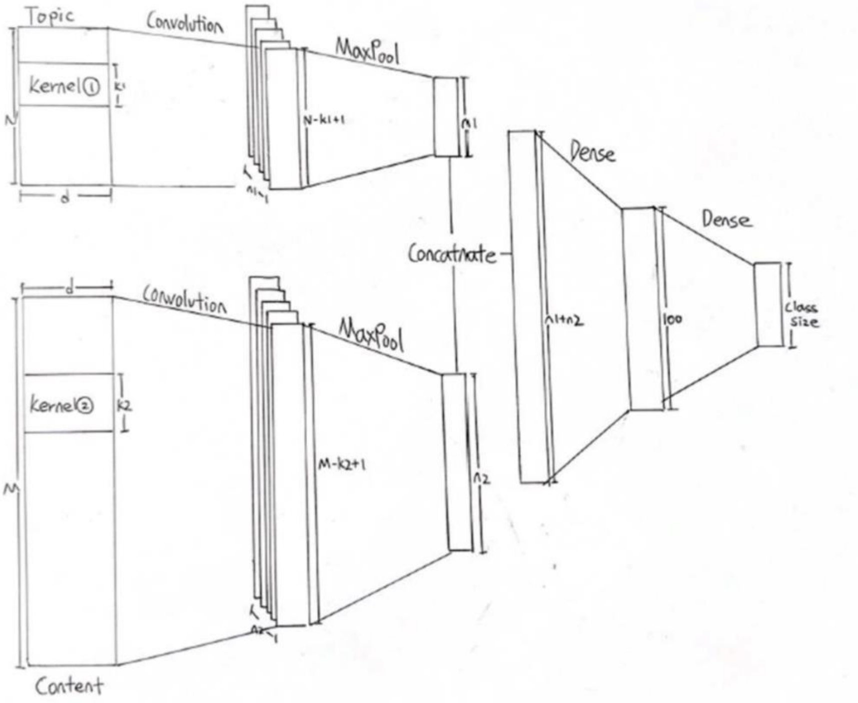
As globalization progresses, we are seeing an increasing volume of people taking the standardized language tests, some widely known tests are IELTS and TOEFL. Learning a new language and getting used to these standardized tests can be a bothersome process, especially when you are on your own. Sometimes the teacher is just not there, and you need to mark your own test, this can be hard if you are new to the language and do not know the metrics of the test.

The structure of these tests is composed of four parts: Listening, Reading, Writing and Speaking. To self-evaluate the performance of the Listening and Reading section is easy, one can just compare the answer with the solution and deduct marks accordingly. For speaking, a human examiner is required to evaluate the performance, so it is quite hard to automate. Considering the scope of this project and the fact that applications of NLP are mainly focused in written texts, we will be attempting to automate the performance evaluation for the writing section.

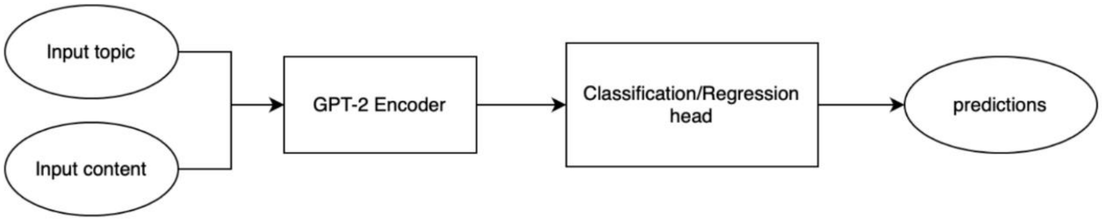
There will be two inputs: the question body of the composition, which is around 25-35 words and the written composition around 250-300 words. We are trying to predict the final discrete score of the composition that is ranging from 1-9.

