ECE1786 - CREATIVE APPLICATIONS OF NATURAL LANGUAGE PROCESSING

FINAL REPORT

Project Title: Newsify

Gopi Revathi Sreenivasan - 1008141388

Aswin Raj Giri - 1007950059

Word Count: 1830

Penalty: 0%

Introduction:

News provides information and knowledge. The headline of the article is an important aspect since it drives users towards reading the article. A strong headline conveys what the users are looking which might be valuable to the users. That is where Newsify comes in where we try to generate fake headlines and news articles with keywords as inputs or prompts from the user. Latest advancements in NLP models have been one of the main motivations for this project and here we try to explore the capabilities of BERT model and GPT2 model in the space of headline and news articles generation.

With latest advancement and capabilities of the transformer models and the possibility of using a pretrained model such as GPT 2 medium, which are used in similar projects solidifies our confidence in utilizing the machine learning based approach.

Newsify Overview – Figure:

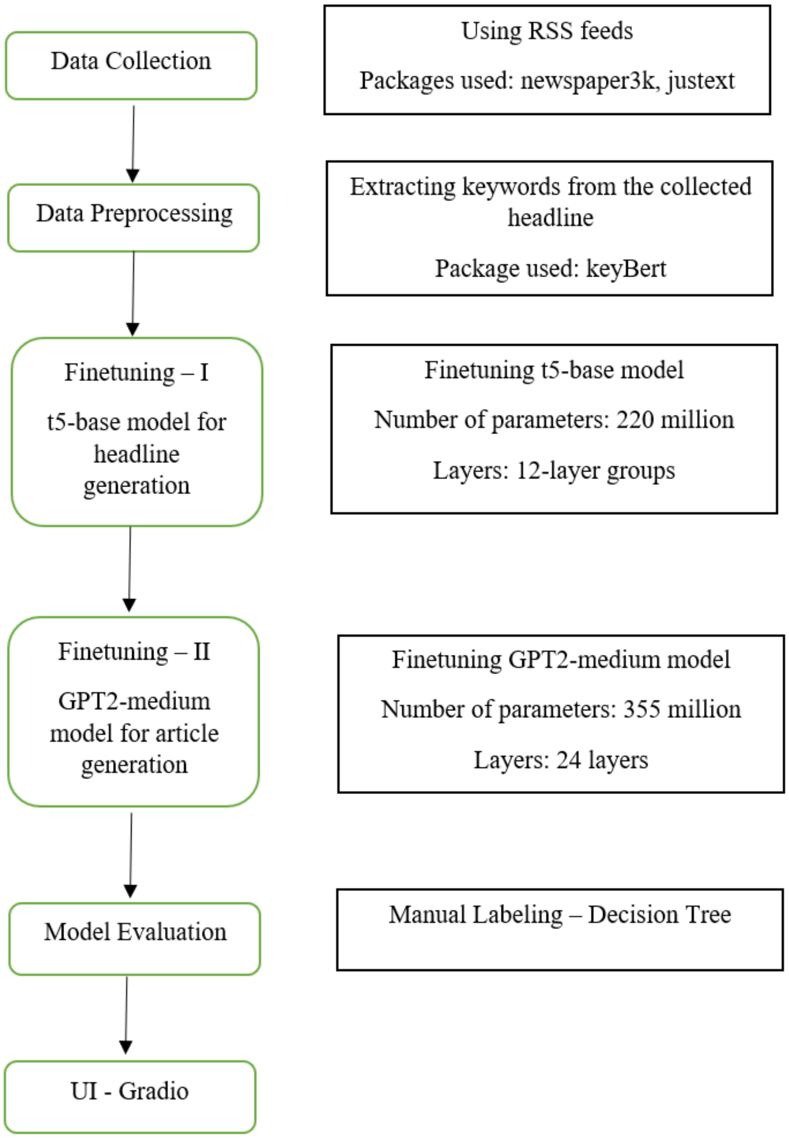

Data Collection
Using RSS feeds
Packages used: newspaper3k, justext
Data Preprocessing
Extracting keywords from the collected headline
Package used: keyBert
Finetuning – I
t5-base model for headline generation
Finetuning t5-base model
Number of parameters: 220 million
Layers: 12-layer groups
Finetuning – II
GPT2-medium model for article generation
Finetuning GPT2-medium model
Number of parameters: 355 million
Layers: 24 layers
Model Evaluation
Manual Labeling – Decision Tree
UI - Gradio

Figure 1. Overview of how Newsify works

Background & Related Work:

Text generation based on the given keywords or sentences is a well researched topic. Our main focus was to finetune the model to generate headlines rather than normal sentences.

