Named Entity Recognition and Headline Text Generation for News Articles

Amirsaman Arabali
saman_arabali@berkeley.edu

Sri Yanamandra
sri.yanamandra@berkeley.edu

Sudhrity Mondal
sudhrity@berkeley.edu

W266 Spring 2022, UC Berkeley School of Information

Abstract

This paper discusses the implementation of Natural Language Processing (NLP) based news headline generation using Named Entity Recognition (NER) and various text summarization techniques. Named entities in news articles (COVID news & CNN news) are detected using a BERT-based NER model and used to select the highest ranked sentences in the article that are then passed through multiple BART/T5 – based summarization models and experiments for training and headline generation. The generated headlines are scored using various metrics, including Rouge, Bleu, Bleurt, and Bertscore. Our simulation results show that the models that are custom-trained on specific topics such as COVID news perform better than the ones custom-trained on a wide range of topics such as general CNN news articles. Our results also show that the models trained on the full articles perform better than the ones trained on the filtered articles or a summary of the articles.

1 Introduction

Machine learning and NLP is becoming an industry-changing technology for various domains. Text data in articles, news, Twitter feeds, social media posts, and web content is increasing exponentially. Analysis and understanding of these texts can be challenging and take time. NLP-based text headline creation is an approach to summarizing paragraphs of text into a sentence for a quick and more straightforward understanding of the text and for text classification. The headline text can include named entities present in the text for easy identification. In this paper, we used two datasets for headline text summarization:

1. COVID news articles
2. CNN news articles

In this work, NER is used to identify named entities, and the highest-ranked sentences are filtered to train the BART/T5 models for the prediction of headline text. Variations of experiments include training the headline summarization models (BART/T5) using full article text and pre-summarized text using extractive, abstractive, and extractive-abstractive techniques. Based on the current research, there is no consensus on how to validate and score text summaries, including headline texts. To address this issue, we used multiple validation metrics (Rouge, Bleu, Bleurt, Bertscore) and manual validation.

2 Background

Text summarization is categorized into two general groups: extractive and abstractive. The extractive creates a summary from the sentences of a document with the highest scores, while the abstractive shorten parts of the document to create a contextual summary. We use both extractive and abstractive summarization in this work. With headline text generation, we attempt to summarize the article using a single sentence.

2.1 Related Work

Text summarization has been the topic of many research articles. A general framework is proposed by (Yang Liu et al. 2019) for both extractive and abstractive text summarization by introducing a novel Bert-based encoder that is able to express the semantics of a document and obtain representations for its sentences. The extractive model is built on top of the encoder by stacking several Transformer layers to extract document-level features. The abstractive model combines the pre-trained Bert encoder with a randomly-initialized Transformer decoder to create an encoder-decoder architecture. A hybrid pointer-generator network is used by (Abigail See et al. 2017) to copy words from the text source by pointing at particular words, which helps with accurately reproducing the information while keeping the capability to produce new words. They use coverage to keep track of summarized text, which discourages repetition. The model has been applied to the CNN/Daily Mail. A NER corpus based on Tweetbank V2 (TB2) was created by (Hang Jiang et al. 2022) and used to train NLP models. They trained a Stanza NER model and achieved competitive performance against non-transformer-based NER systems. In (Yen-Chun Chen et al. 2018), a

hybrid extractive - abstractive summarization architecture is proposed that first selects important sentences from the text and then writes an abstract by reconstructing sentences. They use policy-based reinforcement learning to bridge together the two neural networks. Since their algorithm first operates on the sentences and then on the words, they are able to achieve much faster training/inference speed by enabling parallel decoding of the neural generative model than the previous paragraph encoder-decoder models. A NER-based extractive summarization of the text is implemented in (Štěpán Műller, 2019). LSTM is used as an encoder-decoder to provide an entity name for each token. An attention block is used to let the decoder pay attention to different hidden states of the encoder. The named entities are used to score the sentences, and the text summary is the extracted top-ranked sentences. Important sentences are crucial for the task of pre-training an abstractive summarization (Jingqing Zhang, 2020).

2.2 Motivation

Headline text generation is a special case of text summarization. A standard base summarization includes an encoder-decoder architecture. The decoder generates token by token. One of the important issues with summarization models like BERT is the limited number of tokens. By default, BERT limits the number of tokens to 512. This limitation helps BERT with better memorization and also to keep the simulation time reasonable as transformers have quadratic compute costs in the length of the input text sent in the model. Assuming the average length of a sentence in English is 25 words, that would give 15 sentences, that is a paragraph or two at most. Therefore, the idea is to use NER to filter the important sentences and feed those to the summarization model rather than feeding the entire text.

