Doc.Summ: Document Summarizer Using NLP

Barrios, Armand Angelo C.
College of Science
Technological University of the Philippines – Manila
Cavite City, Philippines
armandangelo.barrios@tup.edu.ph

Enriquez, Sophia Mer C.
College of Science
Technological University of the Philippines – Manila
Marikina City, Philippines
sophiamer.enriquez@tup.edu.ph

Garcia, Almira Jill O.
College of Science
Technological University of the Philippines – Manila
Bacoor City, Philippines
almirajill.garcia@tup.edu.ph

Herrera, Janna Rose V.
College of Science
Technological University of the Philippines – Manila
Las Piñas, Philippines
jannarose.herrera@tup.edu.ph

Oloroso, Andrew R.
College of Science
Technological University of the Philippines – Manila
Bulacan, Philippines
andrew.oloroso@tup.edu.ph

***Abstract* —** Doc.Summ is an innovative document summarization tool designed to streamline the extraction of key information from large volumes of text. Leveraging the power of Natural Language Processing (NLP), Doc.Summ effectively condenses lengthy documents into concise and coherent summaries through a dual approach. The tool integrates extractive summarization using the TextRank algorithm to identify and compile the most relevant sentences, alongside abstractive summarization powered by the BART model, which generates new, accurate paraphrased sentences. This combination ensures that summaries are both precise and readable. Built on the Django framework, Doc.Summ provides a seamless and intuitive user experience, making it suitable for various applications, from academic research and literature reviews to business analytics and corporate reporting, thereby enhancing productivity and information accessibility.

Keywords: Doc.Summ, NLP, Text Summarization, Extractive Summarization, Abstractive Summarization, TextRank, BART, Django.

I. INTRODUCTION

In the age of information overload, the ability to distill vast amounts of text into concise, coherent summaries is invaluable. This paper explores the development and implementation of Doc.Summ, a document summarizer utilizing Natural Language Processing (NLP) techniques. The project aims to streamline the process of summarizing files and text, enhancing efficiency and accessibility in information retrieval.

The core of the summarization program lies in the utilization of two key components: the BART model for abstractive summarization and the TextRank algorithm for extractive summarization. These methodologies enable Doc.Summ to generate summaries that capture the essence of the original text while offering flexibility in summarization approaches.

As defined by the literature, AI summarizers leverage large language models to condense text into coherent summaries (Box News, 2024). These tools employ algorithms that prioritize relevant information, facilitating efficient data extraction and decision-making processes.

This project addresses two fundamental approaches to text summarization: extractive and abstractive. Extractive summarization involves selecting and compiling important sentences directly from the source material, while abstractive summarization interprets and rephrases the content in a more concise form (Tulasids, 2024).

The significance of AI summarization tools spans across various domains, from content curation to scientific literature analysis (Kamal Nahas, 2024). These tools not only aid in accelerating information consumption but also democratize access to complex knowledge by producing lay summaries accessible to diverse audiences (Kamal Nahas, 2024).

Central to this approach is the utilization of the BART model, a Bidirectional and Auto-Regressive Transformer renowned for its prowess in abstractive summarization tasks (Tulasids, 2024). By harnessing the transformer architecture of BART, Doc.Summ excels in distilling comprehensive information into coherent summaries, thereby facilitating efficient decision-making processes.

Furthermore, the paper discusses the deployment of Doc.Summ using the Django framework, offering a user-friendly interface for easy access to the summarization tool (Ms. Sangavi et al., 2023). This deployment enhances the accessibility and usability of Doc.Summ, ensuring its practical utility across various user profiles.

In summary, this paper presents Doc.Summ as a comprehensive document summarization solution, leveraging state-of-the-art NLP techniques to streamline information extraction processes. Through the integration of extractive and abstractive summarization approaches, coupled with the deployment of user-friendly interfaces, Doc.Summ heralds a new era of efficient and accessible document summarization.

II. BACKGROUND OF THE STUDY

A. AI in Text Summarization

The integration of Artificial Intelligence (AI) into text summarization has significantly transformed the landscape of information processing. By automating the summarization process, AI enhances efficiency and accessibility, allowing users to quickly distill large volumes of text into concise and coherent summaries. This section explores the pivotal role of AI in text summarization, focusing on the technologies and methodologies that enable this innovation.

