ISSN No:-2456-2165

AI Powered Innovative News Curation: The AI Approach to Transform and Enhance the Media Landscape with Efficiency and Quality Bridged

Naga Tulasi[1]; Shilpa Sree[2]; Siva Kumar[3]; Indira Kumar[4]; Shirish Kumar Gonala[5]; Bharani Kumar Depuru[6]
[1,2]Research Associate, [3]Mentor, [4]Team Leader, Research and Development, [5,6]Director
Innodatatics, Hyderabad, India

*Corresponding Author: Bharani Kumar Depuru
ORC ID: 0009-0003-4338-8914

Abstract:- In the dynamic realm of the global news industry, media outlets employ extensive efforts to curate and present data gathered from diverse online and offline sources. The intricate procedure involves not only the collection of news but also its condensation into succinct and informative summaries, often accompanied by a meticulous classification based on significance. However, this multifaceted undertaking is inherently time-consuming, demanding substantial manual input. Enter the realm of AI, a transformative force poised to revolutionise the landscape of news processing.

LLM technology, such as advanced models developed through tailored ML, offers a compelling solution to the problems involved in news extraction and creation of summaries. These algorithms, which are designed to handle unimaginable data, can swiftly and accurately analyse manually and programmatically collected data from a myriad of resources, significantly reducing the burden of manual efforts. The implementation of such AI-driven solutions not only expedites the news processing workflow but also introduces a layer of sophistication by discerning and categorising the importance of news items with unparalleled accuracy. Moreover, the integration of Python libraries designed for online data scraping enhances the efficiency of the entire system. These libraries organise the process of gathering information from various sources, offering a seamless flow of data to the LLMs. The combined synergy of this technology not only saves valuable time but also elevates the quality of news reporting, ensuring that the audience receives well-curated, relevant, and timely information.

In essence, the marriage of AI, LLM models, and Python libraries forms a powerful triumvirate, offering an intelligent and automated solution to the issues prevalent in news extraction and summarising it. As the media landscape continues to evolve, these advancements represent a pivotal step towards a future where information dissemination is not only efficient but also nuanced and tailored to the diverse needs of a global audience.

In this paper, our aim is to implement a model that automatically scrapes the content and summarises it, leveraging advanced scraping python packages and LLM models. Through the seamless overlapping of these technologies, we strive to revolutionise NEWS processing, ensuring swift, accurate, and insightful summarization of diverse resources.

Keywords:- *News Summarisation, News Scraping, Large Language Modelling, Natural Language Processing, Artificial Intelligence, Text Extraction, Web Scraping, Generative AI.*

I. INTRODUCTION

In the digital age the ML models based on Transformers are making ground breaking records not only in the field of IT industry, but also other fields like manufacturing biotechnology and many more. We are making an attempt to incorporate this booming concept in the field of media industry. These sophisticated models exemplified by ground breaking examples like Open AI GPT-3, BERT, T5 harness the power of vast Neural Networks comprising billions of weights which are capable of comprehending intricate linguistic nuances LLMs showcase unbeatable proficiency of achieving operations, such as translation summarization and creative content creation etc. In the rapidly evolving landscape of technology Artificial Intelligence ai and large language models (LLMs) are at the forefront revolutionising. How we engage with digital tools this exploration ventures into the significant influence these innovations wield within journalism and media through their Cutting-Edge capabilities in Natural Language understanding and generation. These models herald a new era of storytelling content creation and information dissemination marking a paradigm shift in the nexus of technology and communication [11].

The News industry can exploit the capacity of these models without any expenditure and deploy this into production for better and increased productivity. The time-consuming process like news collection from various online sources have been ultimately made simple using the power of Scrapy Python package. That, summarising the collected articles from various sources demands intervention of highly

skilled subject matter experts, can be replaced by this model, where manual intervention of such subject matter experts is minimal. This entirely automated process will help the media professional focus on analysis, enhancing and enrich the quality and preciseness of the summary [31].

The LLM that has been built performs swiftly in summarising the textual content in an abstract manner, without changing the content of the story. The model scrapes the data from numerous online websites in a few minutes and performs the abstractive summarizations, which is the process of reducing the large amount of textual content into short and quickly readable content while preserving the intrinsic meaning of the content [34, 36].