Below are our contributions:

- Custom-trained a BERT – based NER model on the CORD-19 dataset (Wang et al. 2020)
- Created a pipeline that starts with reducing article size by 50% using custom NER predictions which are then used as inputs to custom train BART/T5 models for headline text generation
- Custom-trained BART/T5 models on the COVID news dataset
- Custom-trained BART/T5 models on the CNN news dataset

Our hypothesis is that reducing the article length by using stacking techniques such as NER sentence filtering and pre-summarization will improve the headline text generation performance when compared with the headline text generated from the first few sentences of the article text.

3 Methodology

We started with COVID news articles first. The text and titles from the COVID dataset were not always consistent with the content of the article. In addition, the article language did not seem to be professionally authored and, at times, incoherent. As a result, we felt the need to use an additional dataset that had professionally authored articles for the study. We chose the CNN news data along with COVID.

We used two separate pipelines to process these two datasets. **Figure 1** shows the schematics of our models developed for headline text generation using the COVID news dataset, and **Figure 2** and **Figure 3** are the schematics of our models for the CNN news dataset. The intuition behind training multiple models on CNN and COVID news is that the COVID dataset is on a specific topic while CNN news carries articles across many topics.

3.1 Datasets

We have used three datasets in this paper:

- Ran Geva's "Free dataset from news/message boards/blogs about CoronaVirus" IEEE Dataport.
- COVID NER data CORD-19 CORD-NER dataset covers named entity annotations for 75 fine-grained entity types for over 500K scholarly COVID articles. (Wang et al., 2020)
- CNN News (from CNN/Daily Mail)

3.2 COVID news

The COVID data comprised news articles from Dec 2019 to March 2020 in 31 JSON files with 3062112 records, of which the first eight files (765528 records) were used. Not all news articles in the dataset were related to COVID, and a subset of the articles was selected based on the presence of words *'covid'*, *'corona'*, *'virus'*, *'mask'*, *'shelter-in-place'*, *'vaccine'* to get a better representation of COVID news articles. The data was pre-processed and split into a train (80%), valid (10%), and test (10%).

Figure 1 shows our approach for the COVID news dataset. **1)** Our first model uses BART to perform an abstractive headline text generation on the news articles. We custom-trained BART-large on the COVID news training dataset and performed headline predictions on the COVID test dataset. **2)** Our second model is a mix of extractive - abstractive headline

summarization. We use a BERT-large model custom-trained on the CORD-19 dataset of COVID-related NEs to perform a named entity recognition task on both COVID train and test data. The results of the NER algorithm are then used to perform an extractive summarization by scoring the sentences based on the predefined weights associated with the named entities they contain and extracting the top high score sentences. Here, the idea is that the sentences that contain the highest number of named entities carry more information about the text and, therefore, should be in the text summary. Those sentences that include "who", "where," and "when" would likely be better candidates to form the text summary rather than sentences that do not include any named entities (Štěpán Műller, 2019). The score of the sentences is calculated by summing up the weights of the named entity of each token. The custom-trained BART-large summarization model takes the filtered sentences as inputs and returns the abstractive headline prediction. For experimentation, we used variations of the full article text, the first three sentences of the articles, and the top 50% sentences containing COVID NEs as input data for the two models described above. Finally, we trained T5-base models and experimented with all the combinations as described for the BART models above.

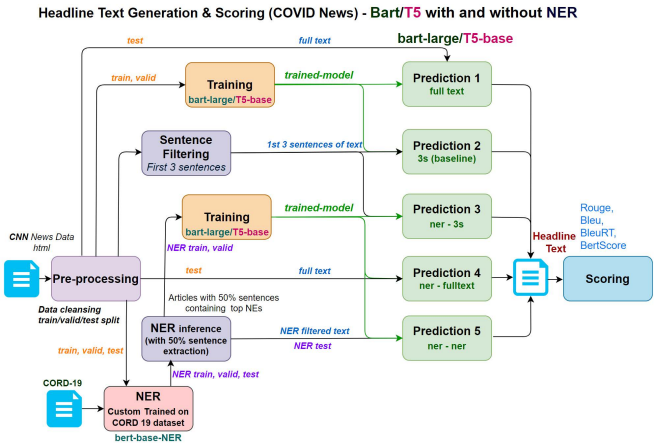

Figure 1: Summarization Flowchart for COVID news

3.3 CNN news

The **CNN news** data comprised 92579 HTML files, each containing a news article from 2006 to 2013. The *title* and *text* from the HTML files were read using a *newspaper.Article* package pre-processed and split similar to the COVID dataset. **Figure 2** and **Figure 3** show the approach used for CNN news headline generation. We custom train a pre-trained Facebook's BART-large model on the CNN news train dataset and then use the trained model to perform headline predictions on the CNN test dataset. In addition, we have also custom-trained a T5-base model and used it for headline prediction. For both scenarios, we executed the training and tests with the full article text and again with 50% of the text with sentences selected based on NER-based scoring of the sentences. As shown in the figures below, our first two approaches use sentence extraction based on position in the article. Our first sentence-as-a-headline approach is rudimentary, although our expectation is to learn quite a bit about the article after reading the first sentence. We then used a 3-sentence approach to headline generation (Prediction 1). This fared better than the previous approach. These two approaches together form our baselines for the CNN news dataset. We then execute the model to generate headline text from the full CNN news article text (Prediction 2).