Our project's model was an improvement towards the LSTM architecture used in "Myanmar News Headline Generation with Sequence-to-Sequence model" by Yamin Thu et. al [1]. This was one of the first papers to generate a headline instead of a normal sentence and they used an LSTM encoder decoder model. The second part of our project which is text generation was inspired from open ai GPT3 model and "Template Controllable keywords-to-text Generation" by Abhijit Mishra et.al [2] where keywords were used for text generation. The main insight this works provides is how the approach is indifferent to the order of the input keywords.

Data Collection and Preprocessing:

For data collection, RSS feeds of various news articles for categories such as sports, business, and technology are fetched which are in XML format. From the XML document, the website links of news articles are collected, from which the nonresponsive and broken links are removed. Then using the python packages, newspaper3k and justext, the headline and the article corresponding to the headline are extracted respectively and stored in a data frame. Finally, after removing all the null and duplicate values, the data frame is saved as an csv file.

For data preprocessing, first we generate keywords using the collected headlines. For this keyBERT model, is used which takes in the headline as the input and generates keywords as output. These extracted keywords are appended with the corresponding headline and news article and saved as an csv file which is used for training/fine tuning the models.

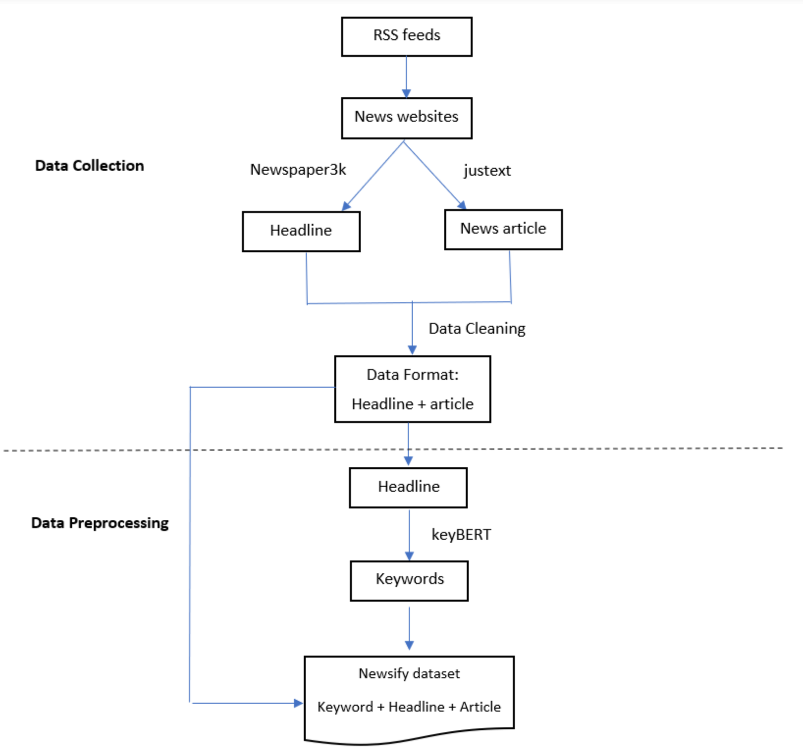


Figure 2. Data Collection and Data Preprocessing

Architecture and Software:

The architecture of the project is composed of two subdivisions:

1. Headline Generation
2. Article Generation

Headline Generation:

The hugging face language model used for headline generation is 'mrm8488/t5-base-finetuned-common_gen' (reference). This is Google's T5-Base model, which takes a sequence of text as input and outputs a sequence of text (seq-to-seq model). This version of T5 has 220 million parameters and 12 layers of encoders and decoders. This model is utilized instead of Google's BERT because BERT-style models can only output either a class label or a span of the input. It is pretrained using 'Common Gen' Dataset, which is used for generative common-sense reasoning. The dataset comprises of keywords as inputs and sentences formed using those keywords as outputs. The Language Modelling head is leveraged since it is a generation task. The pretrained model is finetuned using the collected dataset. The keywords are tokenized and used as encoder inputs along with the attention mask. Similarly, the headlines are tokenized and used as decoder inputs. Experimented with various epoch counts and batch sizes to arrive at the best model. The best model is chosen based on quality of generation. The model is finetuned so that it produces headlines as outputs rather than simple statements/sentences. The model is trained on a GPU having 16GB ram.