The model that we have built, based on T5 architecture, conducts abstractive summarization on scraped articles. Leveraging the “t5-small" pre-trained model, it efficiently processes articles by dividing them into chunks based on a maximum token limit. The task of summary generation is performed through a function that includes LLM in which each batch of tokenized text is passed. To handle articles surpassing the token limit, the model intelligently divides them into manageable batches, by making sure that summarization is concise [31]. The implementation showcases a robust integration of advanced textual data processing techniques, resulting in streamlined and informative summaries. The final outputs are stored in a CSV file, providing a practical solution for news summarization with minimal manual intervention [22].

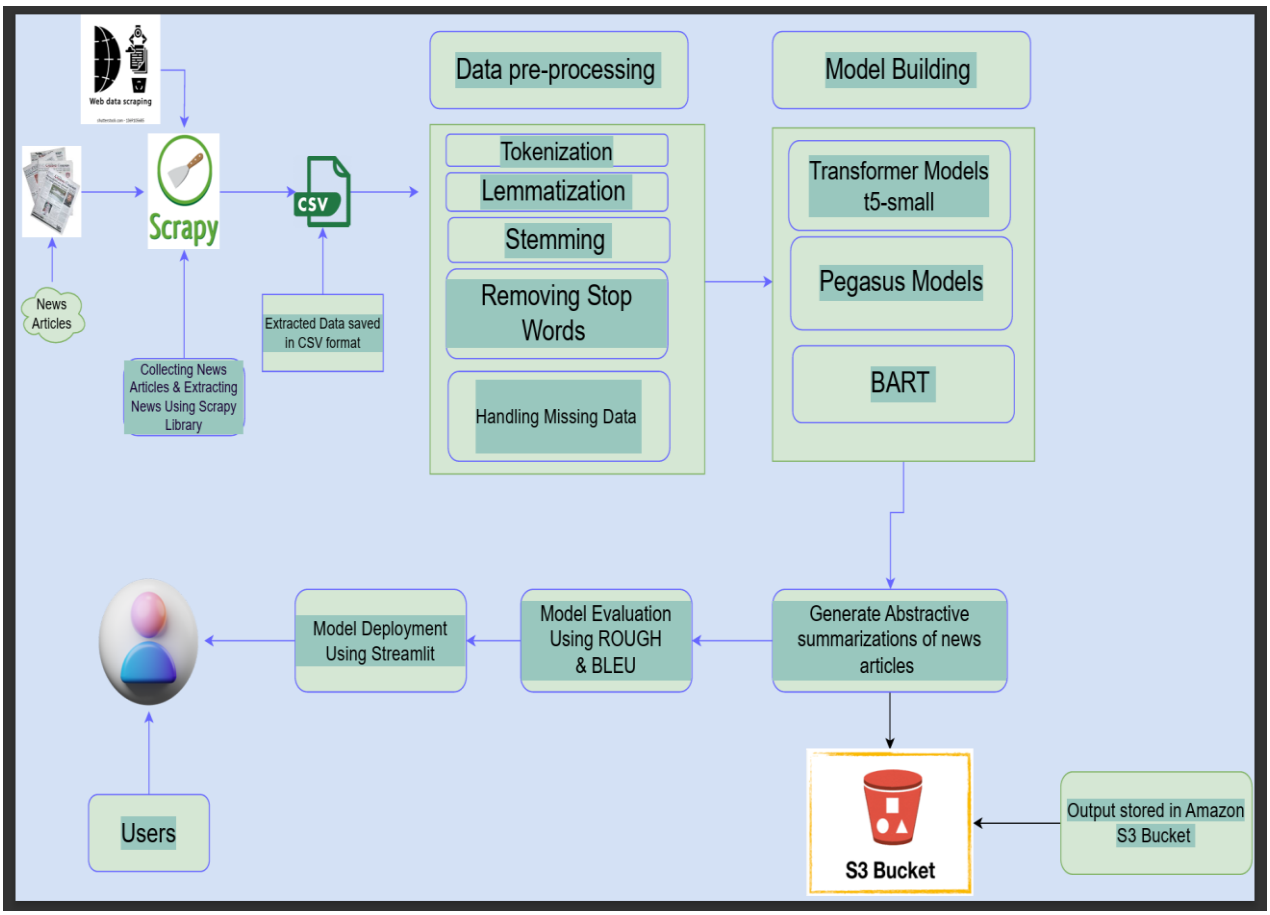

Fig. 1: Architecture Diagram- Offers a Visual Roadmap of its Integral Components and Sequential Steps
Source: Mind Map - 360DigiTMG