We have used a variety of summarization techniques to summarize the original article text into a shorter text and then execute the BART and T5 models to generate headline text. For Prediction 3, 4, and 5, we used summarization of article text using extractive (pre-trained SBERT), abstractive (pre-trained bart-large-cnn), and extractive followed by abstractive summaries, respectively.

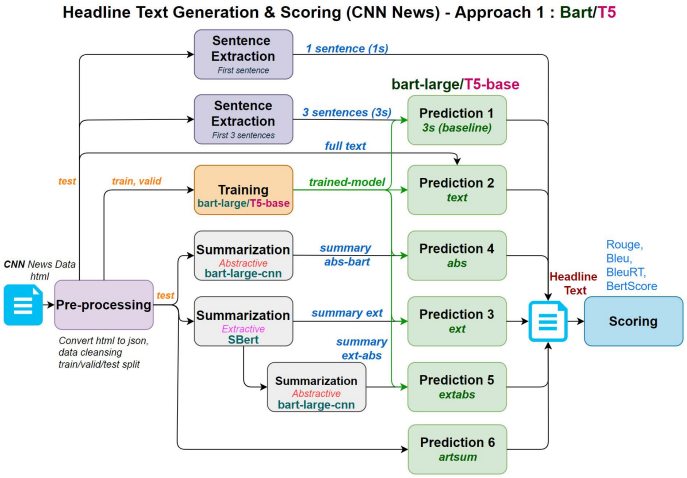

Figure 2: Summarization Flowchart for CNN news – 1

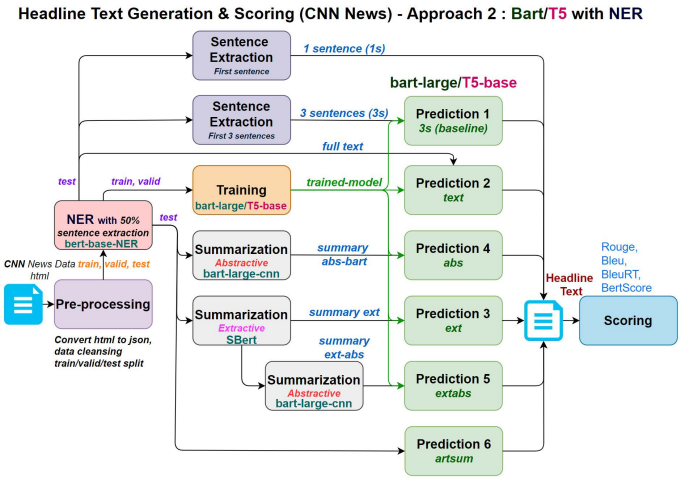

Figure 3: Summarization Flowchart for CNN news – 2

Next, we use a BERT-based (bert-base-NER) model to perform a named entities recognition task on the CNN news data. We do sentence extraction for 50% of the text using NER-based scoring and select sentences with most NERs. We use this to train the BART and T5 models and execute tests as we have done previously. The idea is that the sentences with the highest number of named entities carry more information about the text and, therefore, should be in the text summary. The text is fed to BART and T5 models to perform headline text generation.

3.4 Models

3.4.1 BART and Summarization

In this paper, we have used a BART model from Huggingface to perform abstractive summarization (Simple Transformers). BART is a seq2seq model with the first part of the architecture being a bidirectional encoder and the second part being autoregressive decoders where the previously generated output in each step is fed into the model as an additional input. BART-large has 12 encoder and decoder layers with 400M parameters (Transformers BART).

3.4.2 BERT and NER

The Named entities often provide valuable information regarding what a given text is about. In this project, we have used the NERModel class from Simple Transformers with BERT to perform the NER task and token classification. We custom train the NER model on the COVID dataset. **Figure 8** in the Appendix shows the Named Entities and the distribution of those entities for a sample of 100k COVID records. We use NER to score sentences and filter them based on their importance. BERT is pre-trained on the entire Wikipedia and is fine-tuned by adding one output layer for the NER task, which is done by NERModel.

3.4.3 T5 and Summarization

Text-To-Text Transfer Transformer (T5) offers a vast number of pre-trained models that can be used for language generation, i.e., responses to the questions. We used the T5-base (220 million parameters) model for headline text generation (Simple Transformers).

4 Simulation Results and Discussion

The simulation was done using AWS VM with Ubuntu 20.04 with c5a.8xlarge and g4dn.xlarge-g4dn.4xlarge instance types with Tesla T4 GPU. Due to the lack of a published research baseline for headline text summarization, we used article's first few sentences (1s, 3s) as our baseline.