Illustration

The overall baseline model architecture is illustrated below:



Final model's illustration:



Background & Related Works

This application falls under the Automated Essay Scoring (AES) area of the NLP, this field was invented back in 1960s, where researchers and linguists are working together to define the metrics embedded in an essay, such as essay length, average sentence length etc. and they used multiple linear regression to predict the score of a given essay, one variation of this method is replacing the linear regression with the binary classifier or k-nearest-neighbors to distinguish good or bad essays. (Larkey)

After the invention of the neural networks, this ground-breaking technique is also applied to AES, this removes the need of manually identifying the key features, since deep neural networks such as CNN and LSTM can automatically capture and learn the complex features of essays. It is discovered that CNN are effective for sentence modeling while LSTM are more effective in document modeling. (Dong et al.)

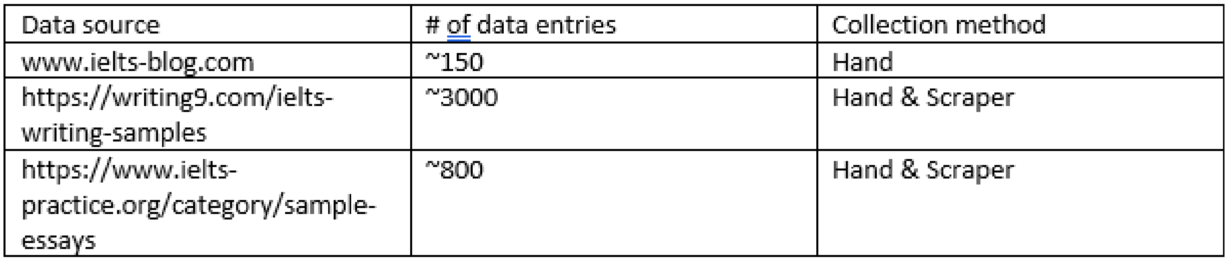
The emergence of transformers also gives AES much more flexibility for processing text sequences, there are various pre-trained models available across the internet and they are all proven excellent in terms of their own usage, a pre-trained model can be fine-tuned on the training dataset to achieve a higher accuracy.

Data and Data Processing

So far, 4000 data entries have been collected, a single data entry is composed of three parts:
Topic: The given composition topic, usually around 25-50 words.
Content: The composition body written based on the topic, around 200-400 words.
Score: A discrete score from 1-9, where all composition below 4 is categorized as "<4", the score set is:

$$S = \{< 4, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8, 8.5, 9\}$$

These data entries were collected either by hand (copy & paste) or by python scraper from different websites, the collection method needs to be readjusted since each website has a different format. The source of data, number of data collected, and collection method is listed below:



Data source	# of data entries	Collection method
www.ielts-blog.com	~150	Hand
https://writing9.com/ielts-writing-samples	~3000	Hand & Scraper
https://www.ielts-practice.org/category/sample-essays	~800	Hand & Scraper

The data has also been cleaned so they are evenly distributed among the 12 classes. This is done by firstly collecting all available data at once, then inspect the score distribution and remove some data from the overpopulated class (7.5). The class distribution of before/after data cleaning is listed below.

	<4	4	4.5	5	5.5	6	6.5	7	7.5	8	8.5	9
Before	457	311	308	320	302	320	394	322	677	321	338	311
After	457	311	308	320	302	320	394	322	297	321	338	310

In addition to this, individual data entries have also been cleaned by replacing all the newline characters (/n, /r) to spaces.

Architecture and Software

The overall architecture of the final model can be described by the illustration above.

The input is composed of the combination of context and the topic, separated by a space token. Then the input is fed into the gpt-2 encoder that outputs a 1024-dimensional representation vector. The encoder is a pretrained gpt-2 medium model from the huggingface. Finally, the representation vector is passed to a classification/regression head.

For the classification head, it's a simple fully connected layer that takes in 1024-dimensional vector and outputs a 12-dimensional vector. This 12-dimensional vector is passed through a softmax layer and cross entropy loss layer.