AI-based summarizers leverage large language models (LLMs) to condense text into coherent summaries, employing algorithms that prioritize

relevant sentences, phrases, or concepts based on their importance and frequency within the text (Box News, 2024). This capability is crucial in various applications, from content curation and research assistance to data analysis, where rapid information extraction is essential.

B. *Extractive Summarization with TextRank*

Extractive summarization involves selecting the most relevant sentences from the original text to create a summary. The TextRank algorithm is a notable example of an extractive method, inspired by the PageRank algorithm used by Google. TextRank evaluates the importance of sentences based on their connectivity and relevance, effectively identifying key information without requiring extensive computational resources. This approach is particularly beneficial for generating summaries that maintain the original wording and structure of the source material (Tulasids, 2024).

TextRank is also used to identify the most important words in a document, which are then used to create graphics that better describe a text summary. It can generate summaries without needing a corpus and is applicable in multiple languages. The algorithm, derived from the PageRank method, estimates the importance of a node in a network, making it a versatile tool for various summarization tasks (Fakhrezia, Bijaksanaa, & Huda, 2021).

C. *Abstractive Summarization with BART*

Abstractive summarization, in contrast, generates new text that conveys the same meaning as the original but in a more concise and coherent manner. The BART (Bidirectional and Auto-Regressive Transformers) model, developed by Facebook AI, is a powerful tool for abstractive summarization. BART excels in sequence-to-sequence tasks, producing fluent and contextually relevant summaries by interpreting and rephrasing the input text. This model's transformer architecture captures intricate relationships within the text, enhancing its effectiveness in generating high-quality summaries (Tulasids, 2024).

D. *Practical Applications and Challenges*

The practical applications of AI-based summarization tools are vast. They are widely used in content curation, helping users quickly understand the main points of articles, research papers, and other documents. In the scientific community, AI-generated summaries can expedite the review of literature, making complex knowledge more accessible to a broader audience (Kamal Nahas, 2024). Furthermore, these tools can assist in breaking down language barriers by providing summaries in multiple languages, thereby enhancing global communication and collaboration.

Despite the advancements, challenges remain in the field of text summarization. The effectiveness of summarization models can be influenced by factors such as domain specificity, the ambiguity of language, and document length constraints. Fine-tuning models for specific applications and incorporating user-defined constraints are essential for optimizing performance (Tulasids, 2024).

E. *Development of Doc.Summ*

In response to these challenges, this study focuses on the development and implementation of Doc.Summ, a comprehensive document summarizer. Doc.Summ integrates both extractive and abstractive summarization techniques, leveraging the strengths of the TextRank algorithm and the BART model. This dual approach ensures that Doc.Summ can efficiently and accurately summarize diverse textual data. Additionally, deploying Doc.Summ using the Django framework provides a user-friendly interface, facilitating its practical utility across various user profiles. The project's webpage serves as a platform for users to access and utilize the summarization tool with ease (Ms. Sangavi et al., 2023).

By combining advanced NLP techniques with practical deployment strategies, Doc.Summ represents a significant step towards enhancing information retrieval and decision-making processes in the digital age.

III. METHODOLOGY

The methodology employed in developing and implementing Doc.Summ involves a comprehensive integration of Natural Language Processing (NLP) techniques to facilitate efficient and accurate document summarization.