4.1 COVID news

The scenarios used to experiment with the COVID news dataset is shown in **Figure 1** and listed in **Table 1**. The developed models are based on BART and T5. NER is used for some models to filter important sentences. **Table 1** has the definition of each scenario with the model used. The headline generation is done for 25K articles in the test COVID news dataset, and the generated headlines are evaluated against the original headline (reference) of the article using various metrics. The results are averaged for all the articles in the test COVID news dataset.

#	Experiment Scenarios		Id	Description & Summarization Models
	Model trained on	Test input data		
0	Full COVID articles	First 3 sentences of article	**3s**	**facebook/bart-large** and **T5-base** Test data: 1st 3 sentences filtered using nltk sent_tokenizer
1	Full COVID articles	Full COVID articles	**text**	**facebook/bart-large** and **T5-base** Test data: Full articles
2	50% of sentences extracted using a custom trained NER model	First 3 sentences of article	**ner-3s**	**facebook/bart-large** and **T5-base** NER model: **BERT-large** trained on CORD-19 data set (Wang et al. 2020)
3	50% of sentences extracted using a custom trained NER model	Full COVID article	**ner-full**	**facebook/bart-large** and **T5-base** NER model: **BERT-large** trained on CORD-19 data set (Wang et al. 2020)
4	50% of sentences extracted using a custom trained NER model	50% of sentences extracted using a custom trained NER model	**ner-ner**	**facebook/bart-large** and **T5-base** NER model: **BERT-large** trained on CORD-19 data set (Wang et al. 2020)

Table 1: Experiments with COVID news dataset

Table 2 shows the averaged metrics for different scenarios experimented on the COVID news dataset. Based on the results, the model based on T5 trained on the full COVID articles and tested on the full COVID articles (id: **text**) gives the best scores among all the scenarios developed for the COVID news dataset. According to our results T5 - based scenarios gave better scores compared to BART - based scenarios.

Another observation is the performance of our developed models based on NER filtering is not good, which indicates we are losing some important

contextual information while performing important sentence filtering. This could mean that the headlines in real life do not have a high density of NEs - potentially making the headlines more readable. Filtering-in sentences with a high density of NEs also loses the context of the whole article and thereby creating a lower quality prediction.

id	model	Bleu	Rouge 1 (f)	Rouge 2 (f)	Rouge l (f)	Bert Score (f1)	BleuRT
text	BART	0.509	0.473	0.362	0.461	0.901	-0.162
full-ner	BART	0.459	0.416	0.308	0.405	0.891	-0.272
ner-ner	BART	0.473	0.432	0.326	0.420	0.894	-0.231
ner-full	BART	0.495	0.458	0.347	0.445	0.899	-0.174
text	T5	**0.617**	**0.597**	**0.494**	**0.588**	**0.924**	**0.019**
3s	T5	0.573	0.551	0.451	0.542	0.915	-0.085
full-ner	T5	0.538	0.506	0.400	0.497	0.908	-0.164
ner-full	T5	0.498	0.457	0.343	0.446	0.900	-0.226
ner-3s	T5	0.473	0.433	0.326	0.424	0.896	-0.286
ner-ner	T5	0.450	0.402	0.290	0.392	0.891	-0.334

Table 2: Experiment Scores for COVID news

4.2 CNN news

The scenarios used to experiment with the CNN news dataset are shown in **Figure 2** and are defined in **Table 3**. The headline text generation is done for 9233 articles using the **test** CNN news dataset. The candidate headlines are evaluated against the original reference headline of the article using various metrics. The results are averaged for all the articles in the test dataset. The summarized data (output of each scenario for a sample article) and generated headlines are provided in **Figure 9** and **Figure 6** in the Appendix, respectively.

Table 4, **Table 5**, **Table 6**, and **Table 7** show the Rouge, Bleu, Bertscores, and BleuRT for different scenarios for the CNN news dataset. **Table 4** and **Table 5** are for BART and T5-based scenarios without any NER filtering, while **Table 6** and **Table 7** are for BART and T5-based scenarios with 50% NER-based filtering.

As we have seen from the results in the previous sections, the highest score during generated headline text evaluation was achieved when the headline text was generated from the full article text instead of from the summaries.

We have also investigated the article text summary (**artsum**) produced by the *newspaper.Article* package. We provided this as input to our trained BART model, and the results obtained are slightly better than the results presented earlier in most cases.