For the regression head, it's a fully connected layer that outputs one single scaler. Finally, the scaler is passed into a customized piecewise function that bounds the range of the scaler (between 3.5 and 9). Finally, the bounded scaler will be used to calculate an MSE loss with the label (for <4, the label score is 3.5).

During training stage, both encoder and the head are trained. During evaluation stage, the prediction of the regression head is the closet score band to the bounded output scaler. For example, if the bounded scaler is 5.6, the model's prediction will be 5.5.

We also built a frontend for the final model using Gradio. The user will need to input the topic and the content to get a prediction score.

Baseline Model

The baseline model has a CNN architecture, it has two separate CNN that are responsible for topic and content respectively, both CNNs have one convolutional layer for producing the feature map and a Maxpool layer for pooled representation. The output of these two CNNs will be concatenated and fed into a fully connected neural network. The addition of feature map sizes of two CNNs would be the input layer size and the neural network has one hidden layer of size 100 and an output layer representing the possibilities of each class. For the regression baseline model, output layer has a single numerical output indicating the final score of the essay and everything thing else remains unchanged.

To train the classification model, the loss function of Cross Entropy Loss is used, where the true output will be a one-hot encoding of the result class. For the regression model, Mean Squared Error was used to calculate the absolute distance from the prediction score to the target score.

Quantitative Results

We evaluated both baseline model and the final model with classification and regression heads on the validation set. We evaluated with two metrics, the exact accuracy and the average absolute distance. Worth noting that the absolute distance is the distance between the prediction score band and the label score band.

Classification Head	Accuracy	Average Absolute Distance
Baseline	21%	1.5
GPT-2 finetune	41%	0.66

Regression Head	Accuracy	Average Absolute Distance
Baseline	14%	1.55
GPT-2 finetune	32%	0.64

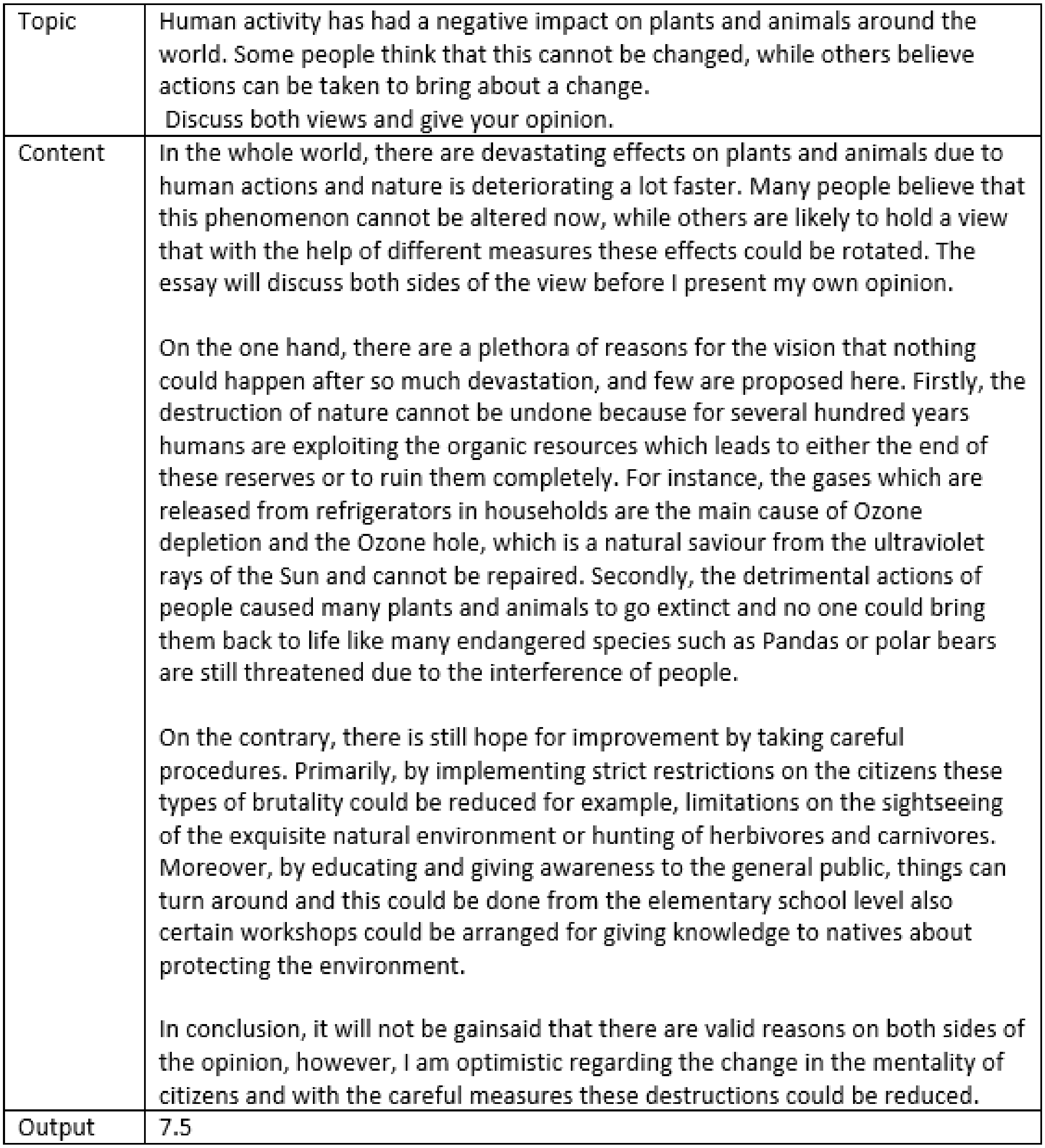
In both classification and regression heads, the gpt-2 finetuned model performs better than the baseline model under both metrics, as expected. For average absolute distance, the regression head for the gpt-2 finetune model performs better than the classification head, but has lower exact accuracy than the classification head. This is also expected, because MSE loss essentially directly optimizes the distance between the output score and the label score, and CE loss directly optimizes the accuracy.