Hyperparameter setting used:

Batch Size: 2
Number of epochs: 15
Learning rate: 5 x 10-5

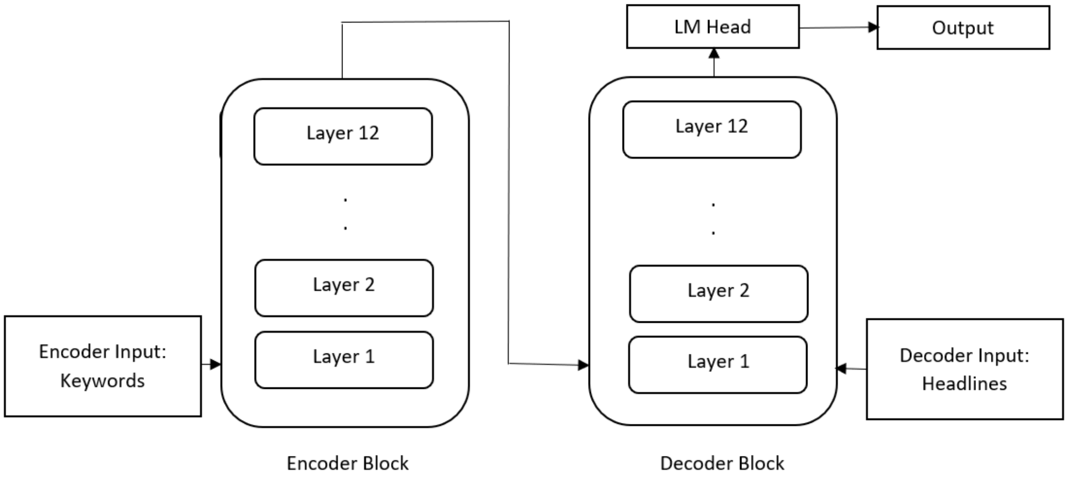


Figure 3. Architecture diagram for headline generation

Article Generation:

'GPT2-Medium', a transformer-based language model created by Open AI is employed for article generation. This model is trained to guess the next word in sentences. When a sequence of continuous text is given as input, the model gives the same sequence along with a newly predicted word as output. This version of GPT2 has 355 million parameters and 24 layers. It is pretrained using 'Web Text' Dataset (causal language modelling), containing 40GB of text which has not been publicly released. A language modelling head is used to generate the article from headline. The finetuning process for this model differs from that of the T5 model. As it is a decoder-based model, it takes only one input. To finetune this model, the headline and the article from the dataset should be combined and given as input. This is achieved by using a special separator token. The headlines and articles are tokenized and fed into the model. Hyperparameters such as batch size and number of epochs are altered to obtain the best model. The length of the sequence produced can be adjusted. The model is trained on 3 distributed GPUs each having 16GB ram.

Hyperparameter setting used:

Batch Size: 2
Number of epochs: 15
Learning rate: 5 x 10-5

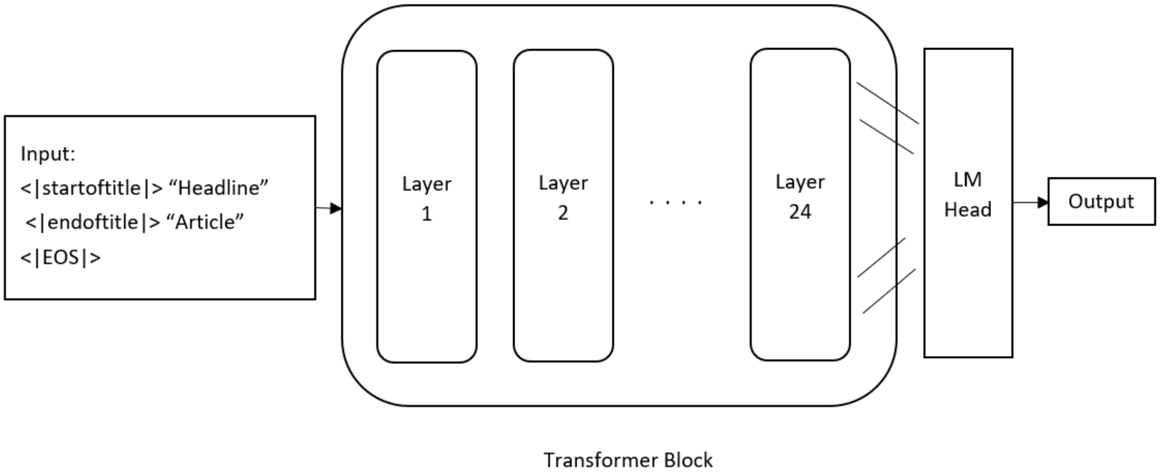
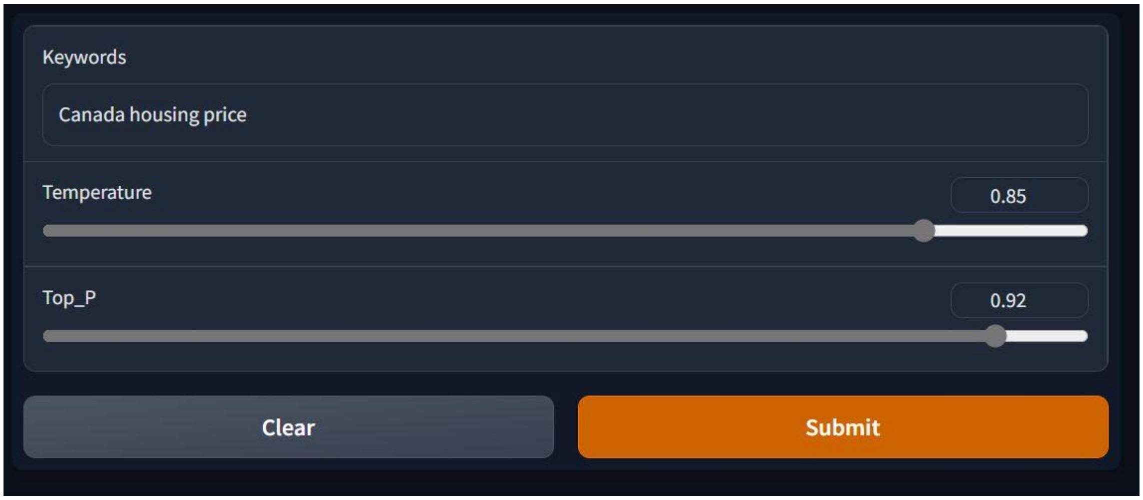


Figure 4. Architecture diagram for article generation

User Interface - Gradio:

Gradio is an intuitive web interface for any machine learning models. This makes it easier for anyone to use someone else's model. The created interface takes keywords as inputs. There are two sliders provided, which can be used to adjust the hyperparameters such as Temperature and Top P. This enables the user to get better results by experimenting with different hyperparameter settings. Once the input is given, it takes about 4 to 5 seconds to generate the headline and an article.

