Comparitive Study on Abstractive Text Summarization Using Transformer Models

Anuj Rithalia
Computer Engineering,
Delhi Technological University
Delhi, India
anujrithalia_co20b10_02@dtu.ac.in

Harsh Arya
Computer Engineering,
Delhi Technological University
Delhi, India
harsharya_co20b4_08@dtu.ac.in

Abstract—**In recent years, there has been a huge increase of data from a wide variety of different sources. This volume of text is a priceless source of knowledge and information that must be skillfully distilled in order to be of use. It can be challenging to find hidden patterns in this vast dataset of different organized or unstructured data in order to gain insights that can be put to use. In order to learn anything important in a reasonable amount of time, the material must be summarized. Everyone wants things that are ready to use in today's hectic world, and this is especially true when it comes to the Internet era. Therefore, the idea of TEXT SUMMARIZATION receives the greatest attention in order to promote empowerment in the realm of computers. Text summary has emerged as a crucial and timely technique for helping and analyzing text information in light of the volume of text material readily available on the Internet. This research represents a comparative study of different transformers models on different datasets to achieve abstractive summarization of text. The various state-of-the-art models Bert (having 12-layer, 768-hidden, 12-heads, 110M parameters), Bart (having 12-layer, 768-hidden, 16-heads, 139M parameters) and Long former (having 12-layer, 768-hidden, 12-heads, ~149M parameters) is applied on datasets like Amazon Food Review (~5,00,000 reviews), CNN Daily Mail (~3,00,000 articles) and Reddit ADHD (~2,00,000 comments). The performance of models is calculated with the help of ROUGE score and comparison among ROUGE values to suggest the best model to carry out abstractive text summarization work. Different abstractive summarization systems are explored and finally some research issues that will be addressed as future work are also identified.**

Keywords—Abstractive text summarization, attention mechanism, bidirectional encoder representation from transformers(BERT), neural networks, transfer models, CNN.

I. Introduction

In the ever-evolving landscape of abstractive text summarization, traditional methods heavily relied on statistical techniques, employing heuristics and dictionaries for word substitution. The advent of deep learning marked a paradigm shift in this field, with pioneering contributions such as the Recurrent Neural Network Encoder-Decoder (RNN) by Nallapati et al. in 2015 [5] [6][. This architecture introduced a novel approach to abstractive summarization, addressing concerns with Sequence to Sequence models [7].

The Lead-3 model was selected as the foundation, despite its drawback of requiring an extensive training dataset. Following this, attention-based summarization (ABS), as proposed by Bahdanau, Cho, and Bengio, revolutionized the way models handle lengthy input strings, utilizing weighted attention mechanisms to focus on essential words.

The Bidirectional encoder-decoder model, known as BiSum, was introduced by X. Wan et al., providing a balanced approach by considering both past and future information. The Pointer-Generator Networks with Coverage Mechanism addressed the issue of repetition in sequence-to-sequence models, ensuring accurate reproduction of information while minimizing redundancy.

The Double Attention Pointer Network, a paradigm introduced by Xhixin Li et al., leveraged a self-focus mechanism to gather key information, enhancing the overall accuracy and logic of summaries[4].Subsequently, the Transformer Architecture, proposed by Vaswani et al., brought forth a revolutionary design devoid of convolutions, relying solely on attention mechanisms. The inherent parallelization capability of the Transformer led to state-of-the-art performance compared to traditional encoder-decoder models.

BERT, or Bidirectional Encoder Representations and Transformers, developed by Devlin et al., presented a breakthrough by pre-training deep bidirectional representations from untagged text [1]. BERT's ability to understand language context from both directions allowed for a more comprehensive understanding of textual nuances[2].

The PEGASUS model, focused on abstractive summarization pretraining with extracted gap-sentences, introduced two key objectives: gap sentence generation and masked language modelling [8]. This dual approach enhanced the model's ability to generate concise and informative summaries.

BART, an acronym for Bidirectional and Auto-Regressive Transformers, employed a novel in-filling approach for machine translation, restoring text spans with a single mask token. Finally, T5, or Text-To-Text Transfer Transformer, implemented a unified framework for various tasks, using relative scalar embeddings to achieve goals such as language modelling, BERT-style tasks, and reshuffling[3].

