National College of Ireland

Transformer based model for News Headline Generation task by incorporating Named Entity Recognition

MSc Research Project
Data Analytics

Steffi Veientlena
Student ID: x21202109

School of Computing
National College of Ireland

Supervisor: Prof. Prashanth Nayak

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Steffi Veientlena
Student ID:	x21202109
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Prof. Prashanth Nayak
Submission Due Date:	14/08/2023
Project Title:	Transformer based model for News Headline Generation task by incorporating Named Entity Recognition
Word Count:	7113
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Steffi Veientlena
Date:	18th September 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Transformer based model for News Headline Generation task by incorporating Named Entity Recognition

Steffi Veientlena
x21202109

Abstract

Business these days are looking for insights from large amount of data from various sources for strategic growth. Textual information, specially news data has seen notable upsurge in the past years. This necessities the effective method to manage and comprehend the abundance of textual content. In order to meet this challenge, Extraction and Abstractive text summarization techniques can be used which allows more streamlined information management and consumption. This research study suggests the novel use of abstractive summarization on news data with a purpose of producing news headline that are appropriate to the context. while the earlier studies focused on attention based models for abstractive text summarization. This study goes further by examining the integration of named entity recognition (NER) with transformer based T5 model, renowned for its efficacy in language understanding and text generation task.The ROUGE and Bert score evaluation metric commonly used evaluate the quality of generated text in comparison with the actual text has been used in this study to examine the robustness of the proposed model.

1 Introduction

Text summarization is one of the NLP techniques that uses various machine learning algorithms to find and extract important information from large corpus of text and condenses them to concise form. Various strategies have been implied to ensure that the summary generated accurately captures the coherence and context of the original text. The goal of this techniques is to capture most important and relevant information from the original text and generate text that is clear and insightful. One of the most common use of text summarization is headline generation. There are two approaches through which text can be summarized mainly known as extractive summarization and abstractive summarization Widyassari et al. (2022). Extractive summarization focuses on extracting specific content or sentence from the original text and display them in generated text result. The typical problem with this kind of summarization is positioning of the sentences in summarized text.where as abstractive text summarization mainly focuses on generating meaningful summaries that capture the semantics of the original summary.

The goal of this study is to create news headline from news articles that accurately convey the content and semantic. This is done by abstractive summarization technique where

the output is created to present crucial information of the text. This study will explore an hybrid architecture that utilises the power of named entity recognition and transformer based model. The initial goal was to assess the function of named entity recognition in this hybrid architecture and see if it outperforms the attention based model. Abstractive text summarization with attention mechanisms have produced encouraging outcomes across diverse application including headline generation taskRehman et al. (2022).

As mentioned,the study suggests hybrid architecture combining a NER model with transformer based model T5.The NER model is used to locate and identify entities such as name, place, organization , locations in the text sentences. These sentences will be identified using NER model from SpaCy package in python. Additionally the study incorporates the T5 (Text to Text Transfer) model which is a text to text framework. T5 model utilizes attention mechanism by which it focuses on long range dependencies and semantic relationship for natural text.Due to which this model has been successfully applied on various NLP task such as text summarization, question-answering and sentiment analysis Wang et al. (2023).The study conducts three experiments in order to understand the effectiveness of the proposed approach. The first experiment will be to directly use the pre-trained model on the test dataset and analyse the result using the evaluation metric. The second will be fine-tune the T-5 model with training set and test the model with testing set and evaluate the result. The third will be to implement the NER and then use the trained model from second experiment and analyse the results.The current research will be heavily drawn from previous studies where abstractive text summarization has worked well on textual data and NER models demonstrated crucial role in entity extraction.By combining both these models the aim is to leverage the strength of both models with the aim of significant improvement in headline generation task.The models have have yielded tremendous outcomes in the past. however, there is notable gap in research on creating news headlines by taking sentences that contain entities and running them through an transformer-based model to measure effectiveness. Further the results of this will be evaluated using BERT score evaluation metric and a comparison with ROUGE tool kit will be evaluated.Thus this research aims at focusing on the research question:

To what extent does incorporating named entity recognition model improve performance of an transformer based model for news headline generation task" The main focus of this project is on the following areas:

- Analysing previous research and results to support or validate the current research.
- Pre-process the data as per the model requirements.
- Identify the NER sentences from the news articles.
- Fine-tune the T-5 model for headline generation task.
- Generating headline from the hybrid model.
- Evaluating the result using Evaluation Metric

The following section is an outline of research flow.A survey of related work in the area of Natural Language Processing is presented in Section 2, with an emphasis on news reporting and text summarization in particular. The primary implementations that were created throughout the research stages are validated by this review. The research

methodology is then covered in detail in Section 3, and the design requirements for the methodologies used in the study are covered in Section 4. The research's application is examined in Section 5, and its evaluation and justification are provided in Section 6. Section 7 of the study presents the research's conclusion as well as any potential follow-up studies.

2 Related Work

2.1 Named Entity Recognition and Text Summarization

The most important step in text summarization of large amounts of material is finding out the key passages and ensure that the summary retains accuracy in context and consistency. By using a named entity recognition paradigm, this can be accomplished in an efficient manner.The NER model is essential for finding entities in texts and classifying them according to entity type. According to a study conducted by Jabeen et al. (2013), it is vital to use Named Entities in social media texts from Twitter to produce summaries that accurately point out key passages in the text. For the purpose of this study, disambiguate named entities are found using an AIDA Service that is open to the public. The topic sentence is created using the recognised entities as part of the summary creation process, and sentences within the text are ranked according to how similar they are. The most prominent sentences are chosen with the least amount of repetition to guarantee there is no repeat sentences in the summary. As a result, the final summary only includes the most pertinent and unique sentences.

Mena and Palazzo (2012)also conducted a study that emphasises the significance of recognising and extracting key entities, particularly in brief context communications like SMS. YAGO KB, a knowledge base with a wealth of information, is used in this research to aid named entity extraction and disambiguation approaches. Clusters of related entities are created when the extracted from the text and matched with the knowledge base. The summary of the text is extracted based on the cluster that has the smallest size, which makes it easier to find the most pertinent information.

