International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

# ADVANCMENTS IN TEXT SUMMARIZATION AND EXTRACTIVE QUESTION-ANSWERING : A MACHINE LEARNING APPROACH

DR D. Eswara Chaitanya M. Tech., Ph.D.
Associate Professor
dept of Electronics and Communication Engineering
RVR & JC College Of Engineering, Guntur
deswarachaitanya@rvrjc.ac.in

Paridala Sai Lokesh (Y20EC0147)
dept of Electronics and Communication Engineering
RVR & JC College Of Engineering, Guntur
sailokeshparidala03@gmail.com

Nandigam Mounika (Y20EC135)
dept of Electronics and Communication Engineering
RVR & JC College Of Engineering, Guntur
mounikanandigam21@gmail.com

Vudumula Gopichand (Y20EC187)
dept of Electronics and Communication Engineering
RVR & JC College Of Engineering, Guntur
vudumulagopichand18@gmail.com

Paritala Bharath Kumar (Y20EC149)
dept of Electronics and Communication Engineering
RVR & JC College Of Engineering, Guntur
bharathchowdary11s18@gmail.com

***Abstract*— In the era of social media platforms, the rapid expansion of data mining in the fields of information retrieval and natural language processing emphasizes the crucial need for automated text summarization. At the current time, pretrained word embedding techniques and sequence to sequence models can be effectively repurposed in the realm of social network summarization to efficiently condense significant information with strong encoding capabilities. However, dealing with the challenge of extended text dependency and efficiently utilizing latent topic mapping presents an increasingly significant obstacle for these computational models. In this document, we propose a topic-focused approach for both extractive and abstractive summarization, integrating Question Answering features by employing BERT and Pegasus pretrained models.**

**A comprehensive analysis of the architectural complexities of the pre-existing models utilized in the creation of our model has been presented. The evaluation of our Text Summarization & Question Answering Model was conducted meticulously on a variety of datasets including Multi news, XSum, and CNN/Daily Mail, with experimental results indicating its achievement of state-of-the-art performance levels based on ROGUE scores. Following this, the verification of our results through human evaluation confirms that the summaries generated by our model are in alignment with the performance standards set by humans across various datasets.**

**Furthermore, we have integrated an Extractive Question Answering task into our model, proposing an architecture leveraging BERT and conducting comparative analyses against alternative language models to gauge its efficacy.**

**Keywords—Natural Language Processing (NLP), Text Summarization, Question & Answering (Q&A), Transformer Models, Pegasus, BERT (Bidirectional Encoder Representations from Transformers), Neural Topic Modeling, Extractive Summarization, Abstractive Summarization, Semantic Representation, Topic Embedding**

## I. INTRODUCTION

Research in text summarization spans various disciplines such as computer science, multimedia, and statistics, with a focus on extractive and abstractive techniques. This process has been heavily analysed across different sectors like computer science, multimedia, and statistics. Presently, there exist two primary summarization methodologies: extractive and abstractive. The extractive approach involves selecting important sentences or rearranging them from the original text, while the abstractive method focuses on creating new words or phrases to convey the main idea. The majority of current methodologies focus on encoding paragraphs and subsequently decoding them using various mechanisms, albeit resulting in significant loss of information.

The advantages of word embedding are limited by small, specific datasets, requiring more extensive contextual information. The future direction of summarization aims to predict words from comprehensive contexts through language modelling and representation learning. Hence, this document tackles a demanding research issue: how to employ preexisting language models for text representation and creation.

Both BERT and Pegasus are notable models in the field of text summarization and question answering tasks. BERT (Bidirectional Encoder Representations from Transformers) has completely altered the dynamics in natural language understanding through the pretraining of a bidirectional language model. It efficiently captures contextual information, facilitating superior performance in various NLP tasks, such as text summarization and question answering.

On the contrary, Pegasus is tailored specifically for abstractive text summarization. In contrast to BERT, which functions bidirectionally, Pegasus employs an encoder-decoder architecture akin to customary sequence-to-sequence models but with significant enhancements. Pegasus generates summaries by considering the entire input document without relying on explicit sentence extraction. This methodology enables Pegasus to create coherent and informative summaries, showcasing its proficiency in handling summarization tasks.