II. Methodology

In our research, we undertook a comprehensive evaluation of Our research methodology encompassed a meticulous comparison of three transformer models—BERT, BART, and Longformer—across three distinct datasets: Reddit ADHD, Amazon Food Reviews, and CNN Daily Mail, with the aim of conducting abstractive summarization of textual data. To ensure the robustness of our findings, we adopted a systematic approach to data collection and preprocessing. This involved selecting datasets based on their relevance, size, and representativeness in the domain of text summarization, followed by standard preprocessing techniques such as tokenization, lowercasing, and removal of special characters and stopwords.

The following explains our proposed methodical approach towards abstractive summarization:

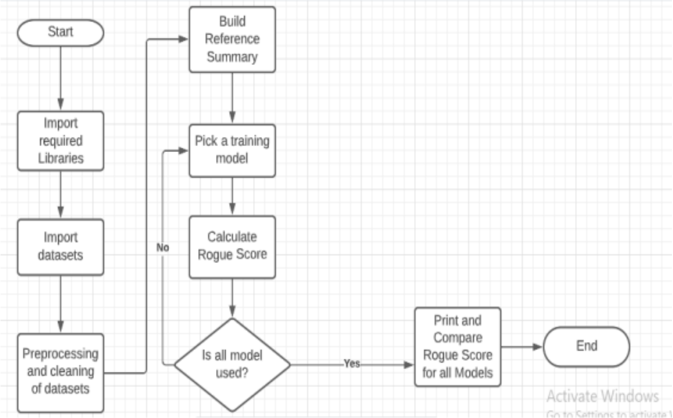


Fig.1. Showcasing adopted methodology in the project

In our evaluation setup, we employed a combination of quantitative and qualitative metrics to assess the performance of the transformer models. Specifically, we utilized evaluation metrics such as ROUGE scores to quantitatively measure the quality of the generated summaries. Additionally, we conducted human evaluation studies to qualitatively assess the readability and coherence of the summaries produced by each model. Throughout the experimentation phase, we fine-tuned the transformer models on the selected datasets and optimized hyperparameters to maximize summarization performance. Cross-validation techniques were also applied to ensure the reliability and generalizability of our results.

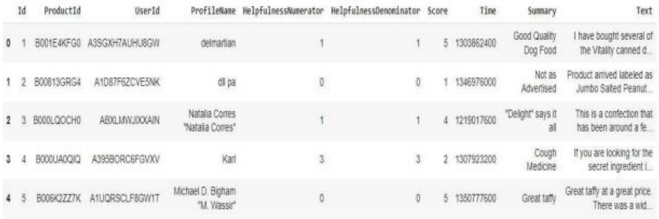
Our methodology also included a thorough analysis of performance variations across different datasets, allowing us to gauge the adaptability and generalization capabilities of the transformer models. We interpreted the results through statistical analysis and discussions, providing insights into the strengths and limitations of each model in summarizing diverse textual content. Moreover, ethical considerations such as data privacy, bias mitigation, and transparency were carefully addressed throughout the research process to uphold ethical standards and ensure the integrity of our findings.

In conclusion, our research methodology facilitated a comprehensive evaluation of transformer models for abstractive text summarization, yielding valuable insights that contribute to the advancement of natural language processing research. By systematically comparing the performance of BERT, BART, and Longformer across multiple datasets, we have provided a nuanced understanding of their efficacy in real-world summarization tasks, thereby informing future developments in this field.

A. *Amazon food reviews*

Amazon Customer Reviews, commonly referred to as Product Reviews, stand as one of Amazon's most prominent features. Originating in 1995, spanning nearly two decades, and accumulating over 100 million reviews from consumers worldwide, this repository serves as a rich source of insights into customer sentiments and product experiences. Recognizing the value of this extensive dataset, we aim to leverage it to advance research across various academic domains, particularly in understanding consumer perceptions of products. Our focus lies in utilizing Amazon reviews to develop a model capable of summarizing textual data effectively. Specifically, we intend to employ the review descriptions as input data and the review titles as target data for our model training. The dataset, conveniently available on platforms like Kaggle, offers a diverse range of reviews spanning four distinct product categories: books, DVDs, electronics, and kitchen/home goods, comprising varying sample sizes for each category.