In our study for news article summarization, the data flow follows a meticulously crafted journey comprising two essential phases. Initially, we embarked on the task of extracting news articles from various online sources, employing the powerful web scraping tool Scrapy, due to its ability to handle vast amounts of data swiftly and efficiently [23,10]. With a focus on speed and accuracy, the extracted data was transformed into text format and seamlessly exported to an S3 bucket, laying the groundwork for the subsequent phase.

Transitioning to the data summarization phase, we entrusted the task to the transformative capabilities of the T5-small model, meticulously fine-tuned with precision using hyperparameters to ensure the generation of abstractive summaries of the desired size. This phase marked the culmination of our efforts in distilling raw textual data into concise and informative summaries. Finally, with the deployment of the complete application using Streamlit, we seamlessly integrated all components, offering a user-friendly interface while ensuring the storage of both the extracted data and their corresponding summaries in the S3 bucket, thus completing the journey of our data processing endeavour with finesse and efficiency.

II. METHODS AND METHODOLOGY

A. *Business Understanding*

Initiating a Data Science study entails a structured approach, and the CRISP-ML(Q) methodology [Fig.2] provides a comprehensive framework [37]. The very first step of this study is to lay down the foundation to the solution which can be done only through the in-depth understanding of business problems, business environment and problems being faced by the industry. After the thorough study of this, it is found that the firm loses most of its time only in online sources to derive useful insight from news articles already published on various reliable pages. Once the insights are derived, it is evaluated by the expert journalist team to ensure credibility and to ensure legal compliance. This is the process which must be completely automated using AI driven sources. From the in-depth study of organisational behaviour, it was very clear that only most of the portions of the process can be automated. Although we incorporate AI in the media industry, zero percent human intervention is not possible at all. Such substitution can lead to misinterpretation of insights derived from various sources. Incorporation of ML algorithms, in the media industry, facilitates customer satisfaction and engagement through personalised marketing strategy like recommendation engine [32].