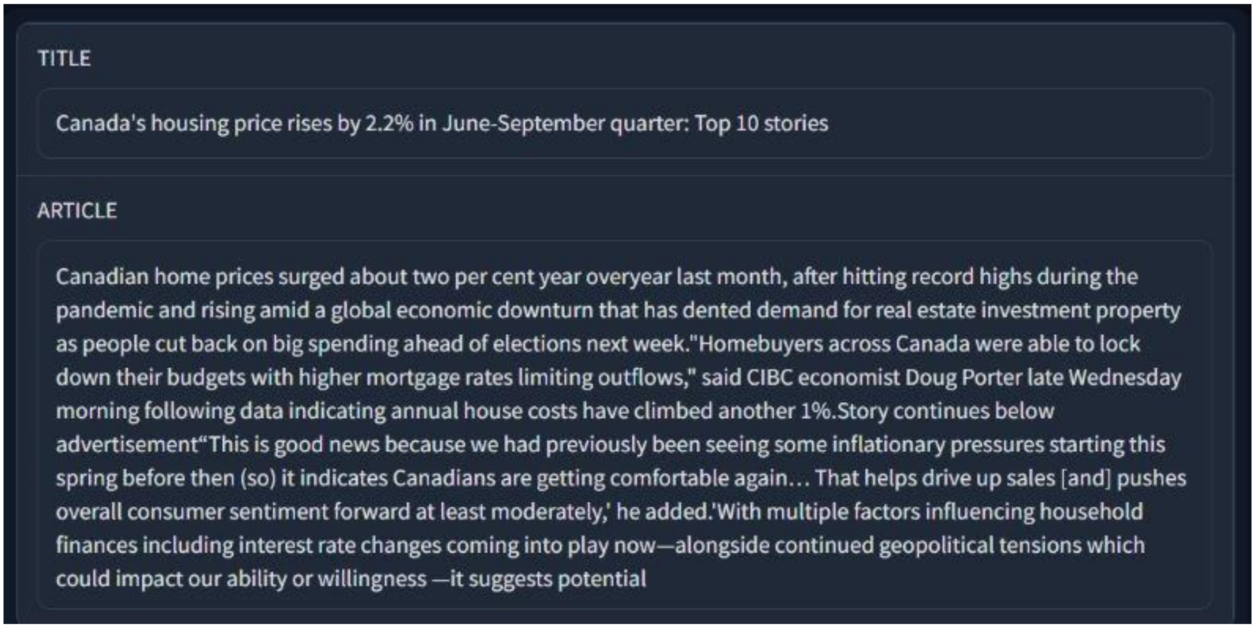

Keywords

Canada housing price

Temperature 0.85

Top_P 0.92

Figure 5. Gradio Web Interface – Input


TITLE

Canada's housing price rises by 2.2% in June-September quarter: Top 10 stories

ARTICLE

Canadian home prices surged about two per cent year overyear last month, after hitting record highs during the pandemic and rising amid a global economic downturn that has dented demand for real estate investment property as people cut back on big spending ahead of elections next week."Homebuyers across Canada were able to lock down their budgets with higher mortgage rates limiting outflows," said CIBC economist Doug Porter late Wednesday morning following data indicating annual house costs have climbed another 1%.Story continues below advertisement“This is good news because we had previously been seeing some inflationary pressures starting this spring before then (so) it indicates Canadians are getting comfortable again… That helps drive up sales [and] pushes overall consumer sentiment forward at least moderately,’ he added.’With multiple factors influencing household finances including interest rate changes coming into play now—alongside continued geopolitical tensions which could impact our ability or willingness —it suggests potential

Figure 6. Gradio Web Interface – Output

Quantitative results:

To get quantitative results, manual labelling approach is used. The test dataset consists of 119 keywords, headlines, and articles. Using the keywords as input, headlines and articles are generated by the T5-Base and GPT2 model respectively. A decision tree classifier is created to manually label the outputs. First, we look at the ‘contextual correctness’, checking if the context provided by the headline is relevant to the context present in the article. Then the outputs are classified on the grounds of grammatical structure. The branching depends on whether the grammatical structure is Excellent/Good/Poor/Very Poor. After this we check if words are repeated very often and based on this, we fix ranks for the outputs. If the rank is 3 or above, it is considered as a good result. If the rank is less than 3, it is considered as a poor result. We got 57 good results out of 119 records in the test dataset, which gives us an accuracy of 47.8%.