This dataset encapsulates not only the customer review texts but also relevant metadata, facilitating comprehensive analysis and exploration. The metadata can be broadly categorized into two components. Firstly, it includes a compilation of comments posted on the Amazon.com marketplace from 1995 to 2015, accompanied by related metadata. This vast collection enables researchers to delve into the characteristics and evolution of customer reviews over time, offering insights into how consumers evaluate and share their experiences on a large scale. With over 130 million client evaluations, this segment provides a robust foundation for studying consumer behavior and preferences. Secondly, the dataset aims to facilitate cross-linguistic and cross-national analysis by incorporating reviews from various Amazon marketplaces and languages. With over 200,000 client evaluations spanning five nations, this subset allows for the exploration of not only product perceptions but also broader consumer preferences across linguistic and cultural contexts.

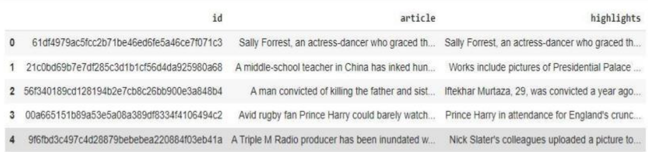


	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	1219017600	"Delight" says it all	This is a confection that has been around a fe...
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient i...
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	1350777600	Great taffy	Great taffy at a great price. There was a wid...

Fig.2. Showcasing review data before preprocessing

B. *CNN/DailyMail news text summmarization*

A dataset for text summarizing is CNN/Daily Mail. In order to test the system's ability to fill in the blanks, human-created abstractive summary points were created from news articles on the CNN and Daily Mail sites as queries (with one of the elements obscured). The stories were then used as the relevant passages. The scripts used to crawl, pull out, and create twins of excerpts and questions from these sites were made public by the authors. As per the scripts, the corpus has 11,487 test pairings, 13,368 validation pairs, and 286,817 training pairs altogether. The training set's source papers average 766 words and 29.74 sentences, while the summaries contain 53 words and 3.72 sentences. Highlight sentences and news items make up the data. The articles are used as the context in the question-answering setting of the data, and entities are gradually hidden in the highlight sentences to create Cloze-style questions, the objective of which is for the model to correctly guess which entity in the context has been concealed in the highlight. The highlight sentences are combined to create a summary of the article in the setting for summarizing. This dataset's goal is to aid in the creation of models that can succinctly describe lengthy text passages. This assignment helps display information effectively when there is a lot of text there. It should be made clear that any summaries obtained by models trained on this dataset are automatically created, though they represent the language used in the articles.

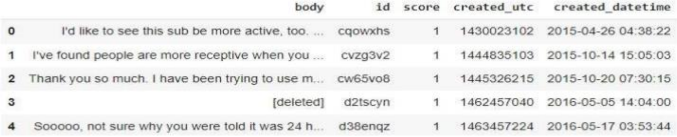


	id	article	highlights
0	61df4979ac5fcc2b71be46ed6fe5a46ce7f071c3	Sally Forrest, an actress-dancer who graced th...	Sally Forrest, an actress-dancer who graced th...
1	21c0bd69b7e7df285c3d1b1cf56d4da925980a68	A middle-school teacher in China has inked hun...	Works include pictures of Presidential Palace ...
2	56f340189cd128194b2e7cb8c26bb900e3a848b4	A man convicted of killing the father and sist...	Iftekhar Murtaza, 29, was convicted a year ago...
3	00a665151b89a53e5a08a389df8334f4106494c2	Avid rugby fan Prince Harry could barely watch...	Prince Harry in attendance for England's crunc...
4	9f6fbd3c497c4d28879bebebea220884f03eb41a	A Triple M Radio producer has been inundated w...	Nick Slater's colleagues uploaded a picture to...

Fig.3. Showcasing CNN data before preprocessing

C. Reddit ADHD dataset

A recently compilec eddit dataset called Reddit TIFU, where TIFU stands for the name of the /r/tifu subreddit. Overall, there are 122,933 text-summary pairs. We propose a novel dataset and a new technique to solve the issue of abstractive summarization. First, we gather the 120K posts that make up the Reddit TIFU dataset through the online debate forum Reddit. In contrast to existing datasets, which often utilize formal documents as source, like news stories, we use such informal crowd-produced posts as text source. Because significant lines are typically found towards the start of the context and positive conclusive candidates are present in the context in comparable types, our dataset may be less biassed as a result. In the second section, it is recommended a brand-new abstractive summarization model called multi-story memory networks (MMN), which has multi-level memory to store text data at various stores of abstraction. We demonstrate that the Reddit TIFU dataset is extremely abstractive and that the MMN counterparts the leading summarization algorithms through quantitative assessment and user studies conducted using Amazon Mechanical Turk. Deep neural networks have been utilized recently in automatic text summarization to produce high-quality abstractive summaries, although the effectiveness of these models heavily relies on ample training data.