In a study by Berezin and Batura (2022), the integration of Named Entity Recognition (NER) with abstractive text summarization is the main objective. This was accomplished by training the Named Entity Recognition model using a domain-specific dataset and the RoBERTa language model. Then, a sequence-to-sequence Transformer architecture known as the BART model was used. The pre-training of BART on a corpus of documents was followed by fine-tuning for the particular purpose of summarization. The model's performance was evaluated using the ROUGE tool set, and the results revealed that the ROUGE score was slightly shy of the most recent state-of-the-art findings.

The studies mentioned above provide valuable insight into named entity recognition models' use in text summarization. They demonstrate how this approach can effectively extract text summaries and handle a number of text classification, text summarization, and text categorization challenges..

2.2 Need for named entity recognition on News data

News articles benefit greatly from Named Entity Recognition (NER), which makes it easier to extract important facts and insights from large amounts of text data. The news industry can increase the effectiveness of tasks like data analysis, information retrieval, and knowledge management by automatically recognising named things.

Named entities were used to summarise the content of Czech news items in a study by Marek et al. (2021). SpaCy's NER Model was the NER model applied to Czech publications. The identified entities were transformed into vectors, and the summary was produced using a Sequence-to-Sequence model with global attention. The news article text and vectors produced from named entities made up the encoder's input. The network could generate the target text while selectively focusing on particular sections of the source text thanks to the global attention technique.

Similar research was done on CNN's Daily Email dataset by Alshibly et al. (2023). Named entities from the news articles were identified using SpaCy's NER Model. Tokenizing the detected entities and determining the frequency of words allowed for the ranking of sentences according to their frequency counts. The sentences with the highest ratings were used to create the summary.Similar research was done on CNN's Daily Email dataset by Named entities from the news articles were identified using SpaCy's NER Model. Tokenizing the detected entities and determining the frequency of words allowed for the ranking of sentences according to their frequency counts. The sentences with the highest ratings were used to create the summary.

Both studies emphasize that incorporating named entity recognition in news dataset for generation of headline can yield advantageous results, facilitating more efficient and informative summarization processes.

2.3 Transformer based model on Text Summarizing task

Jha et al. (2023), made use of transformer model for long text documents along with simple models to ensure document matching in source text and destination text.The goal was to determine whether the document matches as similar or not. This experiment is conducted in order to understand the models understanding on semantics of the documents when large documents are considered.ACL Anthology Network Corpus (AAN), Wikipedia Articles (WIKI), and USPTO Patents (PAT), each including balanced pairs of similar and dissimilar articles, are three common long document datasets on which the models are assessed.On all three datasets, the simple neural models—DSSM, ARC-I, and HAN—perform better than transformer-based models with GloVe and Doc2Vec embeddings. In addition, compared to transformer-based models, simple models require substantially less training time. Simple neural models are more effective for analysing lengthy documents because they use less memory and energy.Additionally, the models' resistance to document length and text perturbation are evaluated. Transformer-based models are less resistant to variations in document length than simple models, which consistently perform throughout a range of document lengths. Similar to this, transformer-based models significantly degrade in performance while simple models remain robust to text disturbance.

Ranganathan and Abuka (2022) In their study made use of transformer based sequence to sequence model for summarizing large corpus of news data.The T5 small's variant was chosen for fine tuning the dataset and summarization task was performed by considering the computational constraint. The model performance and results was compared with other summarization techniques which include BERT based and LSTM models in order to understand the relative strength and weakness of the model.A comprehensive analysis was conducted T5 variants, hyper parameter tuning and comparison of other models. The results was evaluated using ROUGE score tool kit. The conclusion of the paper states that the model performed quite well on dataset considered compared to other summarization models. Further the study conducted by Gupta et al. (2022) shows similar investigation where T5 model was used for summarization task of news articles and was compared with other transformer based model using hugging face library even in this study T5 model performed better than other transformer models indicating how this model is suitable for summarizing the text document especially for news articles.

2.4 Transformer models on headline generation task

The presented research by Bukhtiyarov and Gusev (2020) Utilising improved Transformer-based models, the study focuses on abstractive headline generation and assesses the effectiveness of the methods using Russian datasets. Two pre-trained models are compared in the study: BertSumAbs and mBART. The earlier model was pre-trained on a subset of 25 languages and is a typical Transformer-based model with encoder and autoregressive decoder. The latter uses a randomly initialised Transformer as the decoder and a BERT as the encoder. The encoder has been pre-trained using news and Wikipedia data from Russia.The study uses BLEU score and ROUGE metrics to measure the accuracy of the generated summaries in order to assess the models. RIA and Lenta, two Russian news datasets, are used in the evaluation. According to the results, BertSumAbs performs better than mBART on both datasets, setting a new standard for headline development.Annotators compared generated headlines with headlines that have been authored by humans as part of the human review process. The model's ability to generate high-quality outputs is demonstrated by the fact that BertSumAbs headlines were commonly preferred by annotators over human-generated headlines.The error analysis draws attention to frequent errors made by the models, including repetitions, errors at the subword level, and factual errors. Despite these mistakes, the models consistently generate concise, linguistically accurate summaries.The study concludes by showing the value of optimising Transformer-based models for Russian abstractive headline generation. With its language-specific encoder, BertSumAbs outperforms mBART in terms of performance. The study lays the groundwork for additional research in this area and advances headline generation task in the Russian language.