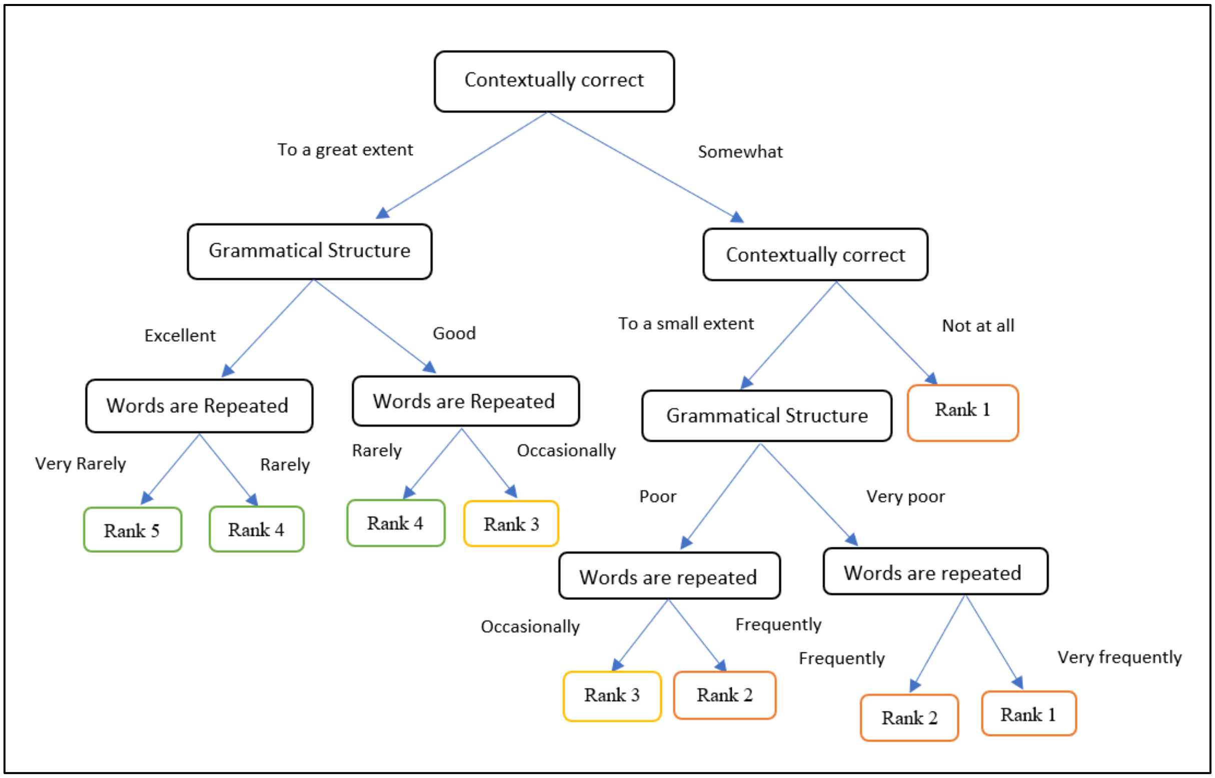


Figure 7. Decision tree for manual labelling

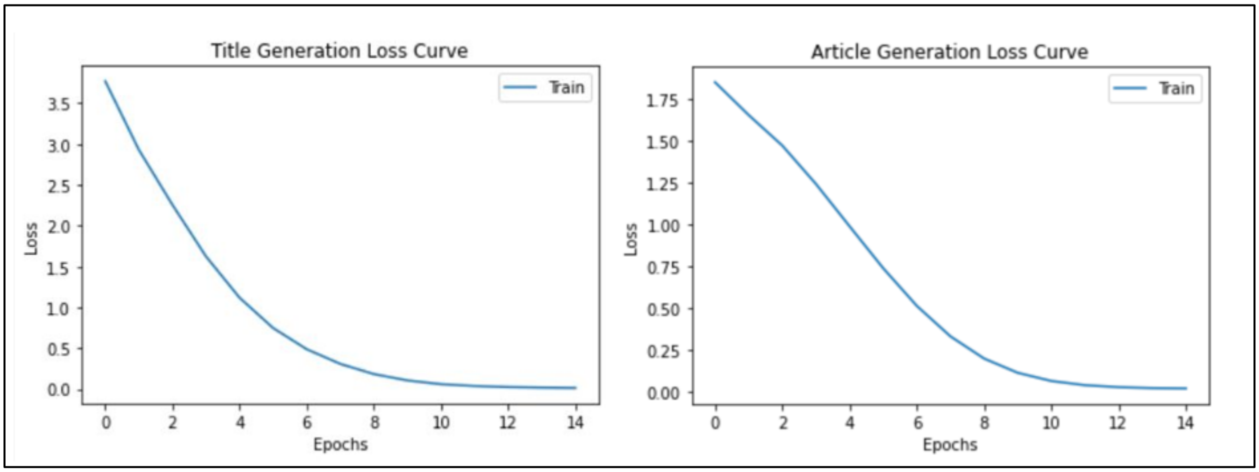
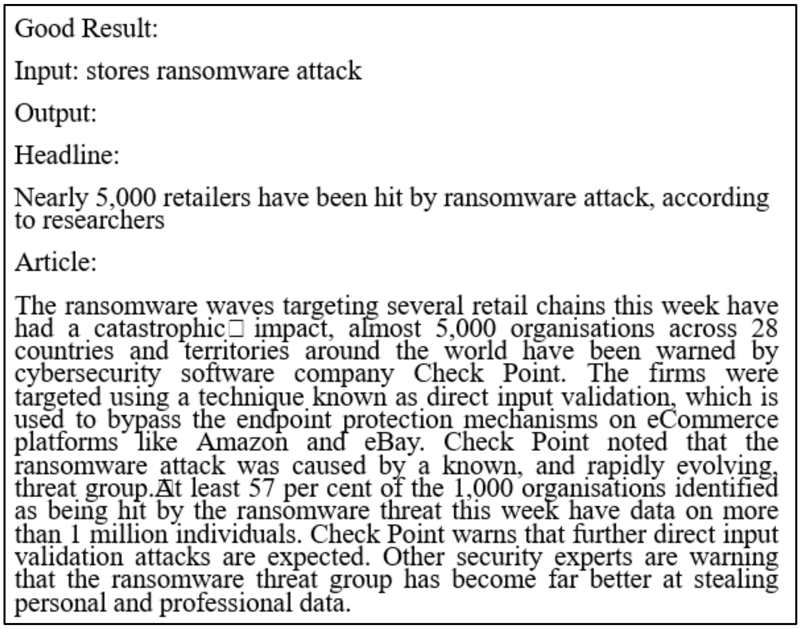


Figure 8. Loss curve for title and article generation

Qualitative Results:

After manual labelling of results on the test dataset, below are examples of a good and bad generation of headline and the article.