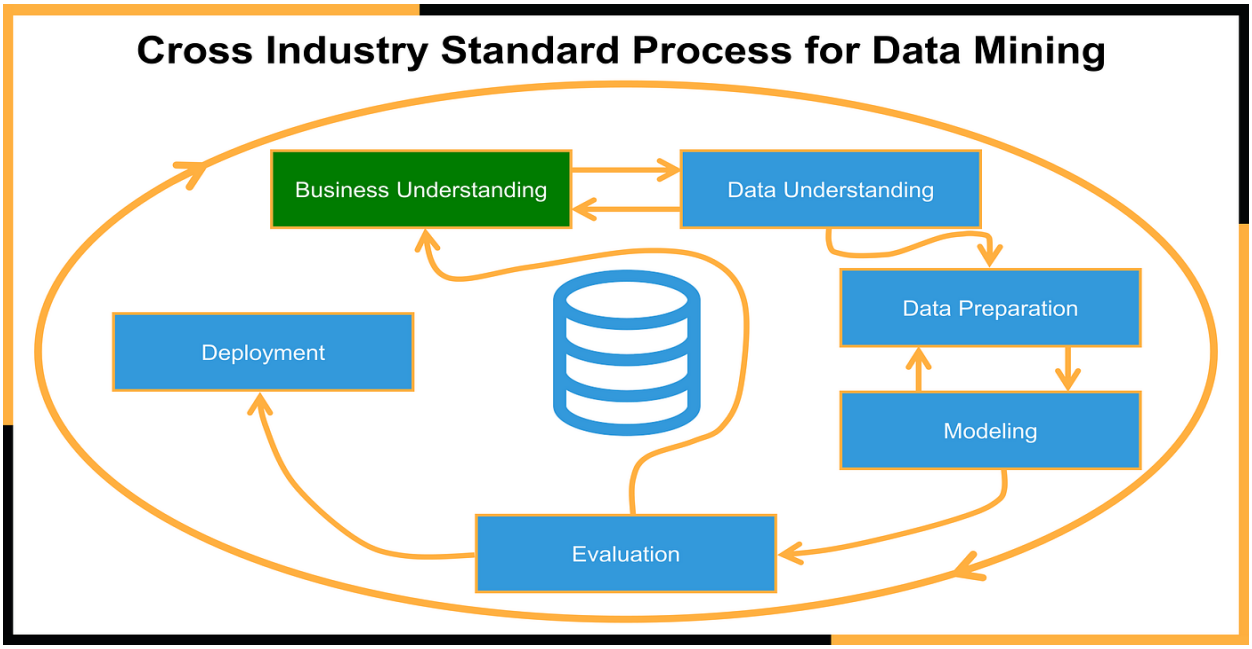

Fig. 2: CRISP-MLQ Methodology
Source: Source 360DigiTMG

➢ *Description about Media Industry and their Work Process*

Newspapers, broadcasts on tv/radio, facebook/x/instagram based news channels, online news publications[7] etc are the various news media that exist now. The objective of these platforms is to collect news and deliver it to the public. Below are the key points of how the media industry works.

- **Content Creation:** Once the is gathered, it is converted into various news forms such as news stories, articles, videos, broadcasts or podcasts. This is done through writing, editing which needs the coordination with other professionals like photographers, videographers, or graphic designers. They will add elements which make the presentation of the content better.[35].
- **Editorial Process:** The editors review those written contents and modify if necessary, to make sure it meets the criteria, such as journalistic standards, objectivity and publication's guidelines. Also they can decide how to arrange them in the outlet by comparing the priority of the various news.
- **Distribution:** During the next stage finalised content is delivered to the public through various channels mentioned already. They are traditional newspapers, news websites, news apps and social media news channels. Among all these social media has gained more popularity which allows quick dissemination to a vast audience.
- **Adapting to New Trends:** The advancement in technology affects people's preferences which in turn make changes in the media industry as well. For example, after the mobile phone usage become very common people are able to access the news websites in the phone, optimising the content for mobile devices as well. Data analytics provide a way to analyse and conclude on the audience taste and behaviour.
- **Challenges:** Like any other industry the media also encounters different challenges such as misinformation, reducing the number of advertisements and the profit in traditional outlets, outbreaking of social media news sources affected inversely, difficulty to balance between the two objectives such as profit and quality of the deliverable, to maintain the news unbiased. The industry is also prone to rapid changes according to the trends in the society and people's preferences.

Summarising collected information to a very good report which people can read easily is a hectic task. It requires in-depth knowledge of the subject and exceptional comprehension skills[31,23].So, the best method is to utilise the capabilities of LLMs to accomplish this task. That will increase the productivity of the print media, since it accomplishes summarisation very quickly without any human involvement.

B. Data Understanding

The collected articles are of raw format which includes special symbols, unwanted numbers, unidentifiable names, emojis, hashtags, etc. Since most of the contents are going to be in the form of text, cleaning this raw content to LLM consumable format requires a lot of effort which necessitates the role of NLP and many other sophisticated python packages. To build an extensive pipeline which will remove any type of unexpected anomalies would also need a deep understanding of the dataset. Machines don't understand the language just like humans do. So the special techniques are required to make the machine understand the context of the input text that is passed to LLM. LLM plays a key role in understanding the relationship between words by performing two different types of logics. One of them is word embedding and the other one is positional embedding.

The positions of each word are very important in LLM, misplacement of the few words can change the context of the entire sentence. Each word is assigned a number and it passed into the Attention layer where the relationship is captured. The semantic relationship between the words of a sentence is captured during the training of the neural network. The word embeddings are the weights or trained parameters of the neural network by passing huge corpus of data

C. Difficulties in Manual Extraction and Summarization

Copy pasting the articles manually is a tedious task, which not only eats out the time of news aggregators but affects the productivity and leads to delayed publication. Hence the news cannot be disseminated to the public for consumption. However, this study attempts to avoid such mistakes by incorporating the highly powerful python tool, Scrapy. This would seamlessly scrap the news articles from multiple news websites and forums, which would take only a maximum of 5 minutes of time. Even though big tech giants like MicroSoft offer APIs, those are not up to the mark and this article also delves into difficulties posed by ML models that summarises the textual contents [21, 29]. Manually summarising the content that consists of various technical terms is very difficult, because that needs strong comprehension and subject knowledge. Take a scientific paper on COVID-19, for instance. Should the summary contain technical terms, or be easily understandable for the general reader? Should it list facts or entice with a compelling narrative? [36]