A similar research conducted by Li et al. (2021) also uses transformer based generative pre-trained model instead of traditional framework by using encoder decoder architecture with attention.Multi-head attention is added to the model to better interpret input tokens by capturing semantic representations and attention distributions. A noteworthy example of an effort to increase content diversity is the integration of sentiment and part of speech features into a rich feature input module. The study offers a pointer generation model as a solution to the out-of-vocabulary issue, illuminating real-world problem-

solving for the creation of short texts. The use of linguistic attributes from n-grams to update hidden states provides a well-thought-out approach to enhancing language coherence. The efficacy of the model is attributed to the decoding procedure, which resembles human reading. According to empirical findings, news headline creation datasets show competitive performance. However, acknowledged difficulties like vocabulary problems and sporadic inaccurate word production suggest there is still space for improvement. It looks promising that the article will improve feature representation and word generation precision. Overall, the research presents a fresh viewpoint on headline creation, but a more thorough investigation of its drawbacks, apparent biases, and comparison with other approaches would enhance its value.

A research gap in the area of creating news headlines has been identified by the literature study. It has been noted that while headlines frequently only contain one sentence, creating them doesn't always require access to the complete content of the article. In order to generate headlines, it is therefore more beneficial to use only those sentences from the article that include key concepts. The suggested research intends to evaluate the effectiveness of a fresh headline creation approach in order to close this gap. This technique combines named entity recognition with an attention-based transformer model. Finding out if this proposed model performs better than conventional headline generation models is the main goal. The study aims to assess the efficiency of the novel strategy in producing more precise and contextually relevant headlines when compared to traditional methods by merging named entity recognition and attention mechanisms.

3 Methodology

3.1 Business Understanding

Data distribution has significantly increased as a result of the amount of news data that is now accessible in many physical or digital formats. Numerous applications rely on summarising engines to provide condensed versions of text, making data summarization a commonly used use case. It's critical for the retrieved news headline to be contextually appropriate because the media business, in particular, frequently uses brief news snippets and attention-grabbing headlines to attract users.This study focuses on investigating how the NER approach may identify significant sentences with semantic properties. These phrases are practised, and their effectiveness is assessed when used to create headlines from news articles. To ensure a smooth and organised execution, the study adheres to the KDD (Knowledge Discovery in Databases) methodology.

3.2 Data Understanding

The BBC dataset that is used as data source for the purpose of this study is news article data set from BBC news. This is a publicly available data set and is used for non commercial and research purposes.The data set is openly available in Kaggle website but was originally sourced from the research work presented by Greene and Cunningham (2006). This data set contains 2225 documents collected from BBC news website for the year 2004 to 2005.The news articles in the data set cover five major subject areas which include technology, business, sports, entertainment and politics. This dataset have been previously used for variety of machine learning tasks like text classification, topic

modelling and text summarization.The data set contains long news article along with their respective headline.This is labelled data pertaining to this research and can also be used to fine tune the machine learning models. The dataset was used adhering to BBC's copyright and terms of using this resource.[1]

3.3 Data Preparation

It is an crucial task to align the data for the purpose of the domain its meant to work for. The news data contain long news article along with associated headline the data was pre-processed which ensures their suitability for machine learning tasks. The pre-processing techniques that was used in this dataset include stemming. Using the Porter algorithm, a popular stemming method, words have been reduced to their root or base form. This method improves the model's capacity to generalise the words by lowering the dimensionality of the data and ensuring that different word variations are regarded as the same word. Stop words that does not add any meaning to the data is also removed from the articles. These terms were removed from the dataset using a list of stop words that includes frequent and pointless words like "the," "and," "is," etc. Eliminating stop words helps to reduce noise and sharpen the emphasis on most crucial content of data.Low frequency terms—more specifically, those that appear in the dataset no more than three times—have been eliminated. Rare terms that might not contain important information for the model are removed in this step.The files in the given archives adhere to certain formats that are appropriate for tasks involving data analysis and machine learning. These pre-processed datasets can now be used in the study to efficiently train and test the machine learning models.

3.4 Model Training

The NER task was carried out with the help of the SpaCy library before splitting the data into train and test sets and fine tuning the machine learning model. Each article was processed by an algorithm that looked for sentences with named entities (NERs). The sentences with NERs were then entered into a new column that had been created. The dataset was then split in to training and testing sets for further process. The training set was used to fine tune the data using T5 model to align the data to news domain. This method can be used to test the model's capacity to judge the coherence of the generated headline that is created.A combination of supervised and unsupervised tasks were used to pre-train the transformer-based encoder-decoder model T5. Each task is transformed into a particular text format with a different prefix since it is developed in a text-to-text format. T5 can handle many jobs by only adding an appropriate prefix to the input, therefore it can perform well on a variety of tasks which includes translation and summarization without the need for substantial fine-tuning.[2]

3.5 Model Evaluation

Following the training phase, the model's effectiveness is assessed using test data from the split. A test article is chosen for the model's input from the test dataset. The "headline" which represent the title provided by the news text's original author, are then contrasted

[1]Dataset: `https://www.kaggle.com/datasets/hgultekin/bbcnewsarchive`
[2]T5-model: `https://huggingface.co/docs/transformers/model_doc/t5`

with the model generated headline. This evaluation procedure aids in determining how closely the headline generated by the model and the title written by a original author of the article correspond. Two evaluation metrics are used to understand the model performance namely ROUGE score and BERT score.

ROUGE score measures the overlapping of n-grams between the headlines created by the model and the actual headline in the data and calculates the ROUGE score.This metric can be downloaded from ROUGE library.It is specialized in evaluating the generated headline in recall oriented manner.ROUGE calculates three main metrics: ROUGE-1 , ROUGE-2 and ROUGE-L. The average of all the ROUGE metrics gives the ROUGE score.[3].Another evaluation metric considered based on its language understanding is BERTScore. Each token in the candidate sentence and each token in the reference sentence are compared to determine how similar they are. This is accomplished by matching terms in the candidate and reference sentences based on cosine similarity and using pre-trained contextual embeddings from BERT models.Furthermore, BERTScore computes precision, recall, and F1 measure, offering helpful insights for assessing different language generation tasks.[4]

4 Design Specification

This section contains comprehensive specifications for all the important entities used in the implementation, as well as an overview of the design flow of the research. The design flow diagram is explain in Figure 1. The basic workflow of the proposed experiment is that the pre-processed data is fed into a function that used named entity recognition from the Spacy library. This function identifies the named entities and picks up these sentences and stores them in new column in the data-frame.Data is then split into train and test sets. The training set is then used to fine tune the T5 model. Once the T5 model is trained it is used to generate the headline.