Good Result:

Input: stores ransomware attack

Output:

Headline:

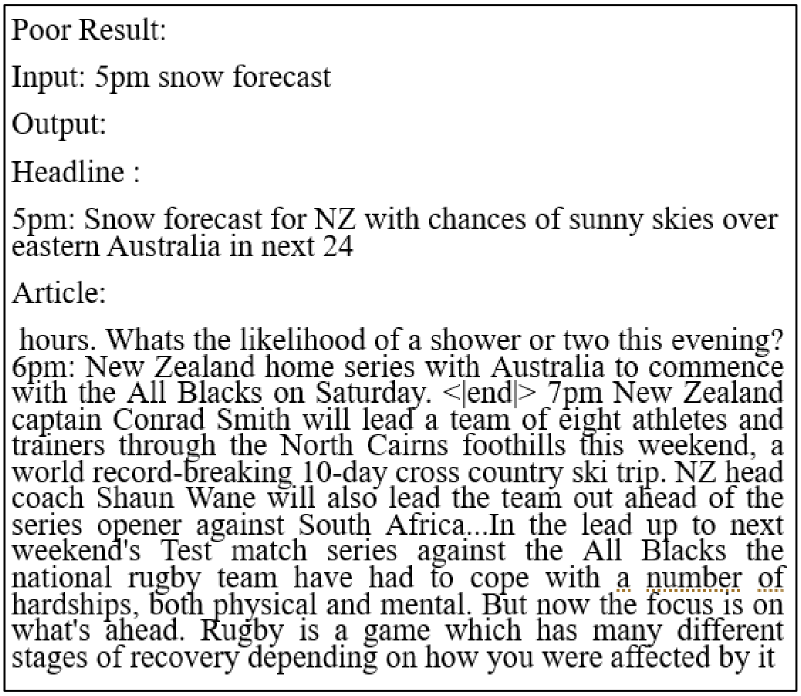
Nearly 5,000 retailers have been hit by ransomware attack, according to researchers

Article:

The ransomware waves targeting several retail chains this week have had a catastrophic□ impact, almost 5,000 organisations across 28 countries and territories around the world have been warned by cybersecurity software company Check Point. The firms were targeted using a technique known as direct input validation, which is used to bypass the endpoint protection mechanisms on eCommerce platforms like Amazon and eBay. Check Point noted that the ransomware attack was caused by a known, and rapidly evolving, threat group.□At least 57 per cent of the 1,000 organisations identified as being hit by the ransomware threat this week have data on more than 1 million individuals. Check Point warns that further direct input validation attacks are expected. Other security experts are warning that the ransomware threat group has become far better at stealing personal and professional data.

Figure 9. Example of a good result

The above example was labelled as Rank 5 based on the decision tree for manual labelling. The attributes that were used for evaluation were contextual correctness, grammatical structure, and repetition of words. In the above example, the generated fake headline and article perfectly makes sense, they are grammatically correct and are well structured around the given keywords. Also, the article has a clear and concise introduction and conclusion by properly explaining the headline.



Poor Result:

Input: 5pm snow forecast

Output:

Headline :

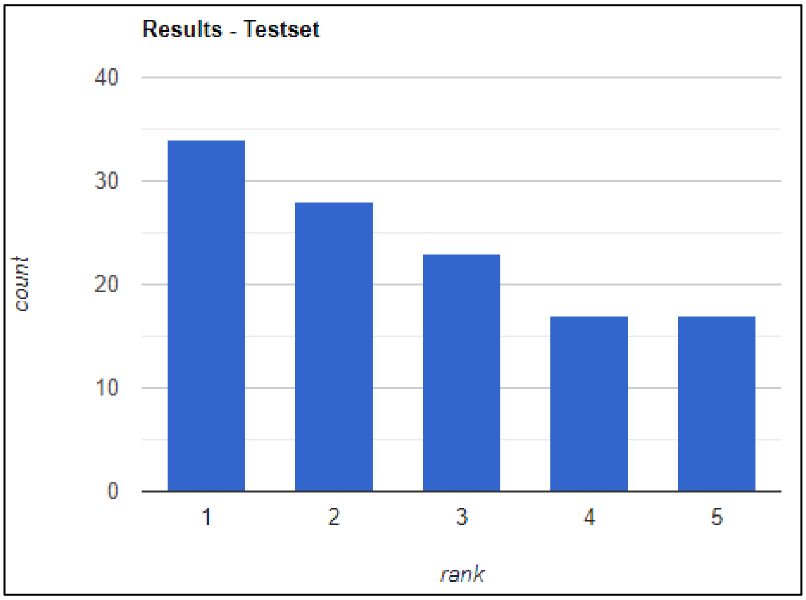
5pm: Snow forecast for NZ with chances of sunny skies over eastern Australia in next 24