#	Experiment Scenarios	Id	Description & Summarization Models
0	Baseline	**1s**	First sentence in article used as headline text and scored
1		**3s**	First 3 sentences in article used for headline text generation and scoring
2	Full text	**text**	Full text from the article is used for headline text generation
3	Extractive Summarization	**ext**	Extractive summary from article text with 10 sentences maximum is used for headline text generation. The model used for summarization is: ***SBertSummarizer('paraphrase-MiniLM-L6-v2')***
4	Abstractive Summarization	**abs**	Abstractive summary from article text with 280 characters maximum is used for headline text generation. The models used for summarization is: ***facebook/bart-large-cnn***
5	Extractive followed by Abstractive	**extabs**	Extractive summary with 10 sentence maximum from article text followed by Abstractive summary from the summarized text with 280 characters maximum is used for headline text generation. The models used for summarization are: ***SBertSummarizer('paraphrase-MiniLM-L6-v2')*** followed by ***facebook/bart-large-cnn***
6	Extractive Summary done by Newspaper3k. Article()	**artsum**	Summary extraction from text using custom keywords; score sentences based on number of occurrences; select top ranking sentences. **(https://github.com/codelucas/newspaper/blob/master/newspaper/nlp.py)**

Table 3: Experiments with the CNN news dataset

id	Bleu	Rouge 1 (f)	Rouge 2 (f)	Rouge l (f)	Bert Score (f1)	BleuRT
1s	0.133	0.142	0.042	0.127	0.834	-0.811
3s	0.314	0.277	0.106	0.259	0.877	-0.667
text	0.350	0.311	0.118	0.290	0.882	-0.595
ext	0.341	0.300	0.111	0.279	0.882	-0.611
abs	0.300	0.258	0.090	0.242	0.874	-0.700
extabs	0.282	0.239	0.079	0.225	0.871	-0.737
artsum	**0.369**	**0.336**	**0.132**	**0.312**	**0.885**	**-0.564**

Table 4: Experiment Scores for CNN news (BART)

id	Bleu	Rouge 1 (f)	Rouge 2 (f)	Rouge l (f)	Bert Score (f1)	BleuRT
1s	0.133	0.142	0.042	0.127	0.834	-0.811
3s	0.290	0.260	0.102	0.245	0.876	-0.767
text	0.316	0.287	0.109	0.271	0.882	-0.721
ext	0.308	0.275	0.104	0.260	0.879	-0.734
abs	0.277	0.241	0.087	0.228	0.872	-0.784
extabs	0.261	0.223	0.078	0.211	0.869	-0.816
artsum	**0.350**	**0.321**	**0.128**	**0.302**	**0.884**	**-0.659**

Table 5: Experiment Scores for CNN news (T5)

id	Bleu	Rouge 1 (f)	Rouge 2 (f)	Rouge l (f)	Bert Score (f1)	BleuRT
1s	0.133	0.142	0.042	0.127	0.834	-0.811
3s	0.313	0.277	0.107	0.259	0.877	-0.667
text	0.341	0.303	0.114	0.282	0.881	-0.595
ext	0.333	0.294	0.110	0.274	0.880	-0.611
abs	0.278	0.233	0.078	0.220	0.872	-0.700
extabs	0.270	0.225	0.075	0.212	0.870	-0.737
artsum	**0.369**	**0.334**	**0.132**	**0.310**	**0.884**	**-0.564**

Table 6: Experiment Scores for CNN (NER - BART)

id	Bleu	Rouge 1 (f)	Rouge 2 (f)	Rouge l (f)	Bert Score (f1)	BleuRT
1s	0.133	0.142	0.042	0.127	0.834	-0.811
3s	0.285	0.260	0.102	0.245	0.876	-0.786
text	0.319	0.287	0.109	0.271	0.882	-0.720
ext	0.304	0.275	0.104	0.260	0.878	-0.747
abs	0.263	0.241	0.087	0.228	0.871	-0.828
extabs	0.258	0.223	0.078	0.211	0.870	-0.838
artsum	**0.349**	**0.322**	**0.129**	**0.303**	**0.885**	**-0.672**

Table 7: Experiment Scores for CNN (NER – T5)

Our best performing model for COVID news is T5 full-full, vs. the best performing for CNN news is BART (**artsum)**. COVID news dataset contains a more homogeneous set of articles that are on a specific topic and just from a 4-month period, while CNN news articles range across an unspecified number of topics over a ten-year period. Note that there would be a higher variation in the use of language as well as the context in CNN when compared to COVID news (due to topic specificity as well as temporality). Based on this, our inference is that a multistage summarization and BART work better for general use-cases while T5 trains well for specific topics better.

Comparing scores across COVID and CNN news headline results, we infer that the approach of custom training topic-specific models is more effective than having more complex stacked models over data across multiple topics. According to the results BART - based scenarios gave slightly better scores compared to T5-base, possibly because BART is trained as a denoising autoencoder dedicated to summarization.