4.1 Name Entity Recognition

One of the key aspect of natural language processing is named entity recognition (NER), which involves the automatic identification and classification of named entities in text, including people, companies, places, dates, and more. The use of NER, which makes it possible to recognise significant entities embedded inside news items, is essential to the scope of this research project. To accomplish this, the task is carried out using the SpaCy library,[5] which is recognised for its NER capabilities.This method effectively separates and identify sentences that contain relevant items. This tactic encourages the inclusion of only NER-identified sentences during the headline creation process. The research aims to increase the accuracy, coherence, and contextually of the generated headlines by using these sentences. This strategy not only improves the quality of generated headlines, but also makes sure that they are closely related to important details in the articles. A crucial first step towards developing a headline generation model that is more complex,

[3]ROUGE: `https://pypi.org/project/rouge/`
[4]BERTSCORE: `https://huggingface.co/spaces/evaluate-metric/bertscore`
[5]SpaCy: `https://spacy.io/api/entityrecognizer`

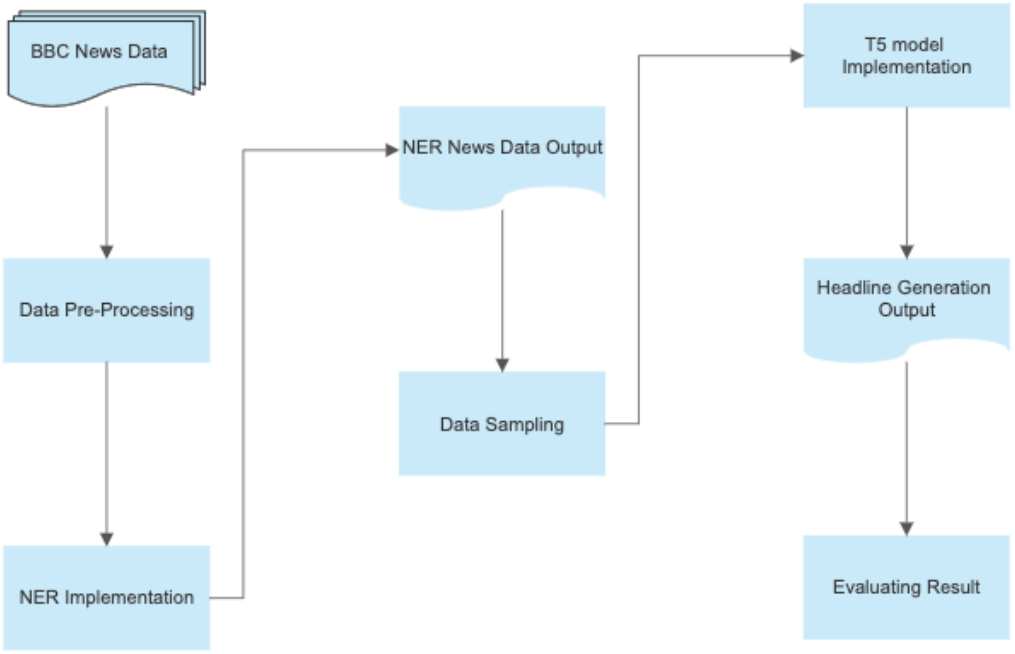

Figure 1: Design Flow

educational, and appropriate and that accurately captures the core of news items is the seamless integration of NER with headline generation.

4.2 Text to Text Transfer Transformer Model

Natural language processing (NLP) has seen a transformative innovation with the development of the Text-to-Text Transfer Transformer (T5) model. T5, which is based on the Transformer architecture, is a flexible and strong model that performs well in a variety of NLP applications. Its unique quality is its capacity to convert different NLP tasks into a common text-to-text format. T5 streamlines the modelling procedure and makes it flexible to tasks ranging from translation and summarization to question-answering and more by considering every activity as a text generation problem.Raffel et al. (2020)The encoder and decoder in T5 are made up of multi-layer Transformer blocks, and the architecture is designed on a conventional encoder-decoder framework. The decoder produces the corresponding output text while the encoder processes the input text. The versatility of T5 is what sets it apart; by simply adding a suitable task-specific prefix to the input, it may be fine-tuned on data sets specific to particular tasks. Due to this feature, T5 can be customised to meet a variety of NLP difficulties without requiring significant architectural modifications.As per the architecture design in Figure 3 Each task we look at requires text input into the model, which is subsequently trained to generate particular target text. Using the same model, loss function, and hyper-parameters for a variety of tasks, including translation, linguistic evaluation, phrase similarity, and document summarization, allows us to preserve consistency. This consistent method not only ensures a standardised evaluation framework but also acts as a constant baseline for the strategies examined in our empirical investigation.

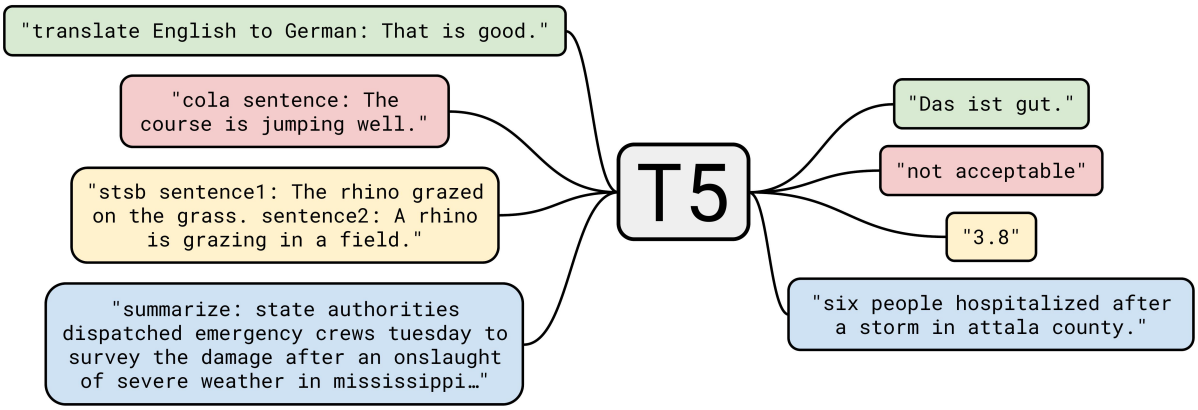

Figure 2: T5 Architecture
Source: Raffel et al. (2020)