In terms of question answering, BERT has demonstrated remarkable performance, especially in tasks where context plays a crucial role. By leveraging its bidirectional nature and fine-tuning on QA datasets, BERT can effectively understand and respond to questions based on the provided context.

In this investigation, it is proposed to refine and investigate the PEGASUS and BERT models for abstractive and extractive text summarization, in addition to the Question Answering task, utilizing a varied dataset. The primary aim of this inquiry is to assess the efficacy of the transformer-based model on this diverse dataset and to pinpoint any prospective enhancements that could be implemented to enhance its capabilities in summarization and question answering. The subsequent sections of this manuscript will refer to our Text summarization and Question Answering model as QuickViewInsight. The organizational framework of the document is delineated as follows: initially, an elucidation of the foundational technologies employed in the development of this model will be provided. Subsequently, an exposition on the architectures of the pre-trained models utilized in this QuickViewInsight model will be presented. The subsequent sections will elaborate on the datasets utilized for training our model and the experimental configuration, culminating in a summary of the principal contributions and references in the concluding section.

## II. Background Work

The progress of Abstractive and Extractive Text Summarization, in conjunction with Question-Answering systems, originates from comprehensive background investigation and preparatory efforts aimed at integrating state-of-the-art technologies and methodologies. This segment provides a summary of the fundamental components and theoretical structures that influence the development and implementation of the system.

### A. *Natural Language Processing Techniques*

Natural Language Processing (NLP) serves as the essential foundation of current language comprehension frameworks, empowering computers to grasp and handle human language. The methods of NLP encompass a broad array of strategies, like syntactic and semantic analysis, tokenization, and language modeling.

### B. *Transformer Architecture*

The transformer architecture denotes a pivotal transformation in the area of Natural Language Processing (NLP), enabling the generation of remarkably scalable and productive models for tasks related to sequences. With a distinct feature of self-attention mechanisms, transformers showcase remarkable ability in capturing extensive dependencies across sequential data, rendering them especially well-suited for tasks such as text summarization and question-answering.

### C. *Tokenization*

Tokenization is an essential procedure within the domain of Natural Language Processing (NLP), which entails the division of textual data into distinct units known as tokens. These tokens play a vital role as the foundational components for further analysis by NLP algorithms, enabling functions like semantic interpretation, sequence annotation, and text synthesis.

### D. *Attention Mechanism*

The mechanism of attention, a fundamental element within architectures based on transformers, empowers models to deliberately focus on significant segments of input sequences while conducting computations. Through the allocation of weights to individual tokens according to their significance, attention mechanisms empower models to concentrate on noteworthy data and filter out unimportant context, consequently boosting performance across a range of natural language processing tasks.

Additional Technologies:

### E. *Streamlit Library*

Streamlit is a widely renowned Python library acclaimed for its straightforward nature and user-friendly interface when developing interactive web applications. By leveraging Streamlit, researchers can seamlessly deploy machine learning models and data analysis workflows, enhancing user-friendly interaction and real-time feedback.

### F. *Python*

Python acts as the predominant language in contemporary data science and machine learning, providing a diverse array of libraries and frameworks to construct advanced NLP applications. Due to its adaptability and clarity, Python enables quick prototyping and testing, rendering it essential in various research and development projects.

## III. Proposed Model

QuickViewInsight is an advanced platform for text processing that aims to enhance the extraction and comprehension of information efficiently. By exploiting advanced natural language processing (NLP) techniques, QuickViewInsight delivers efficient solutions for tasks such as text summarization and question answering. Based on state-of-the-art pre-trained models like Pegasus and BERT, QuickViewInsight represents the latest progress in NLP research, allowing users to create concise and coherent summaries of input documents while preserving essential information.

In the segment dedicated to the proposed model, we delve into the intricate design and functionalities of QuickViewInsight, elucidating the underlying mechanisms of its extractive and abstractive

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

summarization modes, along with its question-answering capabilities. Through a thorough examination of its components and algorithms, our goal is to offer a comprehensive insight into the operation of QuickViewInsight and its potential applications across various domains.

*A. Architectural Components*

QuickViewInsight is a sophisticated platform for text processing distinguished by its complex architectural configuration, carefully constructed to enable efficient summarization and examination of textual data. At its nucleus, QuickViewInsight consists of multiple interconnected elements, each fulfilling a unique yet complementary role in the text processing sequence.