	body	id	score	created_utc	created_datetime
0	I'd like to see this sub be more active, too. ...	cqowxhs	1	1430023102	2015-04-26 04:38:22
1	I've found people are more receptive when you ...	cvzg3v2	1	1444835103	2015-10-14 15:05:03
2	Thank you so much. I have been trying to use m...	cw65vo8	1	1445326215	2015-10-20 07:30:15
3	[deleted]	d2tscyn	1	1462457040	2016-05-05 14:04:00
4	Sooooo, not sure why you were told it was 24 h...	d38enqz	1	1463457224	2016-05-17 03:53:44

Fig.4. Showcasing reddit ADHD data before pre processing

D. Experimental Setup

Python's train test split method divides arrays or matrices into random subsets for the train and test sets of data, respectively. The parameter list includes *arrays, test size, train size, random state, shuffle, and stratify. It returns the list that contains train and test splits of test inputs.

My sister has a friend called John

Really, tell me more about him

She thinks he is so funny

Fig.5. Showing bob who is talking with his AI friend Alice

A pipeline addition for spaCy 2.1+ called NeuralCoref uses a neural network to annotate and resolve coreference clusters. NeuralCoref is adaptable to new training datasets, integrated into the spaCy NLP pipeline, and production-ready. Python/Cython is used to write NeuralCoref, which only has a pre-trained statistical model for English.

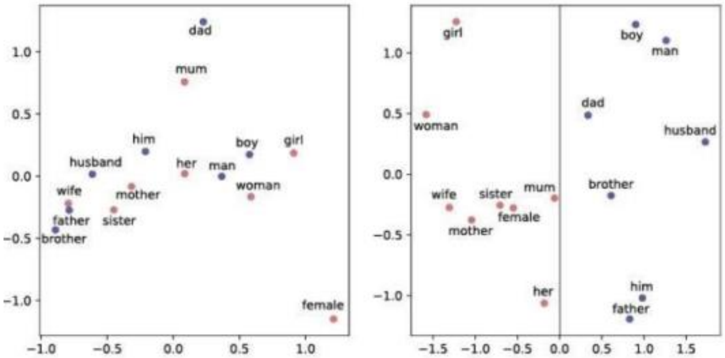


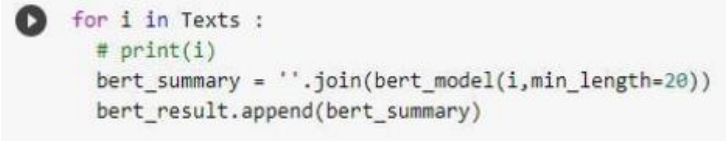
Fig.6. Showing left: initial word embeddings right: trained word embeddings

Coreference resolution is the process of connecting together mentions that refer to actual entities. Extract a list of words that are perhaps references to things in the real world. Calculate a set of features for each mention and each pair of mentions. The next step is to use this set of attributes to determine the most likely antecedent for each mention, assuming there is one. Pairwise ranking is the name of the final phase.

Used Sentencepiece (unsupervised text tokenizer and detokenizer) for tokenizing. We have used it so the vocabulary size is predetermined prior to neural network training. It is an API that offers the encoding, decoding, and training of sentencepiece.

E. Transfer Model

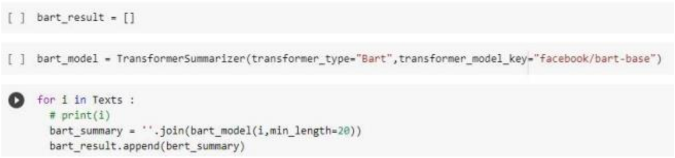
The Bert-base-uncased model consists of 12 layers, 768 hidden units, 12 attention heads, and approximately 110 million parameters. It has been pre-trained on a vast corpus of English data, allowing it to capture intricate language patterns and nuances. This model serves as a powerful tool for various natural language processing tasks, offering robust performance and adaptability.



```
for i in Texts :
  # print(i)
  bert_summary = ''.join(bert_model(i,min_length=20))
  bert_result.append(bert_summary)
```

Fig.7. Showcasing command of loading fine tuned and pretrained BERT model from HuggingFace

The Facebook Bart-base model is characterized by 12 layers, 768 hidden units, 16 attention heads, and around 139 million parameters. Similar to the Bert-base-uncased model, it is pretrained on English language data, enabling it to understand and generate coherent text. The Bart model is particularly adept at tasks such as text summarization and generation.