This research has carefully planned the combination of Named Entity Recognition (NER) and the Text-to-Text Transfer Transformer (T5) paradigm. The sentences that include relevant entities are found and isolated from the news articles after using the SpaCy library for NER to identify them. These NER sentences are then easily included as input into the T5 model. The architecture of the T5 model naturally supports text-to-

text transformation, with the NER sentences acting as the source text and the creation of coherent and detailed headlines as the end goal. The NER phrases serve as the input text for the T5 model, which then translates them into the output headlines in this method, turning the headline generating process into a translation problem. The T5 model's ability to produce coherent and contextually rich text is enhanced by this symbiotic integration, which also gives the model the ability to understand and use contextual entities. The design incorporates a painstaking integration of NER and T5, leading to a revolutionary methodology that improves the precision, relevance, and coherence of generated headlines by carefully choosing which extracted entities to use.

5 Implementation

This research project was implemented in Google Colab notebook environment using Python Programming language. Google Colab provides an online platform That combines a Jupyter Notebook user interface together with access to GPU resources, making it possible to build, run, and analyse code quickly. Using the colab environment the research was implemented which includes code, experiment, and documenting the entire research process in a cooperative and interactive way. This selection of tools enabled efficient investigation and evaluation of the suggested models and approaches by facilitating the seamless integration of diverse libraries, frameworks.

5.1 Data Loading

Data required for this research was sourced from Kaggle website, an open source data science platform where data sets for various machine learning research can be obtained. The data sourced was BBC news articles which contains 2225 documents and contains information such as category, title and content. These articles are related five different areas such as technology, business, sports , entertainment and politics. These data set was mainly used for classification task. The articles collected was from the year 2004 to 2005. Data was stored in csv format to the local system and was then stored in google drive. The research was implemented using colab notebook. Hence the data was easily loaded to colab notebook from the drive. From this the data was analysed to understand the structure that aids EDA.

5.2 Data Pre-Processing

The CSV file that is loaded into colab notebook was then pre-processed before model implementation. Since the news article is textual form of the data it contains noise in various form and eliminating this becomes an important task. The data contained filename column apart from the id column, which did not have much contribution to the data knowledge was removed from the dataset.

5.2.1 Data Cleaning

Several crucial text preprocessing methods were used in the context of data cleaning for the implementation of this research in order to improve the quality and consistency of the textual data. First, all text was converted into a uniform casing format by using a procedure called lower casing. This was crucial for treating all capitalization variations

equally, whether they were in uppercase, lowercase, or mixed case. The next phase entailed removing punctuation, HTML tags, and links. By eliminating characters that did not add sense to the text, a standard text representation was achieved.
Stopwords and frequently occurring words were removed to further improve the data. Frequently used words like "the" and "a" were eliminated since they frequently offered little to no useful information to later studies. Additionally, stemming was carried out, which reduced inflected words to their root form in order to maintain consistency across different spellings of the same term. Lemmatization was used in conjunction with this to change derived words into their base or root form, guaranteeing the inclusion of legitimate root words in the language.Lastly,contraction mapping was used to expand shortened words or syllables, improving readability and maintaining context. Together, these data cleaning techniques were crucial in improving the raw textual data and getting it ready for our research's modelling and downstream analysis. The resulting standardised and clean text data served as the basis for our subsequent work.

5.2.2 Data Sampling

As mentioned earlier the dataset contains 2225 documents. The data is split into train and test using train test split function from scilit-learn library like scikit-learn to split the DataFrame into training, validation, and test sets.Table 2 illustrates the data split

Table 1: Data Split.

Data	**Instances**
Train	1780
Test	222
Validate	223

5.3 NER Implementation

After the news article was cleaned using various pre-processing steps Named Entity Recognition (NER) Model was implemented using spaCy library. A function is created that analyses content of each cleaned article from the dataset. This function is designed to identify sentences in the article that contains named entity using the SpaCy natural language processing tool kit. This procedure aims to extract meaningful sentences that contain entities with precise names, such as individuals, locations, businesses, or other pertinent terms.The step by step implementation of NER is explained in below points

- Every article in the Dataframe is passed through a function which first extracts the content of the article.
- The content is tokenized and parsed for structural analysis using spaCy's NLP capabilities.The article's sentences are each read through to look for Named Entities.
- The function determines whether any constituent token in a sentence belongs to a Named Entity. To accomplish this, spaCy's ent_type property is evaluated.
- Named Entity sentences are thought to include crucial entity-specific information. A list called processed_sentences contains these chosen sentences.

- The chosen sentences are combined to form a coherent paragraph after each sentence in the article has been critically examined. This sentence contains Named Entities, collecting the essential entity-related information. The output sentence is kept in a brand-new DataFrame column called "ner_content." As articles are aligned with their pertinent entity information, processed paragraphs are stored in the 'ner_content' column.