- *Data Description*

In the study, we utilised raw textual content scraped from a variety of news websites, capturing the dynamic nature of global news coverage with daily volumes ranging from 2MB to 200MB. This reflects the variability in the amount of news reported each day, with the number of news articles scraped averaging between 100 to 400, depending on the day's news coverage across different forums. The content is meticulously collected and stored in CSV format, creating a substantial corpus that mirrors the broad spectrum of topics and perspectives found in global news forums. The methodology of collecting the data is designed to ensure the dataset is inclusive, collecting a wide range of articles from reputable sources. This enriches our dataset with diverse perspectives and subjects, making it a rich resource for analysis. The structure of the CSV files is optimised for scalability and ease of access, allowing for efficient handling and analysis, and the potential for seamless integration of additional data as the study progresses. Through this methodical approach, we aim to build a robust dataset that not only reflects the current news landscape but also serves as a critical foundation for subsequent analyses, particularly in applications involving Large Language Models. This foundational dataset is pivotal for exploring trends, biases, and the impact of media in a collective manner, giving invaluable insights into the complexities of consumption of news and dissemination in the digital age.

- *Common Challenges in Data Collection in the Field of Media*

The process of information gathering for news reports pose lot of challenges and those challenges impact the quality, significance and usability of the collected data. Especially in the case of media industry these problems are more significant due to the dynamic nature of content consumption and the involvement of audience[32].Assuring the data accuracy is the is an important concern.

The constantly changing platforms and technologies upkeeping precision of metrics such as viewership, audience demographics, and engagement rates poses a constant challenge. Sustaining a harmony among collecting a valuable data and ensuring user privacy rights is another challenge. This means dealing with changing rules and moral worries, especially about targeted advertising and personalised content delivery.

Another challenging task is - combining the data from different channels including tv, radio, websites, apps, social media etc, blending them and analysing them efficiently. Live data analysis is very important in media decision-making, which needs fast processing and interpretation of data to adapt strategies quickly. Ensuring data-quality, bias related issues, inconsistent standards, and repeated entries which affect the reliability of inferences drawn from the collected information are another set of problems.

The various formats of data text, images, videos and their volume make it even more complex. This complexity may lead to a problem in deriving significant findings from the data. Additionally the data collection methodologies need to be continuously upgraded according to the changes and trends[32].

Some constant hurdles always exist in the process of data collection which hinder the operations and impact the decision-making. Some information sources are inconsistent, so cannot rely on them. Disruptions in the data accessibility, schema modifications, and the massive amount of data can cause substantial amount of downtime, delaying timely insights. Confusing or repeated data makes the problem worse, messing up the accuracy of reports and machine learning models. Other than these technical issues, finding the required and significant data from the ocean of information and dealing with them pose another set of challenges[30]. Dealing with these difficulties needs careful planning, proper monitoring, and innovative ideas to get useful information.

➢ *Various Web Scraping Python Packages*

The diagram [Fig.3] lists the powerful python packages used in this study to test the web scraping.

Fig. 3: Python packages used to do web scraping

➢ *About to Choosing the Scrapy Library*

The powerful python package, Scrapy, that is used to scrap news involves various features such as Item pipeline, middleware, pipeline, spider and crawler. Pipeline helps clean the extracted contents, automatic throttling which enables local machines to prevent the overloading of servers and accelerates the extraction process seamlessly[4]. It has also got ease of extensibility which allows developers to inject user defined functionalities into middle wares and pipeline of scrapy projects. The script utilises the Scrapy Spider class to define the scraping logic, extracting relevant information from the HTML structure of the target website [10]. Upon completion, it uploads the CSV file to an Amazon S3 bucket using the Boto3 library for AWS interactions. The script is structured to run as a standalone application and provides an efficient and automated solution for daily news scraping with data storage on AWS S3.