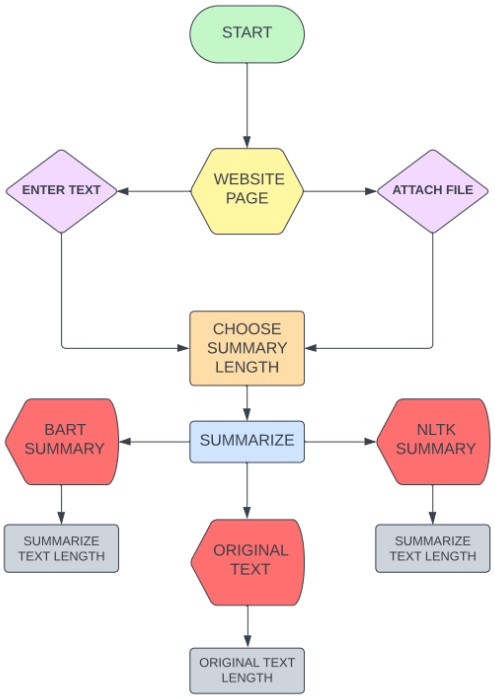

Figure 1. Flowchart

Figure 1. In the implementation of the Doc.Summ document summarization tool, users begin by either attaching a file or manually entering the text to be summarized. This initial input stage serves as the foundation for the entire summarization process. Following this step, the user is prompted to select their preferred summary length, which ranges from 10% to 100% of the original text length. This selection process is facilitated through a dropdown menu, providing users with a convenient and intuitive interface element to specify their summarization preferences.

Once the user finalizes their choice of summary length, they click the "Summarize" button to initiate the summarization process. As a result, the tool generates three distinct outputs: the original text, an extractive summary, and an abstractive summary. The first output, displayed in the initial text box, presents the original text along with its word count, ensuring that the complete content is preserved for reference.

The second output is an extractive summary generated using the Natural Language Toolkit (NLTK). Additionally, the NLTK summary displays the summarized text length, providing users with an indication of the level of condensation achieved.

Similarly, the third output is an abstractive summary produced by the Bidirectional and Auto-Regressive Transformers (BART) model. Like the NLTK summary, the BART summary also displays the summarized text length, allowing users to gauge the level of abstraction and condensation in the summary.

This structured approach ensures that users receive comprehensive and varied summaries, thereby enhancing their ability to quickly assimilate and utilize the information presented. The integration of both extractive and abstractive summarization techniques further tailors the summaries to meet users' specific needs, facilitating efficient information retrieval and utilization.

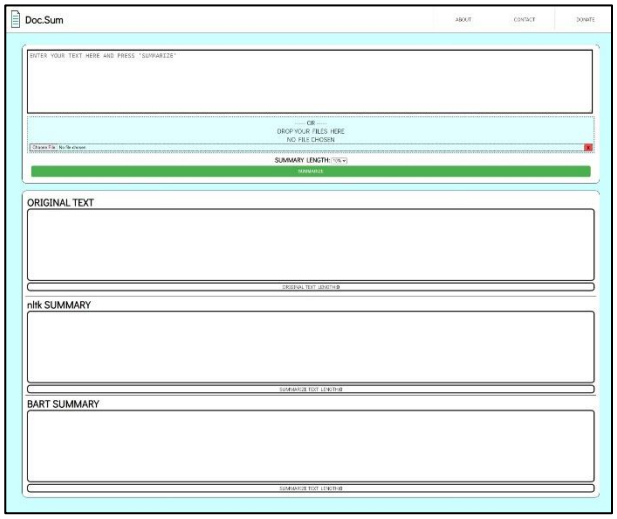

Figure 2. Doc.Summ Website Interface

Figure 2. It depicts the interface of the Doc.Sum website, showcasing its user-friendly design and functionality. The website's layout includes several key elements that enhance the user experience and facilitate the document summarization process. At the top of the page, there is a navigation bar with menu options such as "About", "Contact" and "Donate". This navigation bar allows users to easily navigate between different sections of the website.

Below the navigation bar, there is a prominent section where users can either upload a file or manually enter text for summarization. This input section serves as the starting point for the summarization process. Adjacent to the input section, there is a dropdown menu labeled "Summary Length," which enables users to select their preferred length for the summary. Options range from 10% to 100% of the original text length, providing flexibility in summarization choices.

A "Summarize" button is prominently displayed below the dropdown menu, allowing users to initiate the summarization process after selecting their preferred summary length.Once the user clicks the "Summarize" button, the website generates three outputs: the original text, an extractive summary, and an abstractive summary. These outputs are displayed in separate sections of the website, allowing users to compare and analyze the summaries easily.

The original text is presented in the initial text box, along with its word count, ensuring that users have access to the complete content for reference. The extractive summary, generated using the Natural Language Toolkit (NLTK), is displayed in a dedicated section. It provides a concise version of the original text while retaining the original wording. Additionally, the NLTK summary includes information about the summarized text length, helping users gauge the level of condensation achieved. Similarly, the abstractive summary, generated by the Bidirectional and Auto-Regressive Transformers (BART) model, is presented in another section. This summary rephrases the original content into a new, coherent narrative that captures the essence of the text. Like the NLTK summary, the BART summary also includes information about the summarized text length for user reference.