5.4 Model Building and Training

After the data was split into train and test. The training set of data was used to fine tune the T5 model. A T5 tokenizer was used to tokenize the data and make it to a form that is understandable by the T5 model. To achieve this a class is designed that takes raw article and headlines as its input and transforms them into tokenizers. The methods in the class also converts the tokens input_id's and attention masks. input_id's are where each token of words are converted into its numerical token that models like transformers can easily understand.A distinct ID is given to each token from the model's vocabulary. Each of this input_id's will have an attention mask which indicates which token from the input should be given attention and which can be ignored.By transforming textual data into numerical representations and concentrating attention on pertinent tokens while ignoring padding or extraneous tokens, input IDs and attention masks enable models like Transformers to process and comprehend textual data.

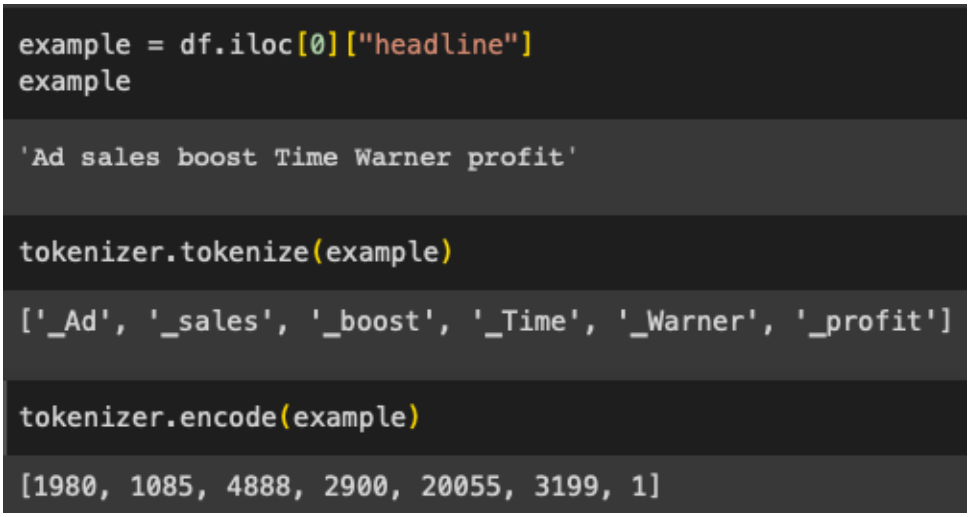

Figure 3: An Example Output from tokenization

After completion of the tokenization task, The model is fine-tuned on the training dataset using the seq2seq model of the T5-small version from the Hugging Face's transformers library.A data loader class is written that facilitates the data loading and preparation process for training and evaluating a T5 model. It serves as a vital link between the training pipeline for a T5 model and the raw dataset.A T5 tokenizer, training, testing, and validation DataFrames, as well as other configuration hyper parameters like batch size and token length limitations, are all initialised, The Dataset class which preprocesses and encodes the text data, is used by the setup function to orchestrate the production of dataset instances. Data loaders that are optimised for each step are produced by

later methods like train_dataloader, val_dataloader, and test_dataloader. These loaders guarantee effective data batching, training-related data shuffles, and simultaneous data loading.

Table 2: Hyperparameter values.

Hyperparameter	**Value**
Evaluation Stratergy	epoch
Learning Rate	0.0001
Train Batch Size	8
Eval Batch Size	8
Train epoch	5
epocs saved	least val_loss
fp32	true

The validation loss at each validation epoch is stored. In order to understand the model performance and convergence during training. Using this data, the validation loss for each epoch is visualized over time to determine if the model is improving or possibly over fitting. the figure Figure 4 displays the loss validation graph for 5 epochs considered during training.The epoch with lowest validation loss is stored and then used for testing purpose.

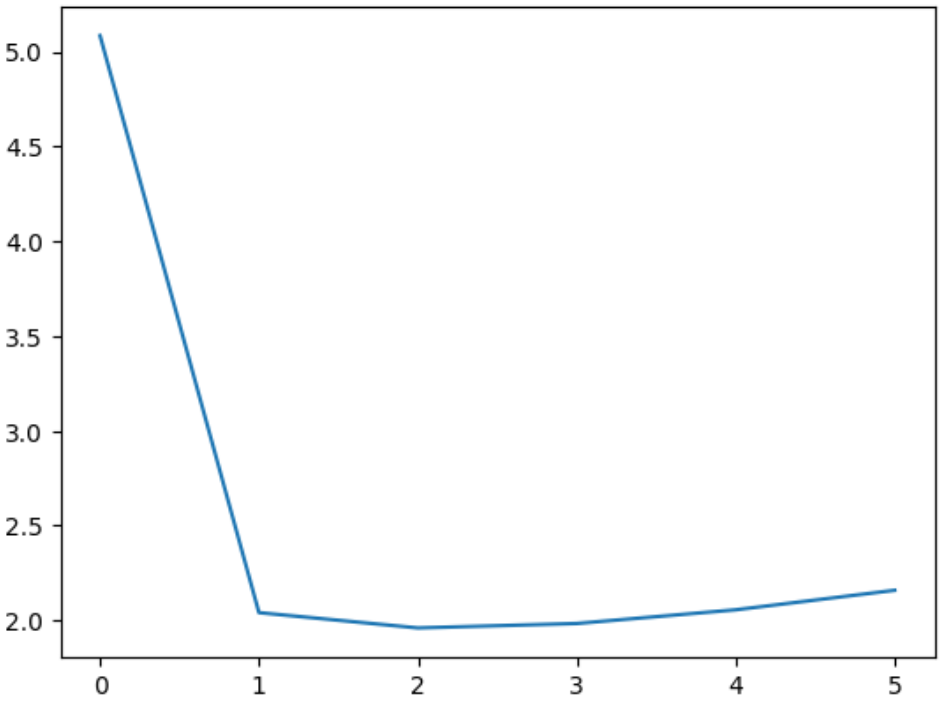

Figure 4: No of epochs vs val_loss

In order to test the ability of the trained model testing set is used. A function named generate_headlines is written which takes the testing dataset, encodes them and uses the trained T5 model to generate headline based on the input. To improve the quality of the generated text, this method makes use of beam search, a technique that investigates a

variety of possible word sequences. The final headline is then created by decoding the generated IDs using the same tokenizer. The created headline, which succinctly captures the core of the supplied text, is what the function ultimately returns. By automating the creation of informative headlines, this method of using pretrained models for text generation demonstrates the effectiveness of natural language processing models in the production of original content. An example of actual headline and generated headline is depicted in the Figure 5

```
sample_row["headline"]

'Ban on hunting comes into force'

generate_headline(text)

'Fox hunting banned in uk'
```

Figure 5: generated headline

6 Evaluation

The evaluation of the results i.e the comparison of generated headline with actual headline is done by using two evaluation metric known as ROUGE tool kit and Bert Score.