We used multiple scoring mechanisms to get a better perspective of our results. Lower Rouge scores indicated not such a strong similarity between candidates and references. Bleu measured the position-independent matches using precision and had lower scores. Higher Bertscore provided better similarity for each token in the candidate sentence with each token in the reference sentence. Negative Bleurt scores aren't typically desirable and indicate fewer semantic similarities. The best performing experiment had the highest Bleurt score.

5 Conclusion

In this project, we have conducted 38+ experiments with COVID and CNN news articles. The test results show that in the case of models trained on specific news topics (as in COVID), the best scores are achieved when the headline text is generated from the full article text without any pre-summarization. Here we fail to reject our null hypothesis. For models trained using articles across multiple topics (as in CNN news), the best scores for headline generation were possible through multiple stages of interim summarization using abstractive and/or extractive methods and then using a final headline summarization model. In this case, summarization using Article Summary (**artsum**) yielded best scores. Manual verification of sampled predictions was reasonable in most cases as seen from the sample (**Fig. 6** & **Fig. 9**) provided in Appendix. Finally, our experiments indicate that there is a no unanimous winner among various transformers (BART and T5) - the use case, model development and stacking methods define how successful a transformer model implementation is.

6 Future Work

Our algorithm uses NER and pre-determined weights to filter sentences within an article before feeding it into the BART model for the headline summarization task. We would like to explore an entity-aware or topic-aware headline summarization where the weights for more important entities are optimized within the NLP training to make the model prioritize its headline summary based on the most important entities. One way to implement this could be by manipulating the attention mechanism to assign more weights to important entities.

Acknowledgments

We would like to thank Prof. Mark Butler and Prof. Joachim Rahmfeld for their valuable guidance and feedback on the project.

References

Yang Liu, Mirella Lapata. 2019. *Text Summarization with Pretrained Encoders, volume 1*. University of Edinburgh, Scotland, arXiv:1908.08345.

Abigail See, Peter J. Liu, Christopher D. Manning. 2017. *Get To The Point: Summarization with Pointer-*

Generator Networks, Computation, and Language. arXiv:1704.04368.

Hang Jiang, Yining Hua, Doug Beeferman, Deb Roy. 2022. *Annotating the Tweebank Corpus on Named Entity Recognition and Building NLP Models for Social Media Analysis*. arXiv:2201.07281.

Yen-Chun Chen, Mohit Bansal. 2018. *Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting*. arXiv:1805.11080.

Štěpán Műller, 2020. *Text Summarization Using Named Entity Recognition*. https://dspace.cvut.cz/bitstream/handle/10467/87671/F3-BP-2020-Muller-Stepan-text_summarization_using_named_entity_recognition.pdf.

Simple Transformers documentations available at: *https://simpletransformers.ai/docs/seq2seq-model/#configuring-a-seq2seqmodel*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, Peter J. Liu, 2020. *PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization*. arXiv:1912.08777

Transformers BART Model Explained for Text Summarization. *https://www.projectpro.io/article/transformers-bart-model-explained/553*.

Xuan Wang, Xiangchen Song, Bangzheng Li, Yingjun Guan, Jiawei Han. 2020. *Comprehensive Named Entity Recognition on CORD-19 with Distant or Weak Supervision*. arXiv:2003.12218.

COVID news Dataset. *https://ieee-dataport.org/open-access/free-dataset-newsmessage-boardsblogs-about-coronavirus-4-month-data-52m-posts*

Appendix

Figure 4 shows a sample COVID news article from the dataset prior to any pre-processing, highlighting the *title* and *text* elements that were used in this project.

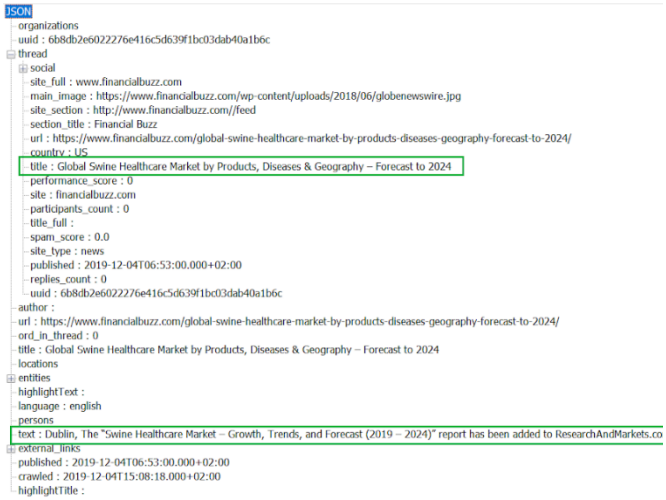

Figure 4: Sample article from COVID 19 dataset

A sample CNN news article prior to pre-processing is shown below, with the *title* and *text* highlighted.

Figure 5: Sample article from the CNN news dataset

Figure 7 shows a sample CNN news article along with the first sentence, the first three sentences, and summaries provided by extractive, abstractive, extractive-abstractive, and **artsum** methods.