The initial processing module functions as the entry point for textual data, meticulously preparing incoming documents through processes like tokenization, sentence segmentation, and encoding. This initial phase guarantees that the text is suitably organized and structured for subsequent scrutiny. Subsequent to the initial processing, the representation layer becomes prominent, utilizing cutting-edge language models such as Pegasus and BERT to capture detailed semantic intricacies embedded within the text. Through the utilization of pre-trained embeddings and contextual encodings, QuickViewInsight adeptly navigates the semantic landscape of textual data, establishing the groundwork for subsequent examination.

The topic induction module serves to enhance the text processing sequence by utilizing Neural Topic Models (NTMs) to uncover latent themes and topics inherent within the text corpus. By identifying patterns and connections among words, this module condenses intricate textual data into coherent thematic groupings, allowing users to acquire deeper insights into the fundamental content.

The core of QuickViewInsight resides in its summarization mechanism, providing both extractive and abstractive summarization capabilities. In the extractive mode, the mechanism utilizes methodologies like LSTM-based sentence classification to pinpoint and extract significant sentences from the input text. Conversely, the abstractive mode employs transformer-based structures complemented by gated networks to produce succinct and logical summaries by paraphrasing and restructuring textual content, effectively capturing the essence of the original text in a concise manner.

Furthermore, QuickViewInsight transcends basic summarization by offering robust question answering functionalities facilitated by fine-tuned BERT models. This component enables users to present queries and retrieve pertinent answers from the input text, easing interactive exploration and information retrieval.

Apart from its advanced processing capabilities, QuickViewInsight features a scalable infrastructure, tailored to accommodate diverse workloads and computational requirements, empowering users to fully exploit the potential of textual data analysis.

*B. PreProcessing and Representation*

The preprocessing phase functions as the fundamental stage in QuickViewInsight's text processing pipeline, where incoming textual data is meticulously prepared to ensure optimal analysis outcomes. By utilizing advanced tokenization methods, the preprocessing module dissects raw text into basic elements, enabling subsequent analysis at a detailed level. Sentence segmentation plays a crucial role in refining the preprocessing pipeline, dividing textual content into coherent units that capture the semantic essence of individual sentences. Encoding, a pivotal step in the preprocessing process, converts tokenized text into numerical representations, facilitating smooth integration with downstream analysis modules. QuickViewInsight ensures that textual data is appropriately structured and formatted for subsequent processing stages through the skilled orchestration of these preprocessing techniques.

Advancing beyond preprocessing, QuickViewInsight's representation layer embodies the core of its text processing capabilities, utilizing sophisticated language models to extract nuanced semantic insights from textual data. Built on cutting-edge models like Pegasus and BERT, the representation layer goes beyond traditional text analysis approaches, delving deeply into the semantic structure of textual content. By leveraging pre-trained embeddings and contextual encodings, QuickViewInsight effectively navigates the intricate nuances present in textual data, empowering users to uncover hidden patterns and derive actionable insights. By seamlessly integrating preprocessing and representation techniques, QuickViewInsight lays the foundation for comprehensive text analysis, enabling advanced summarization, topic modeling, and question answering functionalities.

*C. Utilization of PEGASUS Model*

The integration of PEGASUS within the QuickViewInsight platform represents a crucial architectural element aimed at enhancing the capabilities of text summarization. PEGASUS, recognized as a cutting-edge transformer-based model, has demonstrated remarkable proficiency in condensing extensive textual documents into concise and coherent summaries. Positioned as a foundational component within the architectural framework of QuickViewInsight, PEGASUS plays a vital role in facilitating the extraction of crucial insights from voluminous textual data with unparalleled efficiency.

The architectural arrangement of PEGASUS within QuickViewInsight showcases a sophisticated combination of transformer-based layers, intricately designed to navigate the complex semantic aspects inherent in textual material. Essentially, PEGASUS exploits self-attention mechanisms to facilitate a comprehensive understanding of textual relationships and context. By refining representations iteratively through multi-layered

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

transformer structures, PEGASUS effectively processes input sequences, capturing essential information while maintaining coherence and relevance. In its pursuit of achieving proficient abstractive summarization, PEGASUS employs various advanced methodologies and techniques. A key element of its effectiveness lies in the utilization of pre-trained transformer models, which exploit extensive textual corpora to extract fundamental patterns and linguistic subtleties. Furthermore, PEGASUS incorporates innovative attention mechanisms that allow for dynamic focus allocation to pertinent aspects of the input text. By dynamically attending to relevant content and utilizing contextual embeddings, PEGASUS produces summaries that encapsulate the core essence of the original text while adhering to linguistic norms and coherence.