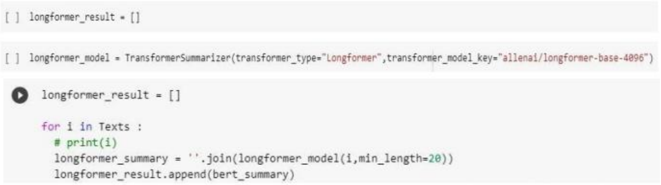
```
bart_result = []
```

```
bart_model = TransformerSummarizer(transformer_type="Bart",transformer_model_key="facebook/bart-base")
```

```
for i in Texts :
  # print(i)
  bart_summary = ''.join(bart_model(i,min_length=20))
  bart_result.append(bert_summary)
```

Fig.8. showcasing command of loading fine tuned and pretrained Bart model from HuggingFace

The Longformer - base - 4096 model features 12 layers, 768 hidden units, 12 attention heads, and approximately 149 million parameters. Designed specifically for handling large documents, this transformer model excels in scenarios where extensive context is crucial for accurate processing and understanding. Its capacity to efficiently process lengthy texts makes it a valuable asset for tasks involving comprehensive document analysis and comprehension.



```
longformer_result = []
```

```
longformer_model = TransformerSummarizer(transformer_type="Longformer",transformer_model_key="allenai/longformer-base-4096")
```

```
longformer_result = []

for i in Texts :
  # print(i)
  longformer_summary = ''.join(longformer_model(i,min_length=20))
  longformer_result.append(bert_summary)
```

Fig.9. showcasing command of loading of fine tuned and pretrained longformer model from HuggingFac

III. Results and discussion

To calculate the result in abstractive summarization, there is a need for an evaluation mechanism that can measure how well the summary is being constructed. Initially, a reference summary is required, against which the candidate summary (generated by fine-tuned models) can be compared.

ROUGE (Recall Oriented Understudy for Gisting Evaluation) calculates the similarity between the candidate document (C) and the reference document (R). It comprises several metrics including ROUGE-1, ROUGE-2, and ROUGE-L.

ROUGE-1 precision measures the ratio of unigrams that appear in C and R to the number of unigrams in C. ROUGE-1 recall measures the ratio of unigrams that appear in C and R to the number of unigrams in R. The F1-Score for ROUGE-1 is calculated using precision and recall.

ROUGE-2 precision calculates the ratio of 2-grams in C and R to the number of unigrams in C, while recall measures the ratio of 2-grams in C and R to the number of unigrams in R. The F1-Score for ROUGE-2 is obtained using precision and recall. The formula can b given as :

F1-Score=2*(Precision *Recall)/(Precision +Recall)

ROUGE-L calculates the longest common subsequence between the model output and reference, where the subsequence need not be consecutive. Precision, recall, and F1-Score for ROUGE-L are computed similarly to ROUGE-1 and ROUGE-2 but with n-grams.

ROGUE	BERT	BART	LONGFORMER
ROGUE-1	0.09511007821048226	0.01604696470140127	0.025897894597087093
ROGUE-2	0.027314732488950128	0.00076639693245325	0.000452652736770383
ROGUE-L	0.08603196592134225	0.01404157515872987	0.02362351537612145

Fig.10. ROGUE value on amazon food review dataset

ROGUE	BERT	BART	LONGFORMER
ROGUE-1	0.08988495153522745	0.09017589592208931	0.03184596273105984
ROGUE-2	0.053620504690352046	0.05365830153168063	0.00515935748621485
ROGUE-L	0.07271776047600165	0.07280144694250731	0.048965231470384517

Fig.11. ROGUE value on CNN/daily mail dataset

ROGUE	BERT	BART	LONGFORMER
ROGUE-1	0.09654871235478977	0.0258974656321478	0.04741258963214745
ROGUE-2	0.047891452368745129	0.0003459771239415	0.00093698521475648
ROGUE-L	0.08945721589321478	0.0095874123657412	0.032587416985741232

Fig.12. ROGUE value on reddit ADHD dataset

The above figures depict the ROUGE values for different datasets upon applying various models including BERT, BART, and Longformer. Observations reveal performance variations across datasets and models. For instance, BERT outperformed BART and Longformer on the Amazon Food Review Dataset, while BART excelled on the CNN/DailyMail Dataset.