We believe average absolute distance is a better metric to evaluate this task because two essays with the same quality can very easily be given different scores by different graders, and this reflects in our training dataset as well. A prediction will be good enough if it’s within an acceptable range.

Qualitative Results

We evaluated our model on several pieces of writings samples that are outside of our dataset.

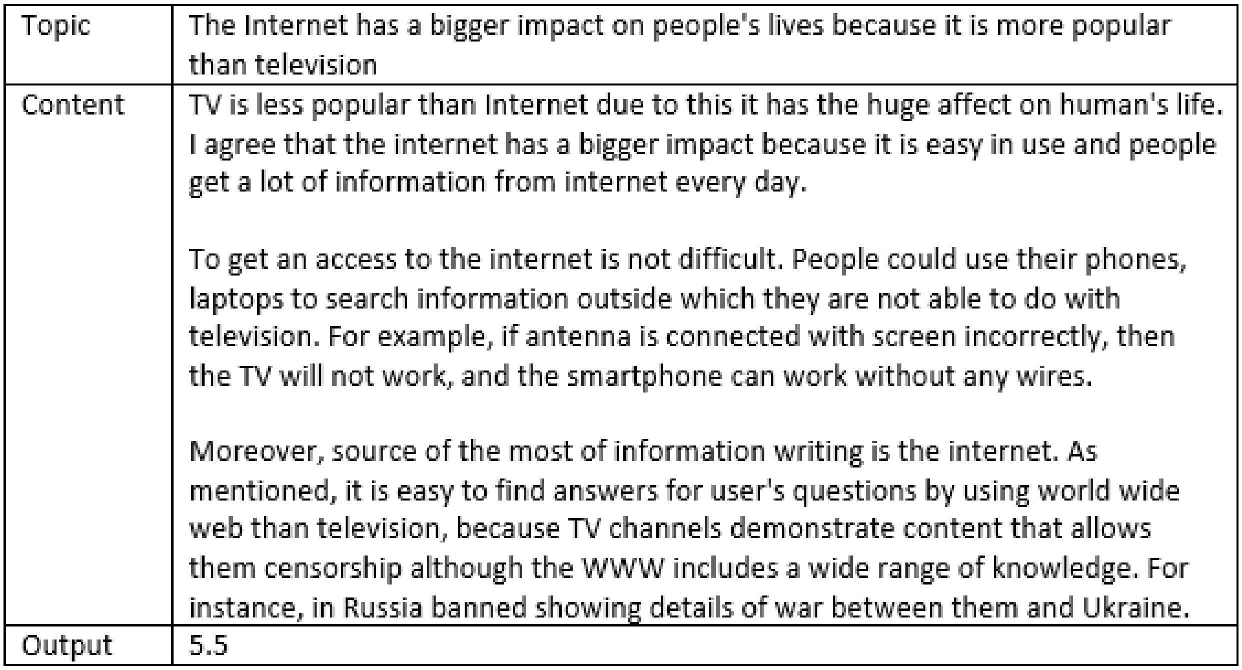
First essay:



Topic	Human activity has had a negative impact on plants and animals around the world. Some people think that this cannot be changed, while others believe actions can be taken to bring about a change. Discuss both views and give your opinion.
Content	In the whole world, there are devastating effects on plants and animals due to human actions and nature is deteriorating a lot faster. Many people believe that this phenomenon cannot be altered now, while others are likely to hold a view that with the help of different measures these effects could be rotated. The essay will discuss both sides of the view before I present my own opinion. On the one hand, there are a plethora of reasons for the vision that nothing could happen after so much devastation, and few are proposed here. Firstly, the destruction of nature cannot be undone because for several hundred years humans are exploiting the organic resources which leads to either the end of these reserves or to ruin them completely. For instance, the gases which are released from refrigerators in households are the main cause of Ozone depletion and the Ozone hole, which is a natural saviour from the ultraviolet rays of the Sun and cannot be repaired. Secondly, the detrimental actions of people caused many plants and animals to go extinct and no one could bring them back to life like many endangered species such as Pandas or polar bears are still threatened due to the interference of people. On the contrary, there is still hope for improvement by taking careful procedures. Primarily, by implementing strict restrictions on the citizens these types of brutality could be reduced for example, limitations on the sightseeing of the exquisite natural environment or hunting of herbivores and carnivores. Moreover, by educating and giving awareness to the general public, things can turn around and this could be done from the elementary school level also certain workshops could be arranged for giving knowledge to natives about protecting the environment. In conclusion, it will not be gainsaid that there are valid reasons on both sides of the opinion, however, I am optimistic regarding the change in the mentality of citizens and with the careful measures these destructions could be reduced.
Output	7.5

This is an essay with good quality. It has little grammatical error and has a lot of discussions and contents. As a result, our predictor gives a score of 7.5, which is a very high score for the composition test.

Second essay:

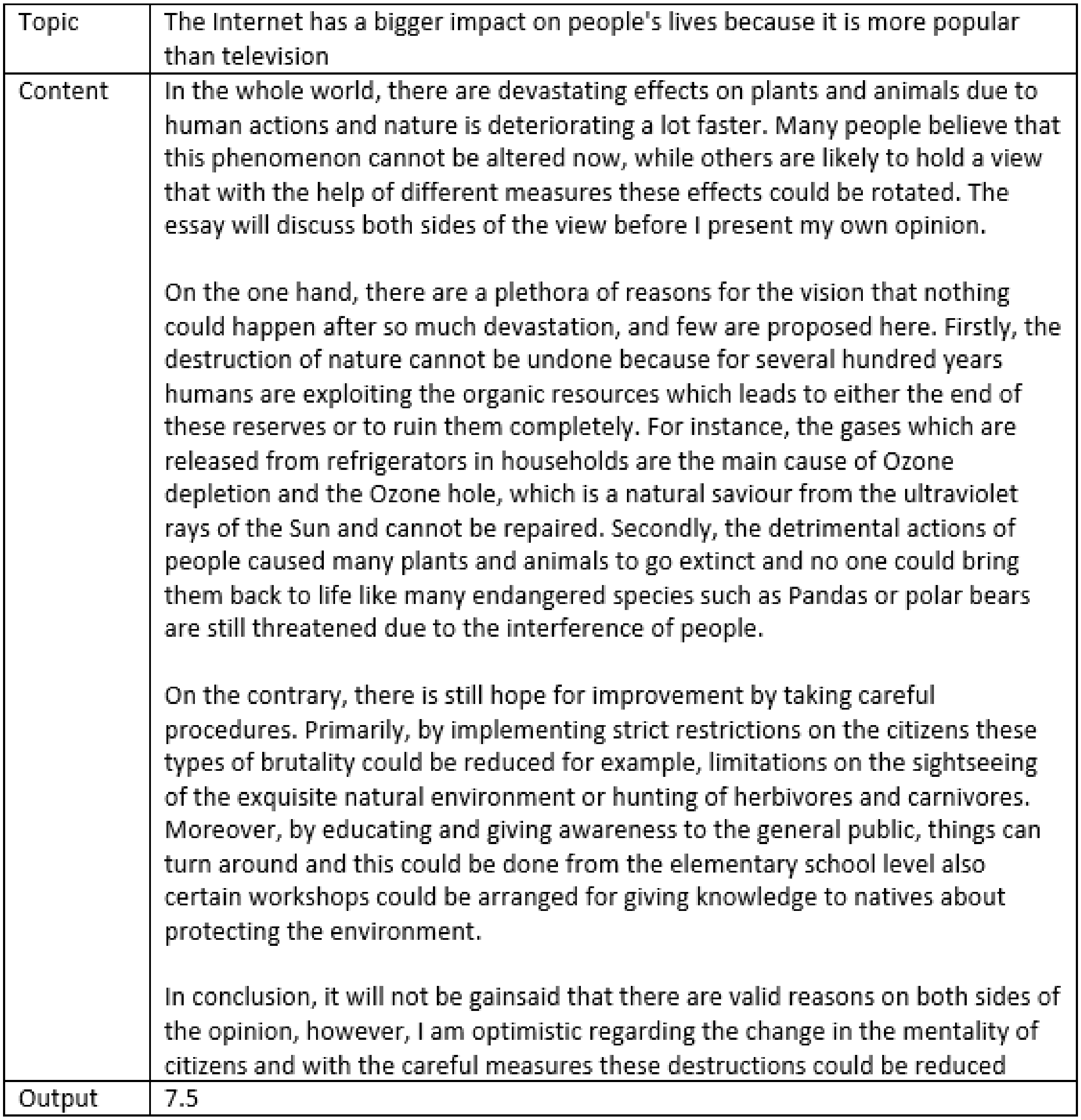


Topic	The Internet has a bigger impact on people's lives because it is more popular than television
Content	TV is less popular than Internet due to this it has the huge affect on human's life. I agree that the internet has a bigger impact because it is easy in use and people get a lot of information from internet every day. To get an access to the internet is not difficult. People could use their phones, laptops to search information outside which they are not able to do with television. For example, if antenna is connected with screen incorrectly, then the TV will not work, and the smartphone can work without any wires. Moreover, source of the most of information writing is the internet. As mentioned, it is easy to find answers for user's questions by using world wide web than television, because TV channels demonstrate content that allows them censorship although the WWW includes a wide range of knowledge. For instance, in Russia banned showing details of war between them and Ukraine.
Output	5.5

This is an essay that's poorly written compared with the first one. It has much more grammatical errors even in the first paragraph, and it does not discuss the topic very well. The prediction given is 5.5, which is a lackluster score (Most university requires 6.5 for undergraduate admission).