Additionally, PEGASUS incorporates specialized techniques tailored specifically for summarization tasks within the QuickViewInsight environment. These methods cover a variety of methodologies, like masked language modeling, diverse output generation, and fine-tuning techniques to fulfill domain-specific needs. Through the integration of these inventive approaches, PEGASUS enables QuickViewInsight to provide abstractive summaries that go beyond mere content extraction, offering nuanced insights and actionable intelligence to users.

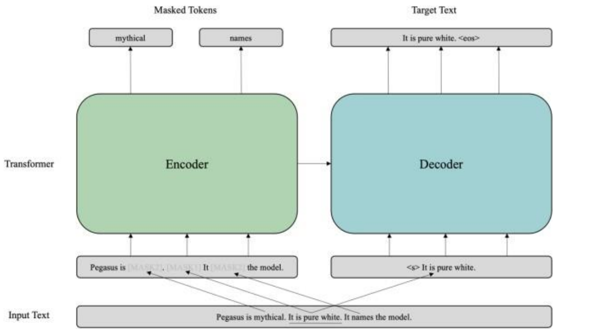

Figure 1: Overall architecture of the PEGASUS model

*D. Utilization of BERT Model*

In the midst of the extensive realm of natural language processing, BERT (Bidirectional Encoder Representations from Transformers) stands out as a pivotal model, known for its flexibility and efficacy in various text-related assignments. Situated within the structural framework of QuickViewInsight, BERT plays a crucial role in enabling question-answering tasks and enhancing the contextual comprehension of textual data.

Central to BERT is a bidirectional transformer design, recognized for its inability to grasp contextual specifics from both preceding and subsequent tokens in a sequence. This bidirectional processing capability equips BERT with a refined understanding of textual relationships and subtleties, allowing it to discern intricate semantic nuances and deduce underlying meanings from input text.

The integration of BERT in QuickViewInsight transcends simple token-level encoding, encompassing a range of advanced methodologies customized for question-answering tasks. By leveraging pre-trained BERT models fine-tuned on question-answering datasets, QuickViewInsight adeptly utilizes BERT's contextual embeddings to identify pertinent passages and extract accurate answers to user inquiries. Moreover, BERT's adaptability extends to various domains and tasks within QuickViewInsight. Through the refinement of pre-trained BERT models on domain-specific datasets, QuickViewInsight enhances BERT's capabilities, allowing it to recognize domain-specific nuances and provide tailored responses that meet users' specific needs.

Beyond its proficiency in question-answering tasks, BERT acts as a foundational element for diverse text summarization initiatives within QuickViewInsight. By incorporating BERT's contextual embeddings into summarization workflows, QuickViewInsight leverages BERT's comprehensive comprehension of textual context to distill crucial insights and produce concise summaries that capture the essence of lengthy documents. Essentially, BERT's inclusion in QuickViewInsight represents a merging of state-of-the-art natural language processing techniques, empowering the platform to offer unparalleled insights and facilitate seamless interaction with textual data. Through its multifaceted capabilities in question-answering and text summarization, BERT enhances QuickViewInsight's effectiveness, elevating it to new levels of sophistication and usefulness in the realm of information retrieval and analysis.

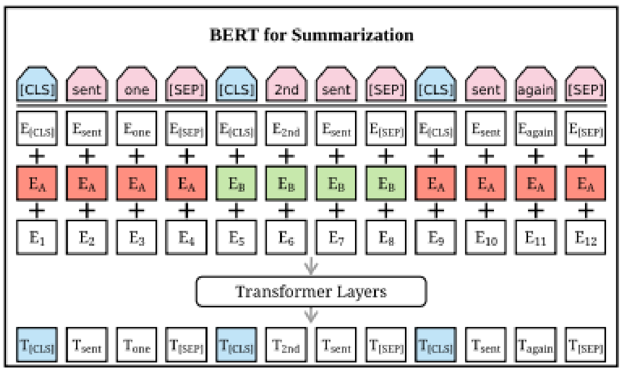

Figure 2: Architecture of the original BERT model.