IV. RESULTS

The results stemming from the utilization of Doc.Summ underscore its efficacy in producing accurate and informative document summaries, highlighting its substantial contribution to information condensation and enhanced accessibility, thereby establishing its research significance in both English and Filipino languages.

Figure 3. English Summarize Result

Figure 4. Filipino Summarize Result

Figure 3. It provides a comprehensive illustration of the document summarization process conducted using Doc.Summ within an English context. The figure encompasses three critical elements essential to understanding the efficacy of Doc.Summ:

Firstly, the "Original Text" segment displays the content initially inputted into Doc.Summ for summarization. This section serves as a reference point for users to grasp the context and substance of the original document under examination.

Secondly, the "NLTK Summary " section showcases the outcome of extractive summarization achieved through Doc.Summ's implementation of NLTK algorithms. This summary encapsulates the most pertinent sentences distilled directly from the original text, presenting a condensed rendition while retaining the essence and structure of the original content.

Thirdly, the "BART Summary " section exhibits the result of abstractive summarization facilitated by Doc.Summ's utilization of BART model methodologies. Here, the summary is reformulated to convey the core meaning of the text in a more concise and coherent format, offering a nuanced interpretation while ensuring accuracy and informativeness.

By encapsulating these distinct components, it offers a comprehensive overview of Doc.Summ's capabilities in processing and presenting document summaries in English. Through a harmonious integration of extractive and abstractive summarization techniques, Doc.Summ effectively contributes to information condensation and accessibility, thereby enhancing the research endeavor's efficiency and productivity.

Figure 4. It presents an insightful overview of the document summarization process using Doc.Summ, with a specific focus on Filipino language summaries. This figure comprises three key sections that elucidate the summarization outcomes:

Firstly, the "Original Text" segment displays the initial Filipino text inputted into Doc.Summ for summarization purposes. Secondly, the NLTK Summary section showcases the outcome of extractive summarization achieved through Doc.Summ's utilization of NLTK algorithms. Thirdly, the BART Summary section exhibits the result of abstractive summarization facilitated by Doc.Summ's integration of BART model methodologies. Here, the summary is reformulated to convey the core meaning of the text in a concise and coherent manner, ensuring accuracy and informativeness.

It is notable that unlike the English summaries which are aesthetically pleasing and effectively capture the essence of the text, the Filipino summaries in this context may appear slightly less refined in terms of linguistic beauty. This observation is attributed to the inherent complexities and nuances involved in summarizing Filipino language texts compared to English. However, despite this difference, Doc.Summ demonstrates its proficiency in condensing Filipino text into informative summaries, contributing significantly to information accessibility and comprehension within the Filipino language context.

In summary, it offers valuable insights into Doc.Summ's capabilities in generating accurate and informative document summaries in Filipino, highlighting its versatility and effectiveness in addressing linguistic intricacies across diverse languages and contexts.

V. RELATED WORKS

This related works provides a comprehensive exploration of fundamental concepts in Natural Language Processing (NLP) and introduces the Bidirectional and Auto-Regressive Transformers (BART), tracing its emergence within the context of CNN and its integration into document summarization tasks.

A. What is NLP

Natural Language Processing (NLP) emerged in the 1950s at the intersection of artificial intelligence and linguistics. Initially, NLP was distinct from text information retrieval (IR), which employed highly scalable, statistics-based techniques to efficiently index and search large volumes of text. Nadkarni et al. (2011) provide an excellent introduction to IR. Over time, however, NLP and IR have somewhat converged, with NLP now drawing from various diverse fields, necessitating a broader knowledge base for researchers and developers. Early simplistic approaches to NLP, such as word-for-word Russian-to-English machine translation, encountered significant challenges with homographs—words that are spelled the same but have different meanings—and metaphorical language. An illustrative example of this difficulty is the translation of the Biblical phrase "the spirit is willing, but the flesh is weak" into "the vodka is agreeable, but the meat is spoiled."