Figure 6 shows the headline summarization using different approaches outlined under Section 4.2 for the CNN news dataset. The article is talking about the impact of big data. The term "big data" has been frequently used throughout the article. That is why the term is part of the headline generated by the extractive, abstractive, and extractive-abstractive methods. However, the actual human-created headline-only contains the term "data."

Our NER model was trained very well on the labeled COVID articles dataset. It gives about 85% accuracy with a 68.8% F1 score. The distribution of Named Entities for a sample of 100k records for the COVID 19 dataset is shown in **Figure 6** in Appendix. **Figure 9** shows a sample result of the NER model for predicting the named entities of the sentence: ["Coronaviruses are a group of related RNA viruses like sars that cause diseases in mammals and birds"].

Approach	Headlines
headlines	Moneyball How businesses using data outsmart rivals
predict_text	Big data revolution coming business
predict_ext	Big data revolution How retailers track customers future
predict_art	Big data new oil
predict_3s	Billy Beane The man changed baseball
predict_abs	Big data new oil Big data new energy
predict_extabs	Big data revolution The rise Amazon Netflix Amazon
predict_1s	Editor note Smart Business explores ways companies thinking smart thrive digitized world .

Figure 6: Sample headlines generated using the approaches discussed for the CNN news dataset.

```
[[{'Coronaviruses': 'B-CORONAVIRUS'},
  {'are': 'O'},
  {'a': 'O'},
  {'group': 'O'},
  {'of': 'O'},
  {'related': 'O'},
  {'RNA': 'B-CHEMICAL'},
  {'viruses': 'O'},
  {'like': 'O'},
  {'sars': 'B-CORONAVIRUS'},
  {'that': 'O'},
  {'cause': 'O'},
  {'diseases': 'O'},
  {'in': 'O'},
  {'mammals': 'B-EUKARYOTE'},
  {'and': 'O'},
  {'birds': 'B-LIVESTOCK'}]]
```

Figure 7: Sample NEs generated using our NER model.

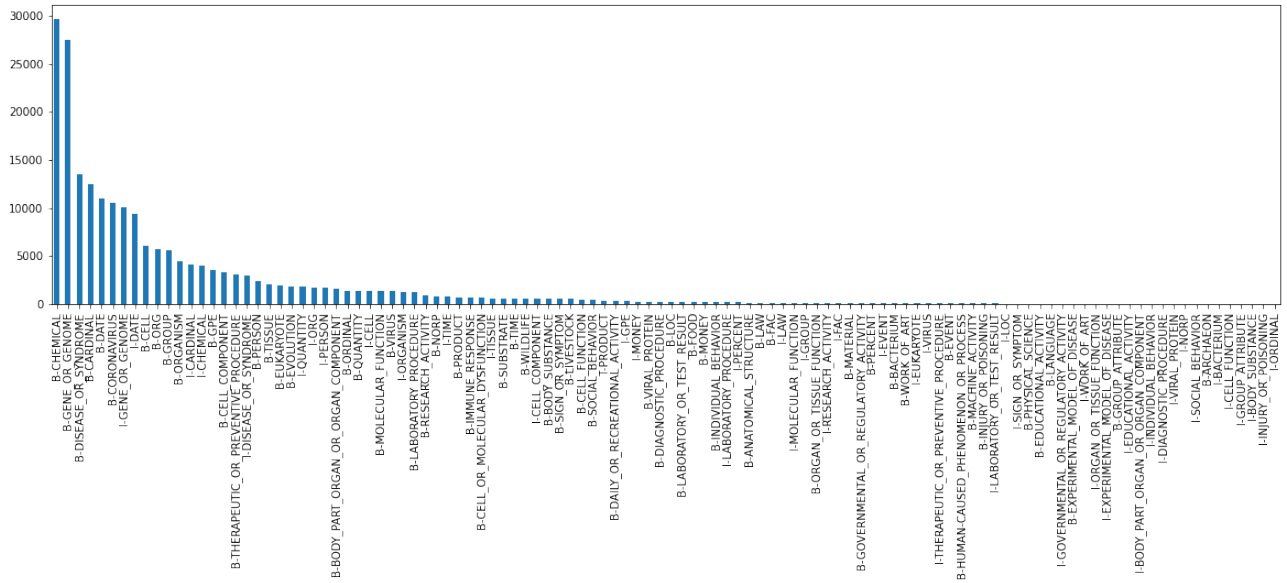

Figure 8: The distribution of Named Entities for a sample of 100k records for the COVID 19 dataset.