*E. Integrations and Interactions*

The fusion of results from the Pegasus and BERT models is essential within the proposed framework of QuickViewInsight, enabling the platform to seamlessly incorporate advanced text summarization and question-answering capabilities. By carefully coordinating these pre-trained models, QuickViewInsight provides in-depth insights and enables natural interactions with textual data .

At the core of this integration is a sophisticated coordination mechanism that blends the outcomes of Pegasus and BERT models. Through capitalizing on Pegasus' proficiency in abstractive summarization and BERT's expertise in question-answering,

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

QuickViewInsight amalgamates these outcomes to provide users with comprehensive insights derived from textual data. By combining the condensed summaries from Pegasus with the precise answers from BERT, QuickViewInsight delivers a comprehensive view of textual content, empowering users to extract actionable insights effectively. Moreover, the synergy between the pre-trained models and other elements of QuickViewInsight, such as neural topic modeling and joint learning optimization, demonstrates a mutually beneficial interaction aimed at maximizing the platform's effectiveness. By intricately engaging with neural topic modeling, the outcomes from Pegasus and BERT models are placed in context within the broader thematic landscape, allowing QuickViewInsight to identify underlying topics and themes present in textual data. This contextual comprehension enhances the summarization and question-answering processes, improving the significance and consistency of the insights provided to users.

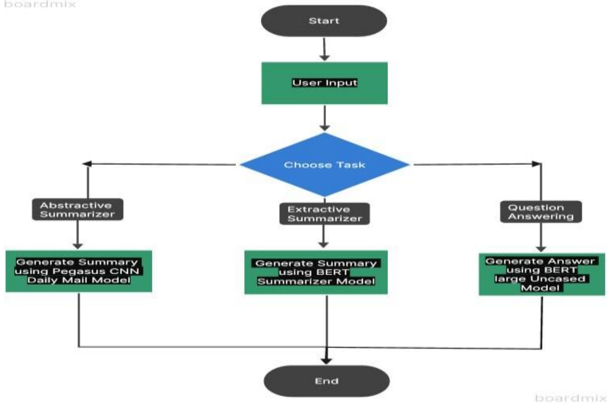


Figure 3: Flow chart of experimental setup used for model.

Additionally, the optimization techniques implemented in QuickViewInsight involve joint learning approaches that continuously enhance the relationships between pre-trained models and supporting elements. Through iterative adjustments to the parameters of Pegasus and BERT models alongside neural topic modeling and other components, QuickViewInsight ensures the smooth alignment of these diverse elements, cultivating a unified environment conducive to optimal performance. Essentially, the integrations and interactions within QuickViewInsight epitomize a comprehensive approach to text analysis, where the outcomes from pre-trained models are amalgamated and positioned within a broader context of thematic comprehension and optimization.

In *conclusion*, QuickViewInsight's significance is grounded in its capacity to democratize the retrieval of intricate textual information, enabling users from various fields to efficiently and accurately extract insights. The modular structure of QuickViewInsight allows for smooth integration into existing workflows, providing unmatched adaptability and scalability. Consequently, QuickViewInsight acts as a driver of innovation, facilitating well-informed decision-making and knowledge exploration within a range of industries.

## IV. Experiment and Results

### A. Datasets

For the purpose of downstream summarization, we exclusively utilized publicly available abstractive summarization datasets, obtaining access to them via the Hugging Face Summarization Datasets platform. This platform offers openly reproducible code for processing datasets and partitioning them into train, validation, and test sets. In cases where no predefined split was available, we adhered to an 80/10/10 train/validation/test ratio. Moreover, when a validation split was not present, we allocated 10% of the training set for validation purposes.

**XSum** (Narayan et al., 2018) is composed of 227k BBC articles spanning from 2010 to 2017, encompassing a diverse range of topics, accompanied by professionally crafted single-sentence summaries.

The **CNN/DailyMail** (Hermann et al., 2015) dataset comprises 93k articles from CNN and 220k articles from the Daily Mail, both of which include bullet point summaries. Our utilization aligns with the non-anonymized version referenced in See et al. (2017).

**Multi News** (Fabbri et al., 2019) serves as a multi-document summarization dataset, featuring 56k pairs of news articles and their summaries composed by human writers sourced from newser.com.

### B. Experiments

The adjustment of the model for the downstream task of text summarization was carried out by adhering to and implementing the original model, relying exclusively on publicly available abstractive summarization datasets.