The analysis from reviewed literature reveals several key findings related to extractive and abstractive summarization methods. Extractive summarization faces challenges such as including unnecessary information, missing important details, and redundancy. Abstractive summarization, on the other hand, struggles with coherence, semantic connections, and linguistic depth.

Understanding the semantics of the text and applying natural language generation principles are crucial for developing effective abstractive summaries. Additionally, the balance between inclusiveness and coherence remains a challenge in abstractive summarization.

IV. Conclusion

There is now an enormous amount of information available due to the Internet's rapid expansion. Humans find it challenging to summarize lengthy texts.Thus, in this era of information overload, there is a great need for automatic summarizing tools and technologies. Therefore, a system that allows a user to quickly find and obtain a summary document is needed. Extractive text summarization is easy to carry out as it just extracts important keywords or sentences and rearranges to provide the summary. However, text summarization using abstractive techniques is more effective because they result in a summary that is difficult to generate but semantically linked. Abstractive summarization techniques yield extremely coherent, cohesive, information-rich, and minimally redundant summaries. However, the training data used in transformer models could be characterized as fairly neutral, Bert model can still have biased predictions. The Facebook/Bart-base model used in this project is mostly meant to be fine-tuned on supervised datasets. These conclude that there is a scope of further improvements in the form of training data or unseen data (Giga words and DUC 2004 test sets). Although it is not feasible to use and explain all transformer models on different datasets, we believe it offers a useful perspective on current developments and trends in automatic summarizing techniques and describes the state-of-the-art in this field of study.

REFERENCES

[1] Sanjan S Malagi, Rachana Radhakrishnan, Monisha R, Keerthana S, Dr. D V Ashoka, 2020, An Overview of Automatic Text Summarization Techniques, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) NCAIT – 2020 (Volume 8 – Issue 15),

[2] T. Ma, Q. Pan, H. Rong, Y. Qian, Y. Tian and N. Al-Nabhan, "T-BERTSum: TopicAware Text Summarization Based on BERT," in IEEE Transactions on Computational Social Systems, vol. 9, no. 3, pp. 879-890, June 2022, doi: 10.1109/TCSS.2021.3088506.

[3] M. Lewis et al., “BART: Denoising Sequence-toSequence Pre-training for Natural Language Generation, Translation, and Comprehension,” arXiv:1910.13461v1.

[4] Z. Li, Z. Peng, S. Tang, C. Zhang, and H. Ma, “Text Summarization Method Based on Double Attention Pointer Network,” IEEE Access, vol. 8, pp. 11279- 11288,Jan. 2020.

[5] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and BBing Xiang, “Abstractive text summarization using Sequenceto-sequence RNNs amd beyond”, arXiv:1602.06023v5, Aug 2016.

[6] Sumit Chopra, Michel Auli, and Alexander M. Rush, “Abstractive sentence summarization with attentive recurret neural networks”, NAACL-HLT, 2016.

[7] R. Nallapati, B. Zhou, C. dos Santos, C. glar Gulcehre, and B. Xiang, “Abstractive text summarization using sequence-to-sequence RNNs and beyond,” arXiv preprint arXiv:1602.06023, 2016.

[8] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, “PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization,” arXiv:1912.08777v3.

[9] D. Suleiman and A. A. Awajan, “Deep Learning Based Abstractive Text Summarization: Approaches, Datasets, Evaluation Measures, and Challenges,” Mathematical Problems in Engineering, Hindawi, vol. 2020, pp. 1-29,Aug. 2020.

[10] T. Shi, Y. Keneshloo, N. Ramakrishnan, and C. K. Reddy, Neural “Abstractive Text Summarization with Sequence-ToSequence Models: A Survey”, http://arxiv.org/abs/1812.02303, 2020.