Summary	Content
text	Editor note Smart Business explores ways companies thinking smart thrive digitized world . When Billy Beane subject 2004 book Moneyball took general manager Oakland Athletics late 1990s revolutionized way baseball teams run . At time managers scouts relied experience identify promising new players Beane successfully used abermetrics statistical analysis baseball see value players teams passed over turning Athletics team capable competing biggest names sport . Now industry watchers say similar statistics revolution going business world . Technological advances giving rise huge amounts data consumers supply chains world events businesses use make better decisions gain competitive edge . It called big data analytics ignore warned risk left behind . Watch Funding 80 startups The rise big data Big data new oil says Andy Cutler director strategy SAS consulting firm specializing big data analytics . The folks going get good value going able refine turn useful products . Firms like Amazon Netflix forefront revolution gathering huge amounts data generated customers analyzing predict customers want buy watch future . This goes beyond personalized purchase suggestions Amazon currently working technology deliver products even ordered them least keep warehouses nearby anticipation says Phil Simon author Too Big Ignore The Business Case Big Data . Netflix effectively built entire business model analyzing customer data says . Read Groceries sent right trunk car Netflix track every view every click attempt understand customers want . The development smartphone technologies nearfield communication means analyzing customer data longer limited online world . Tailored buying experience Arne Strauss analytics professor Warwick Business School says retailers developing ways monitoring customers enter physical stores allowing optimize store layout even change onshelf promotions depending customer walking by . But predicting customer behavior one application . Major banks including HSBC use big data monitor predict fraud cardholders staff setting datamining systems collect patterns look anomalies . The investment bank Morgan Stanley uses statistical models measure impact market events bank real time . Improving efficiency large logistical operations another use . Read 7 insider tips new startup cities Delivery firm UPS spends 1bn year gathering data fleet trucks ensure efficient delivery routes . Meanwhile major supermarkets UK turning big data analytics help provide sameday grocery delivery services huge logistical challenge involving predicting customers likely want order it ensuring coordinated delivery times protect supermarkets alreadythin profit margins . Consultants academics say businesses jump big data bandwagon risk falling behind . But challenges . Too much information Some industry fear backlash consumers uncomfortable amount individual data gathered them stifling development areas . The speed technological development also created skills gap . Professor Thierry Chaussalet runs business intelligence analytics masters course University Westminster London says many businesses simply understand big data technology available offer . Stephen Mills associate partner big data analytics IBM agrees needs culture change . The technology easy bit says . The hard part change culture business processes make use new source data . The books changed world best business brains Technology gets makeover fashion goes futuristic How master Mobile advertising
sentence_1s	Editor note Smart Business explores ways companies thinking smart thrive digitized world .
sentence_3s	Editor note Smart Business explores ways companies thinking smart thrive digitized world . When Billy Beane subject 2004 book Moneyball took general manager Oakland Athletics late 1990s revolutionized way baseball teams run . At time managers scouts relied experience identify promising new players Beane successfully used abermetrics statistical analysis baseball see value players teams passed over turning Athletics team capable competing biggest names sport .
summary_ext	Editor note Smart Business explores ways companies thinking smart thrive digitized world . When Billy Beane subject 2004 book Moneyball took general manager Oakland Athletics late 1990s revolutionized way baseball teams run . Technological advances giving rise huge amounts data consumers supply chains world events businesses use make better decisions gain competitive edge . It called big data analytics ignore warned risk left behind . Firms like Amazon Netflix forefront revolution gathering huge amounts data generated customers analyzing predict customers want buy watch future . Tailored buying experience Arne Strauss analytics professor Warwick Business School says retailers developing ways monitoring customers enter physical stores allowing optimize store layout even change onshelf promotions depending customer walking by . The investment bank Morgan Stanley uses statistical models measure impact market events bank real time . Improving efficiency large logistical operations another use . Meanwhile major supermarkets UK turning big data analytics help provide sameday grocery delivery services huge logistical challenge involving predicting customers likely want order it ensuring coordinated delivery times protect supermarkets alreadythin profit margins . Stephen Mills associate partner big data analytics IBM agrees needs culture change .
summary_abs	Smart Business explores ways companies thinking smart thrive digitized world. Technological advances giving rise huge amounts data consumers supply chains world events businesses use make better decisions gain competitive edge. It called big data analytics ignore warned risk left behind . Watch Funding 80 startups The rise big data Big data new oil.
summary_extabs	Smart Business explores ways companies thinking smart thrive digitized world. Big data analytics ignore warned risk left behind . Firms like Amazon Netflix forefront revolution gathering huge amounts of data.
summary_art	It called big data analytics ignore warned risk left behind . Watch Funding 80 startups onceThe rise big dataBig data new oil says Andy Cutler director strategy SAS consulting firm specializing big data analytics . The development smartphone technologies nearfield communication means analyzing customer data longer limited online world . Consultants academics say businesses jump big data bandwagon risk falling behind . Stephen Mills associate partner big data analytics IBM agrees needs culture change .

Figure 9: A sample CNN news article followed by a summarization using the approaches discussed for the CNN news dataset.