The utilization of the Transformers library from Hugging Face was evident in the experiments conducted for text summarization, question answering, and model training. The input documents were constrained to 100 tokens, with summaries not exceeding 12 tokens. Beam search was executed with a num-beam value of 4, while maintaining a batch size of 6.

To facilitate comparison, identical news paragraphs were inputted into each model, resulting in varied outputs. The evaluation of accuracy was further conducted through the assessment of ROUGE scores.

The input news sample used for comparison purpose is

LOS ANGELES (AP) — In her first interview since the N
BA banned her estranged husband, Shelly Sterling says
she will fight to keep her share of the Los Angeles Clip
pers and plans one day to divorce Donald Sterling.
Los Angeles Clippers co-owner Shelly Sterling, below,
watches the Clippers play the Oklahoma City Thunder

along with her attorney, Pierce O'Donnell, in the first half of Game 3 of the Western Conference... (Associated Press) Shelly Sterling spoke to Barbara Walters, and ABC News posted a short story with excerpts from the conversation Sunday. NBA Commissioner Adam Silver has banned Donald Sterling for making racist comments and urged owners to force Sterling to sell the team. Silver added that no decisions had been made about the rest of Sterling's family. According to ABC's story, Shelly Sterling told Walters: "I will fight that decision." Sterling also said that she "eventually" will divorce her husband, and that she hadn't yet done so due to financial considerations. ||||| Shelly Sterling said today that "eventually, I am going to" divorce her estranged husband, Donald Sterling, and if the NBA tries to force her to sell her half of the Los Angeles Clippers, she would "absolutely" fight to keep her stake in the team. "I will fight that decision," she told ABC News' Barbara Walters today in an exclusive interview. "To be honest with you, I'm wondering if a wife of one of the owners, and there's 30 owners, did something like that, said those racial slurs, would they oust the husband? Or would they leave the husband in?" Sterling added that the Clippers franchise is her "passion" and "legacy to my family." "I've been with the team for 33 years, through the good times and the bad times," she added. These comments come nearly two weeks after NBA Commissioner Adam Silver announced a lifetime ban and a $2.5 million fine for Donald Sterling on April 29, following racist comments from the 80-year-old, which were caught on tape and released to the media. Read: Barbara Walters' Exclusive Interview With V. Stiviano Being estranged from her husband, Shelly Sterling said she would "have to accept" whatever punishment the NBA handed down to him, but that her stake in the team should be separate. "I was shocked by what he said. And -- well, I guess whatever their decision is -- we have to live with it," she said. "But I don't know why I should be punished for what his actions were." An NBA spokesman said this evening that league rules would not allow her tol hold on to her share. "Under the NBA Constitution, if a controlling owner's interest is terminated by a 3/4 vote, all other team owners' interests are automatically terminated as well," NBA spokesman Mike Bass said. "It doesn't matter whether the owners are related as is the case here. These are the rules to which all NBA owners agreed to as a condition of owning their team." Sherry Sterling's lawyer, Pierce O'Donnell, disputed the league's reading of its constitution. "We do not agree with the league's self-serving interpretation of its constitution, its application to Shelly Sterling or its validity under these unique circumstances," O'Donnell said in a statement released this evening in response the NBA. "We live in a nation of laws. California law and the United States Constitution trump any such interpretation." If the league decides to force Donald Sterling to sell his half of the team, Shelly Sterling doesn't know what he will do, but the possibility of him transferring full ownership to her is something she "would love him to" consider. Related: NBA Bans Clippers Owner Donald Sterling For Life "I haven't discussed it with him or talked to him about it," she said. The lack of communication between Rochelle and Donald Sterling led Walters to question whether she plans to file for divorce. "For the last 20 years, I've been seeing attorneys for a divorce," she said, laughing. "In fact, I have here-- I just filed-- I was going to file the petition. I signed the petition for a divorce. And it came to almost being filed. And then, my financial advisor and my attorney said to me, 'Not now.'" Sterling added that she thinks the stalling of the divorce stems from "financial arrangements." But she said "Eventually, I'm going to." She also told Walters she thinks her estranged husband is suffering from "the onset of dementia." Since Donald Sterling's ban, several celebrities have said they would be willing to buy the team from Sterling, including Oprah Winfrey and Magic Johnson. Sterling remains the owner, though his ban means he can have nothing to do with running the team and can't attend any games. Silver announced Friday that former Citigroup chairman and former Time Warner chairman Richard Parsons has been named interim CEO of the team, but nothing concrete in terms of ownership or whether Sterling will be forced to sell the team. Parsons will now take over the basic daily operations for the team and oversee the team's president. Read: What You Need to Know This Week About Donald Sterling ABC News contacted Donald Sterling for comment on his wife's interview, but he declined.