Article:

hours. Whats the likelihood of a shower or two this evening? 6pm: New Zealand home series with Australia to commence with the All Blacks on Saturday. <|end|> 7pm New Zealand captain Conrad Smith will lead a team of eight athletes and trainers through the North Cairns foothills this weekend, a world record-breaking 10-day cross country ski trip. NZ head coach Shaun Wane will also lead the team out ahead of the series opener against South Africa...In the lead up to next weekend's Test match series against the All Blacks the national rugby team have had to cope with a number of hardships, both physical and mental. But now the focus is on what's ahead. Rugby is a game which has many different stages of recovery depending on how you were affected by it

Figure 10. Example of a poor result

The above example was labelled as Rank 1 based on the decision tree used for manual labelling. Here, the generated headline does not make proper sense and it is contextually very poor. The generated fake article talks about a topic that is completely irrelevant to the generated headline or the input keywords. It is also grammatically incorrect and has special characters in between which does not make any sense.



Rank 5 – Best Result

Rank 1 – Poor Result

Figure 11. Bar graph for the test set results

Discussion and Learnings:

From our results, we conclude that both the models perform reasonably. But there's room for improvement in several areas. We noticed that when names are given as keywords (for example: Cristiano voted Henry), the generated output was not good. The reason for this behaviour is unknown. The quality of generation can also be improved by expanding the scope of data (we have limited our scope to fields such as technology, business, and sports).

Robust models such as T5-Large (770 million parameters) for headline generation and GPT2-XL (1.5 billion parameters)/ GPT3 for article generation can be used to see if it yields an increase in accuracy. The GPT2-medium can take a maximum of 1024 tokens as inputs, because of this constraint many articles were truncated. This might have affected the performance. The use of single model (GPT2-XL/ GPT3) for both the headline and article generation can also be experimented with.

Manual labelling approach to assess the generated output can be biased and prone to human error. This approach heavily depends on the perspective of the person labelling the outputs. This can impact the outcome. A classifier can be trained on the manually labelled data and used for evaluating the result. This reduces the risk of bias and human error.

Individual Contribution:

Gopi Revathi Sreenivasan:

- Collected 261 RSS Feeds used for extracting news websites
- Responsible for training the 'mrm8488/t5-base-finetuned-common_gen' model used for headline generation
- Responsible for training the 'GPT2-Medium' model used for article generation
- Responsible for manually labelling 19 results
- Responsible for final report

Aswin Raj Giri:

- Responsible for Data Preprocessing and keyword extraction
- Responsible for constructing the decision tree and manual labelling
- Responsible for Gradio implementation.
- Responsible for final report

References:

1. Y. Thu and W. P. Pa, "Myanmar News Headline Generation with Sequence-to-Sequence model," IEEE Xplore, Nov. 01, 2020. https://ieeexplore.ieee.org/document/9295017 (accessed Dec. 13, 2022).
2. A. Mishra, M. F. M. Chowdhury, S. Manohar, D. Gutfreund, and K. Sankaranarayanan, "Template Controllable keywords-to-text Generation," arXiv:2011.03722 [cs], Nov. 2020, Accessed: Nov 15, 2022. [Online]. Available: https://arxiv.org/abs/2011.03722.
3. M. Grootendorst, "KeyBERT," GitHub, Oct. 28, 2022. https://github.com/MaartenGr/KeyBERT (accessed Nov 15,2022).
4. "Newspaper3k: Article scraping & curation — newspaper 0.0.2 documentation," newspaper.readthedocs.io. https://newspaper.readthedocs.io/en/latest/# (accessed Nov. 15, 2022)
5. J. Pomikálek, "jusText: Heuristic based boilerplate removal tool," PyPI. https://pypi.org/project/jusText/ (accessed Nov. 15, 2022).
6. C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," Jul. 2020. [Online]. Available: https://arxiv.org/pdf/1910.10683.pdf

Permissions:

Video Upload

Aswin Raj Giri: Yes

Gopi Revathi Sreenivasan: Yes

Final Report Public Availability

Aswin Raj Giri: Yes

Gopi Revathi Sreenivasan: Yes

Code Public Availability

Aswin Raj Giri: Yes

Gopi Revathi Sreenivasan: Yes