The ROUGE family of metrics essentially measures the amount of time that the generated text and the reference text's n-grams (constant sequences of n words) coincide. ROUGE-1 (unigrams), ROUGE-2 (bigrams), and ROUGE-L (longest common subsequence) are the three primary ROUGE measures. These metrics reveal how effectively the generated text reproduces the key ideas and logical structure of the reference text. Better text generating performance in terms of content overlap is indicated by higher ROUGE scoresLin (2004).Comparatively, BERTScore assesses text production using contextual embeddings from BERT (Bidirectional Encoder Representations from Transformers) models. BERTScore calculates similarity scores for each token in the candidate (produced) text with tokens in the reference text, in contrast to conventional n-gram-based metricsZhang et al. (2020). This method considers the meaning and context of words, leading to a more thorough review. By calculating precision, recall, and F1 score, BERT Score offers a more complex assessment of text quality.

6.1 Experiment Pre-trained Model

The first experiment was conducted using the T-5 model directly i.e without fine tuning the model with the train data. The test data was passed to T-5 model and the results are evaluated.

Table 3: ROUGE Score with Pre-trained Model

ROUGE	**F1**	**Precision**	**Recall**	**Score**
ROUGE-1	0.09	0.08	0.10	0.06
ROUGE-2	0.009	0.008	0.01	0.009
ROUGE-L	0.09	0.08	0.10	0.06
Average				0.1

For ROUGE evaluation the result seems to be pretty low. The average score for ROUGE with the pre-trained model is 0.1 which indicates that there is no much similarity between the actual headline of the news article and the generated headline. The average of ROUGE-1 , ROUGE-2 and ROUGE-L also seems to be low.

Table 4: BERT Score

AvgP	**AvgR**	**AvgF1**	**Avg**
0.844	0.851	0.847	0.84

In order to better understand the context of the generated headline the Bert score evaluation metric was used this gives the context or semantic score of the generated text using cosine similarity. The Bert score is 0.84. Which indicates that nearly 84 percent of text matches with the actual headline with respect to its context.The further experiment was conducted by pre-training the T-5 model with the training set and then generate the headline.

6.2 Experiment without NER Implementation

For the purpose understanding if the proposed hypothesis works better than the existing methodology. The model was implemented on news article without taking into consideration the NER and the NER sentences from the article but training the T-5 model with the training set first. The evaluation methods are then applied on the generated headline to understand and interpret the results. The results of both evaluation metric is displayed in the table below.

Table 5: ROUGE Score without NER Implementation

ROUGE	**F1**	**Precision**	**Recall**	**Score**
ROUGE-1	0.30	0.31	0.29	0.3
ROUGE-2	0.09	0.09	0.08	0.08
ROUGE-L	0.28	0.29	0.28	0.28
Average				0.2

The results of ROUGE score is described in Table 5.The results when compared with experiment 1 seems to be better,there is a variable amount of increase in from 0.1 to 0.2.

The results shows low score of ROUGE-2 when compared with ROUGE-1 and ROUGE-L. This represents that the bigram of the generated headline poorly overlap with the actual headline.However F1 value of ROUGE-1 displays shows sufficiently higher value representing of 0.3 indicating 30% of the unigrams in generated headline are overlapping that of the actual headline. The average ROUGE score here is 0.2 indicating there is 20% similarity between the headlines.

Table 6: BERT Score

AvgP	**AvgR**	**AvgF1**	**Avg**
0.898	0.897	0.897	0.89

The Bert evaluation score is also generated for the purpose of understanding the semantics of the generated headline and the actual headline when NER sentences are not taken into consideration in Table 6. The average results of the BERT score metric is 0.89. This automatic evaluation metric compares each token in the candidate sentence to each token in the reference sentence to determine how similar the two sentences are. The total BERT Score is then calculated by averaging these token-level similarity ratings. With a BERT Score of 0.89, it is possible to infer that, on average, the generated sentences and the reference sentences have many similarities in common in terms of word choice and context. This suggests that the text generation model has performed well in terms of the BERTScore criterion.

6.3 Experiment with NER Implementation

The final experiment conducted is the proposal of this thesis i.e by including only NER sentences to the from the article to the study.In order to evaluate the results and compare them with experiments done before ROUGE and BERT score metrics are used.The results of the metric are displayed in the table below.

Table 7: ROUGE Score with NER Implementation

ROUGE	**F1**	**Precision**	**Recall**	**Score**
ROUGE-1	0.41	0.42	0.29	0.3
ROUGE-2	0.17	0.17	0.17	0.1
ROUGE-L	0.40	0.40	0.40	0.4
Average				0.3

When ROUGE evaluation metric is calculated for the generated headlines when implemented with NER the scores show a slight higher value compared to that of previous values. The ROUGE-1 and ROUGE-l values are 0.3 and 0.4 indicating 30% of unigram overlap with the generated sentences and 40% of the longest sentences match with that of the actual headline. The ROUGE-2 value is 0.1 which is slight higher then that of the previous experiment but does not indicate much difference.The average score for ROUGE metrics in 0.3 indicating that there is 30% similarity of words of generated headline with actual headline.