From this input text we have extracted a range of 2000 – 5000 tokens for different tasks.

The outputs for different tasks of our model is tabulated below :

| Task | Output |
| --- | --- |
| Abstractive summarization | Shelly Sterling spoke to Barbara Walters, and ABC News posted a short story with excerpts from the conversation Sunday. NBA Commissioner Adam Silver has banned Donald Sterling for making racist comments and urged owners to force Sterling to sell the team. Sterling added that the Clippers franchise is her "passion" and "legacy to my family" Silver announced Friday that former Citigroup chairman and former Time Warner chairman Richard Parsons has been named interim CEO of the team. Nothing concrete in terms of ownership or whether Sterling will be forced to sell the team. |
| | Shelly sterling says she will |

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

| | |
|---|---|
| Extractive Summarization | fight to keep her share of t he los angeles clippers. nba commissioner adam silver has banned donald sterling for making racist comment s. sterling : " i will fight that decision. i'm wondering if a wife of one of the owners of the team, and there's 30 owners, did something like that ". Nba bans clippers o wner Donald sterling for lif e. sterling's lawyer says she 's " shocked " by the league' s interpretation of its consti tution. " we live in a nation of laws. California law and t he united states constitutio n trump any such interpret ation, " she says. Former Ci tigroup chairman richard p arsons has been named int erim CEO of the team. pars ons will take over the team and oversee the team's pre sident. sterling has been th e owner, though he can't at tend any games. sterling wi ll now take over at the tea m, but his ban means he ca n have nothing to do with r unning. |
| Question-Answering | Question: what does shelly sterling said? Answer: " eventually, i am g oing to " divorce her estran ged husband, donald sterlin g, and if the nba tries to for ce her to sell her half of the los angeles clippers, she wo uld " absolutely " fight to ke ep her stake in the team. " i will fight that decision, " sh e told abc news ' barbara w alters today in an exclusive interview. " to be honest wi th you, i ' m wondering if a wife of one of the owners, a nd there ' s 30 owners, did something like that, said th ose racial slurs, would they oust the husband? or woul d they leave the husband in ? " sterling added that the c lippers franchise is her " pa ssion " and " legacy to my fa mily. " |

*C. Evaluation Results*

This section examines the proposed summarization models and hypotheses. The experimental analysis was carried out while considering two key factors: quality and quantity. Assessing the quality and readability of summary models can pose challenges due to variations in individuals' perceptions of an ideal summary. It should be emphasized that validation accuracy is not a suitable metric for evaluating summarization tasks. Consequently, the Rouge score system has commonly been employed as the primary criterion for assessing a summarizer model in various studies.

**ROUGE metrics**: The examination carried out in the tests involved computing the ROUGE-1, ROUGE-2, and ROUGE-L scores for every summary, as outlined in the source. The evaluation process utilized ROUGE-N (1, 2) and ROUGE-L metrics.

**ROUGE-N**: The notion of ROUGE-N is focused on evaluating the presence of n-grams in the candidate and reference summaries, with 'n' indicating the particular number of words for comparison. In the case of 1-gram metrics, the ratio of single words matched to the total words in the reference summary is calculated. For 2-gram metrics, the focus is on the ratio of two consecutive words matching in both summaries. In scenarios involving the summarization of multiple texts, the average of all n-gram values is employed.

$$\text{ROUGE-N} = \frac{\sum\limits_{S \in ReferenceSummaries} \sum\limits_{gram_n \in S} Count_{match}(gram_n)}{\sum\limits_{S \in ReferenceSummaries} \sum\limits_{gram_n \in S} Count(gram_n)}$$

$$\text{ROUGE-N}(Avg) = \frac{\sum R_n}{num.of documents}$$

Generally, evaluation measures are based on precision or recall. Precision is defined as the ratio 38 of matching words in candidate and reference summaries to the number of words in the candidate summary.