Inputs that don't work as well as expectation:

As we experimented with more inputs, an unexpected problem has occurred. The model seems to not capture the relationship between topic and content well. The model seems to rely heavily on the quality of writing in the content and neglects whether the content discusses well around the topic. An example is shown below. The content is the same as the first essay but the topic is randomly swapped with a different one.



Topic	The Internet has a bigger impact on people's lives because it is more popular than television
Content	In the whole world, there are devastating effects on plants and animals due to human actions and nature is deteriorating a lot faster. Many people believe that this phenomenon cannot be altered now, while others are likely to hold a view that with the help of different measures these effects could be rotated. The essay will discuss both sides of the view before I present my own opinion. On the one hand, there are a plethora of reasons for the vision that nothing could happen after so much devastation, and few are proposed here. Firstly, the destruction of nature cannot be undone because for several hundred years humans are exploiting the organic resources which leads to either the end of these reserves or to ruin them completely. For instance, the gases which are released from refrigerators in households are the main cause of Ozone depletion and the Ozone hole, which is a natural saviour from the ultraviolet rays of the Sun and cannot be repaired. Secondly, the detrimental actions of people caused many plants and animals to go extinct and no one could bring them back to life like many endangered species such as Pandas or polar bears are still threatened due to the interference of people. On the contrary, there is still hope for improvement by taking careful procedures. Primarily, by implementing strict restrictions on the citizens these types of brutality could be reduced for example, limitations on the sightseeing of the exquisite natural environment or hunting of herbivores and carnivores. Moreover, by educating and giving awareness to the general public, things can turn around and this could be done from the elementary school level also certain workshops could be arranged for giving knowledge to natives about protecting the environment. In conclusion, it will not be gainsaid that there are valid reasons on both sides of the opinion, however, I am optimistic regarding the change in the mentality of citizens and with the careful measures these destructions could be reduced
Output	7.5

The predictor still gives 7.5 as the score, even though the content discusses a different topic.

Discussion and Learnings

The overall performance for both models falls into expectation, the baseline model provided a rough guess for the result while the refined gpt-2 model presented a more accurate outcome, the metrics used (accuracy and absolute distance) is not very fair for this application because writing task is very subjective so as the marking process, the exact same essay might get graded into different scores, but they should be in the same range. To improve this, we can make the categorization coarser, for example, the score of 5, 5.5 and 6 can be categorized into “5-6”, this way the model can provide a higher accuracy while keeping the evaluation consistent.

To make the project better, we can support the model with more training data, 4000 data entries is plenty but it can still be improved with a larger amount of data, we can also improve the accuracy by using a larger language model such as GPT-3, this will help the model to adapt to various of topics.

To resolve the issue described in the qualitative analysis, we can create examples in the dataset that have mismatched topic and content and assign the example a low score. We think finetuning using GPT-3 would also result in better performance in these examples because GPT-3 has the ability to do zero-shot learning and is able to capture relationship between sentences very well.

Individual Contribution

Ziqin:

- Proposed the project idea
- Hand collected and labelled around 100 data samples and used scraper to collect around 1000 data samples from different websites
- Constructed and trained baseline model.

Chuntung:

- Collected around 2000 data samples using scrapper and manual work.
- Constructed and trained the gpt-2 finetuned model.
- Proposed to experiment with regression head.
- Gradio implementation

Reference

Dong, Fei, et al. Attention-based Recurrent Convolutional Neural Network for Automatic Essay Scoring. 2017. Attention-based Recurrent Convolutional Neural Network for Automatic Essay Scoring, https://aclanthology.org/K17-1017.pdf.

Larkey, Leah S. Automatic Essay Grading Using Text Categorization Techniques. 1998. 290941.290965, https://dl.acm.org/doi/pdf/10.1145/290941.290965.