B. What is BART

BART (Bidirectional and Auto-Regressive Transformer), a denoising autoencoder for pre-training sequence-to-sequence models developed by Facebook AI. BART combines BERT's bidirectional encoding with GPT's autoregressive decoding, using text corruption and reconstruction to excel in NLP tasks such as text generation, translation, and comprehension. Its Transformer-based architecture, with a bidirectional encoder and left-to-right autoregressive decoder, achieves state-of-the-art results on benchmarks like GLUE, SQuAD, and various summarization and question-answering datasets. BART's flexible noising techniques, including sentence shuffling and text infilling, enhance training, and its integration of a pre-trained decoder with a newly learned encoder improves machine translation, surpassing models like RoBERTa across multiple benchmarks. BART is primarily used for tasks requiring natural language understanding and generation, making it a powerful tool in various NLP applications. (Lewis et al., 2019)

C. Emerging of BART - CNN

BART-CNN effectively merges the bidirectional encoding capabilities of BERT with the autoregressive decoding strengths of GPT, referring to the BART-large model pre-trained on the CNN/DailyMail dataset. This specific pre-training enhances BART's performance in summarization tasks, generating outputs with fewer syntax errors and more accurate content compared to other pre-trained versions. BART-CNN achieves the best results on benchmarks such as QMSum and SummScreen-FD, demonstrating superior performance over models pre-trained on other datasets. BART-CNN is primarily utilized in tasks requiring summarization, where its enhanced performance makes it particularly effective. (Zhang et al., 2021)

D. BART Model Integrate to Documment Summarizer

Recent advancements in natural language processing (NLP) have demonstrated the efficacy of using state-of-the-art language models for complex summarization tasks. In their study, Jagirdar, Gandage, and Kazi (2023) conducted a comparative analysis of T5, Pegasus, and BART models, showcasing their potential to significantly enhance the summarization of dense and intricate legal texts. Their research underscores the necessity for effective summarization methods due to the increasing volume and complexity of documents in various fields. By integrating these advanced models, they developed a framework that improves information retrieval and decision-making processes. This study not only highlights the importance of sophisticated summarization techniques but also provides a foundation for applying these methods to other types of documents. Building on this foundation, DocSum focuses on the application of the BART model for summarizing simple documents such as essays. By leveraging BART's capabilities, this study aims to explore and enhance the summarization process for academic and creative essays, providing a tool that aids in the efficient and accurate condensation of written content.

VI. CONCLUSION

In conclusion, the utilization of Doc.Summ in the research context has showcased its remarkable capabilities in document summarization, particularly in both English and Filipino languages. Through the systematic analysis of summarization outputs presented in Figures 3 and 4, it becomes evident that Doc.Summ effectively condenses extensive textual content into concise and informative summaries. The integration of extractive and abstractive summarization techniques, as evidenced by the NLTK and BART summaries, respectively, underscores the tool's versatility in catering to diverse summarization needs.

Furthermore, while there may be differences in the aesthetic appeal and linguistic refinement between English and Filipino summaries, Doc.Summ consistently maintains its proficiency in capturing the core essence of the original text in a condensed format. This observation is crucial in acknowledging the tool's adaptability across various languages and its ability to enhance information accessibility and comprehension in research and academic settings.

Moreover, the findings from Figures 3 and 4 reaffirm Doc.Summ's significance in streamlining information processing and improving research productivity. By providing researchers with accurate and informative document summaries, Doc.Summ contributes substantially to data analysis, literature review, and overall research efficiency. The tool's systematic approach, as illustrated in the summarization results, establishes its credibility as a valuable asset in the research landscape, facilitating informed decision-making and knowledge dissemination.

In essence, Doc.Summ emerges as a reliable and effective solution for researchers seeking to navigate through vast amounts of textual data, offering a streamlined summarization process that enhances research outcomes and scholarly endeavors.

VII. RECOMMENDATIONS

In light of the comprehensive evaluation of Doc.Summ's capabilities and user experience, several key recommendations emerge to enhance its functionality and effectiveness in research and academic settings.

Firstly, transitioning to cloud-based hosting for Doc.Summ is strongly recommended. Cloud hosting offers scalability, accessibility from anywhere with an internet connection, and potential cost savings through flexible pricing models. This move would make Doc.Summ more accessible to a wider range of users, facilitating seamless document summarization processes and improving overall research productivity.