$$Precision = \frac{Num.of matchingn - units}{Num.of wordsincandidatesummary}$$

Recall-based models take into account the number of matched n-units divided by the number of n-grams in the reference summary.

$$Recall = \frac{Num.of matchingn - units}{Num.of wordsinreferencesummary}$$

**ROUGE-L**: Rouge-L investigates the longest common subsequence of terms that align between the candidate and reference summaries. A benefit of utilizing Rouge-L as opposed to n-gram ROUGE metrics is that sentences lacking a continuous match in words can still influence the overall summary evaluation. The n-gram metric does not necessitate predefinition prior to the commencement of the evaluation process, as it automatically identifies continuous n-units. The precision, recall, and F-measure values for Rouge-L are as specified.

$$P_{LCS} = \frac{LCS(C,R)}{words_{candidate}}$$

$$R_{LCS} = \frac{LCS(C,R)}{words_{reference}}$$

Another metric based on both precision(p) and recall(r) known as F-measure(f) is evaluated as follows

$$F_{LCS} = \frac{(1+\beta^2)R_{LCS} * P_{LCS}}{R_{LCS} + \beta^2 * P_{LCS}} \text{where } \beta \leq 1$$

Below Mentioned table is categorizing the ROUGE scores for all 3 models on CNN/Dailymail datasets.:

| Model Name | Rogue-1 | Rogue-2 | Rogue-l |
|---|---|---|---|
| Pegasus | 0.578 | 0.482 | 0.537 |
| Bert for summarization | 0.543 | 0.397 | 0.465 |

## CONCLUSIONS

Throughout the analysis, three unique models were matched up against one another in relation to text summarization and question answering tasks on the Multi News dataset: the BERT model customized for summarization, the BERT model optimized for question-answering, and the PEGASUS model. Analysis of the results revealed that each model displayed the ability to produce high-quality summaries, yet exhibited individual strengths and weaknesses. Despite the BERT model's success in grasping the semantic essence of the input text, it encountered issues when trying to generate a range of compelling summaries. Conversely, the PEGASUS model, specifically designed for abstractive text summarization, demonstrated proficiency in creating highly informative and varied summaries; however, the extensive computational resources required for training and operation of this model were substantial.

On the Multi News dataset, the PEGASUS model surpassed the BERT models in both ROUGE scores and evaluations conducted by human evaluators. This suggests that the PEGASUS model emerges as the preferred option for abstractive text summarization on this dataset. Still, the elevated computational demands of the PEGASUS model when compared to the BERT models may introduce a constraint in certain contexts. The decision on the appropriate model will rely on the particular requirements of the task, along with the resources for model training and deployment. Evaluating the BERT model for question-answering, its ability to provide accurate responses based on the given text was measured using rogue metrics, illustrating satisfactory performance.

## REFERENCES

[1] S. Narayan, S. B. Cohen, and M. Lapata, “Don’t give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization,” 2018, arXiv:1808.08745. [Online]. Available: http://arxiv.org/abs/1808.08745

[2] H. Zhang, J. Xu, and J. Wang, “Pretraining-based natural language generation for text summarization,” 2019, arXiv:1902.09243. [Online]. Available: http://arxiv.org/abs/1902.09243

[3] A. Srikanth, A. S. Umasankar, S. Thanu, and S. J. Nirmala, “Extractive text summarization using dynamic clustering and coreference on BERT,” in Proc. 5th Int. Conf. Comput., Commun. Secur. (ICCCS), Patna, India, Oct. 2020, pp. 1–5, doi: 10.1109/ICCCS49678.2020.9277220

[4] Y. Liu et al., “RoBERTa: A robustly optimized BERT pretraining approach,” 2019, arXiv:1907.11692. [Online].Available: http://arxiv.org/abs/1907.11692

[5] PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization Jingqing Zhang, Yao Zhao, Mohammad Saleh, Peter J. Liu
https://arxiv.org/abs/1912.08777

[6] Leveraging BERT for Extractive Text Summarization on Lectures
Derek Miller https://arxiv.org/abs/1906.04165

[7] A Comparative Study of Transformer-Based Language Models on Extractive Question Answering
Kate Pearce, Tiffany Zhan, Aneesh Komanduri, Justin Zhan
https://arxiv.org/abs/2110.03142