Table 8: BERT Score

AvgP	**AvgR**	**AvgF1**	**Avg**
0.919	0.918	0.919	0.91

The BERT evaluation metric is also calculated for the experiment conducted with NER. The average value of the result is 0.91 which indicates there is 91% similarity between the actual headline and generated headline with respect to similarity and context.

6.4 Discussion

The research was conducted to generate headline for news articles using Text to Text Transfer Transformer model incorporating the NER model. As discussed in the literature review the previous research papers this task was attained using various machine learning models including transformer models with the purpose of text summarization.T5 model was selected for this experiment for its versatile language model architecture with the ability to handle wide range of natural language processing tasks.The idea was to enhance the ability of generating headline for news article hence the implementation of NER model along with the T5 transformer model is taken into consideration. The model was designed and implemented as per the discussed in the sections above.

Two evaluation metric ROUGE and BERT Score was calculated in order to understand the sentence formation and semantic context of generated headline when compared with the actual headline of the new article.The experiment was conducted both with NER implementation and without NER implementation. The results of the model show comparable higher values with each experiment conducted.The similarity seems to be highest with last experiment where NER sentences are used. This can be an indication that for the task of headline generation the implementation of NER gives necessary information for the model to generate more accurate summaries or headlines that are similar in terms of context and sentence formation.

7 Conclusion and Future Work

The primary objective of this research was to investigate how the Text-to-Text Transfer Transformer (T5) model may be used to produce headlines from news stories. The goal of the study was to determine whether using T5 for headline generation task and incorporating Named Entity Recognition (NER) will improve the quality of the generated headlines. The focus of the study was on whether the T5 model could produce informative and cogent headlines that accurately summarise news items after being adjusted and given NER.A thorough literature analysis conducted at the start of the study highlighted the importance of automated headline production and the potential of transformer-based models like T5. The tasks were to pre-train T5, incorporate NER, generate headline and assess the model's effectiveness using common metrics like ROUGE and BERTScore. Data pre-processing, dataset development, model architecture design, and evaluation procedures were all included in the implementation process.

The study's findings showed encouraging progress. The refined T5 model, using NER

data, demonstrated the capacity to produce coherent and contextually relevant headlines when compared with the other experiments conducted. The generated headlines showed a substantial degree of closeness and quality to actual title of the article which was human-authored according to the ROUGE and BERT Score evaluations. This is in keeping with the goals of improving the headline quality through context enrichment led by NER.The main outcomes of this study fall into two categories. First off, news item headlines may be efficiently generated using the improved T5 model with the help of NER. The headlines produced also met excellent ROUGE and BERT Score parameters, indicating a significant degree of alignment with reference headlines. This highlights the importance of NER integration as a source of context enrichment in addition to reiterating the usefulness of T5 for creating headlines.

This research does have certain restrictions, though. Because of the model's reliance on training data and NER's domain-specific nature, headline production may be inaccurate or biased. In some circumstances, the model's effectiveness may also be hampered by its inability to comprehend intricate news themes in depth.The research was also conducted on the small dataset for the current study. Future research could concentrate on overcoming these constraints by looking into how to lessen bias, improve model interpret-ability,use more input data and optimise models for only certain specific news domains.

8 Acknowledgment

The researcher would like to express gratitude to Prof. Prashanth Nayak for his insightful advice and help with this project. His steadfast assistance and subject-matter knowledge helped to maximise this research's technical and intellectual potential.

References

Alshibly, I., Al-Shorfat, S., Otair, M., Shehab, M., Tarawneh, O. and Daoud, M. S. (2023). Text summarization of news articles based on named entity recognition using spacy library.

Berezin, S. and Batura, T. (2022). Named entity inclusion in abstractive text summarization, pp. 158–162.

Bukhtiyarov, A. and Gusev, I. (2020). Advances of transformer-based models for news headline generation, *CoRR* **abs/2007.05044**.
URL: *https://arxiv.org/abs/2007.05044*

Greene, D. and Cunningham, P. (2006). Practical solutions to the problem of diagonal dominance in kernel document clustering, pp. 377–384.

Gupta, A., Chugh, D., Anjum and Katarya, R. (2022). Automated news summarization using transformers, pp. 249–259.

Jabeen, S., Shah, S. and Latif, A. (2013). Named entity recognition and normalization in tweets towards text summarization, pp. 223–227.

Jha, A., Samavedhi, A., Rakesh, V., Chandrashekar, J. and Reddy, C. K. (2023). Transformer-based models for long-form document matching: Challenges and empirical analysis.

Li, P., Yu, J., Chen, J. and Guo, B. (2021). Hg-news: News headline generation based on a generative pre-training model, *IEEE Access* **9**: 110039–110046.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries.
URL: *https://api.semanticscholar.org/CorpusID:964287*

Marek, P., Müller, Konr, J., Lorenc, P., Pichl, J. and Jan (2021). Text summarization of czech news articles using named entities, *Prague Bulletin of Mathematical Linguistics* **116**(1): 5–26.

Mena, S. and Palazzo, G. (2012). Input and output legitimacy of multi-stakeholder initiatives, *Business Ethics Quarterly* **22**(3): 527–556.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer.

Ranganathan, J. and Abuka, G. (2022). Text summarization using transformer model, pp. 1–5.

Rehman, T., Das, S., Sanyal, D. K. and Chattopadhyay, S. (2022). Abstractive text summarization using attentive GRU based encoder-decoder, pp. 687–695.

Wang, M., Xie, P., Du, Y. and Hu, X. (2023). T5-based model for abstractive summarization: A semi-supervised learning approach with consistency loss functions, *Applied Sciences* **13**(12).
URL: *https://www.mdpi.com/2076-3417/13/12/7111*

Widyassari, A. P., Rustad, S., Shidik, G. F., Noersasongko, E., Syukur, A., Affandy, A. and Setiadi, D. R. I. M. (2022). Review of automatic text summarization techniques methods, *Journal of King Saud University - Computer and Information Sciences* **34**(4): 1029–1046.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. and Artzi, Y. (2020). Bertscore: Evaluating text generation with bert.