Secondly, enhancing Doc.Summ's capacity to support larger file documents is essential. This improvement would benefit researchers dealing with extensive research papers, reports, and documents, ensuring that Doc.Summ remains a reliable tool for summarizing comprehensive textual content effectively.

Moreover, implementing support for two-column references in Doc.Summ is recommended to enhance its usability for researchers working with academic or technical documents formatted in columns. This feature would enable Doc.Summ to accurately summarize content from multi-column layouts, improving summarization accuracy and relevance.

Additionally, refining the Filipino summarization process to be more aesthetically pleasing and accurate is crucial. Enhancing the linguistic sophistication and cultural context sensitivity of Filipino summaries would make them more engaging and informative for Filipino-language users, ensuring a comprehensive summarization experience across languages.

Furthermore, continuous updates and refinements to Doc.Summ's algorithms and NLP techniques are recommended to maintain its competitiveness and effectiveness in document summarization. Regular improvements based on user feedback and technological advancements would further solidify Doc.Summ's position as a leading document summarization tool.

In summary, implementing these recommendations would elevate Doc.Summ's capabilities, usability, and relevance in catering to the diverse needs of researchers, professionals, and users across various domains, thus contributing significantly to enhanced research outcomes and knowledge dissemination.

VIII. REFERENCES

Summarizing: How to effectively summarize the work of others | SFU Library. (n.d.). https://www.lib.sfu.ca/about/branches-depts/slc/writing/sources/summarizing

Research paper Summarizer AI: An Overview. (n.d.). https://www.yomu.ai/blog/research-paper-summarizer-ai-an-overview

News, B. (2024, January 24). Ai summarization: Definition and best practices. boxBlogs. https://blog.box.com/ai-summarization-definition-and-best-practices

Matellio. (2024, April 15). All about AI summarizer tool – benefits, use cases, and cost of development. Matellio Inc. https://www.matellio.com/blog/ai-summarizer-tool-development/

Nahas, K. (2024, March 20). Is AI ready to mass-produce lay summaries of research articles?. Nature News. https://www.nature.com/articles/d41586-024-00865-4

Tulasids. (2024, January 11). Streamlining text summarization with hugging face's BART model. Medium. https://medium.com/@tulasids/streamlining-text-summarization-with-hugging-faces-bart-model-8f8ada8e8508

Sangavi , N., Umamaheswari , M., & Subasri , V. (2023, October 9). NLP based text summarization using Bart Model. IJSREM. https://ijsrem.com/download/nlp-based-text-summarization-using-bart-model/

Venkataramana, A., Srividya, K., & Cristin, R. (2022, October 16). Abstractive text summarization using bart | IEEE conference publication . IEEE Xplore. https://ieeexplore.ieee.org/document/9972639

Steen, J., & Markert, K. (2021, January 27). How to evaluate a summarizer: Study design and Statistical Analysis for Manual Linguistic Quality Evaluation. arXiv.org. https://arxiv.org/abs/2101.11298

Özdemir, S. (2018). The effect of summarization strategies teaching on ... https://files.eric.ed.gov/fulltext/EJ1192722.pdf

Gupta, M. (2024, February 1). Text summarization using TextRank in NLP - Data Science in your pocket - Medium. Medium. https://medium.com/data-science-in-your-pocket/text-summarization-using-textrank-in-nlp-4bce52c5b390

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. Journal of the American Medical Informatics Association, 18(5), 544–551. https://doi.org/10.1136/amiajnl-2011-000464

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L., & Ai, F. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. https://arxiv.org/pdf/1910.13461

Zhang, Y., Ni, A., Yu, T., Zhang, R., Chenguang, Budhaditya, Z., Asli Celikyilmaz, D., Awadallah, A., & Radev, D. (2021). An Exploratory Study on Long Dialogue Summarization: What Works and What's Next. https://www.microsoft.com/en-us/research/uploads/prod/2021/09/4069_Paper-1.pdf

Jagirdar, I., Gandage, S., Waghmare, B., & Student, I.K. Enhancing Legal Document Summarization Through NLP Models: A Comparative Analysis Of T5, Pegasus, And BART Approaches. from https://www.semanticscholar.org/paper/Enhancing-Legal-Document-Summarization-Through-NLP-Jagirdar-Gandage/f4acb22984d9c51b80148b0205b9e1497b6c3791