➢ *How did we Overcome this Data Preprocessing in this Study?*

While developing the news article summarization study, the very first step of the study is to scrap the data from different sources using the python package, Scrapy. The scrapping process allowed us to gather a diverse range of news articles for subsequent summarization [25]. The collected data was then meticulously stored in a structured manner using the ultimate scraping tool Scrapy and this forms the cornerstone of the goal which we are trying to achieve.

Cleaning the raw data has always been very challenging and it is no exception to this study as well. The LLM has to understand the context of the text we are passing into the model, only then it will be able to create the short summary from the big article. In order to make the LLM understand the context, it is crucial to remove all anomalies that are present in the scraped data, because anomalies in the data can completely flip the meaning of the entire content, which may result in poor summarisation. So, the removal of anomalies has to be performed every time the model is executed. The problem associated with this is that the model may face unique anomalies all the time, so, the model must be able to rectify the type of unwanted contents such as emojis, hashtags etc.. A highly sophisticated pipeline, which includes tasks like word-tokenization, spell checking, ambiguity removal, etc, has been implemented to handle this situation [20, 25].

The key to the success of this study is the model that we have chosen, which is T5 and was trained by Google with a huge number of parameters. This T5 model is not only suitable for summarisation but also for many other operations such as translation, Question answering, Sentiment Analysis and many more. Significant amount of fine tuning is was needed to leverage its benefits in order to achieve our goal. During the fine-tuning, some of the parameters of the original model will be replaced by the new parameters which are learnt, this will allow our LLM learn the relationship that exists in our content. This will allow the LLM to generate

the most appropriate summary about the article that is passed into the model [24, 18].

In nutshell, the study involves three steps. The first step is to scrap the news from online platforms and the second stage is to preprocess the data scraped from various sources. The final step is to pass the cleaned textual data into the T5 model, where the big data is shortened in the form summary. Pioneering success in news article summarization, this method masterfully navigates data variability and noise. It showcases the power of web extracting, data preprocessing, and cutting-edge transformer models in NLP studies.

Essence of news articles into concise formats. Key objectives include condensation, information extraction, and enhanced comprehension, facilitating quicker understanding of the article's content. [3]

➢ *Identification of Key Sentences for Summarization:*

Identifying key sentences that encapsulate the main points, events, or insights of a news article is critical for effective summarization. NER, Stop Words, Stemming and Lemmatization fundamental blocks of key sentence identification and NER perform a key role in the news extraction process by identifying and classifying entities such as person names, organisations, and locations [33, 31].

➢ *Encoding Techniques in NLP*

Encoding techniques like One-Hot Encoding and Word Embeddings transform raw textual data into a structured format, enabling machine learning models to process and understand the text effectively.

➢ *Feature Extraction in NLP*

Machines understand human languages using mathematical and statistical calculations. Some of them are TFIDF, word embeddings and positional embedding. These methods convert words into vectors that have magnitude and direction, where the angle and position of the words represent intensity, polarity, sensitivity. This helps machine understand and interpret content passed into the LLM [33,29,26].

D. Model Building

➢ *Large Language Models*

These days LLMs are very popular and they have played a major role in the tasks like NLP and in the field of ai the history of this begins in the 1960s [13]. with the help of Eliza the MIT researcherJoseph Weizenbaum developed chatbot, This chatbot generating responses similar to the human responses although basic utilised pattern recognition to engage in simulated conversations by transforming inputs given by the user into questions and it generates output based on predefined rules despite its imperfections eliza laid the groundwork for subsequent research in NLP and the evolution of more LLMs.

Many innovations have propelled the trajectory of LLMs, the arrival of LSTM networks revolutionised in 199. Its neural network capabilities enable the construction of deeper and more intricate architectures, which are capable of handling the substantial data stanford's core NLP suite introduced in the year of 2010, providing researchers with essential tools and algorithms for addressing intricate NLP assignments which are sentiment analysis and NER. A transformative moment occurred in 2011, with the launch of Google brain offering researchers access to robust computing resources datasets and advanced features like word embeddings. This facilitated improved situations comprehension by using NLP Google brains contributions paved the way for most advancements including the introduction of transformer models in 2017.

In the early stages diminutive models like GPT-2 models, with great abilities of neural networks showcasing their potentials to comprehend and craft coherent and it generates responses that are similar to the human text. The game-changer was BERT in 2018. With the help of transformers, it has the advantages for long-range text dependencies boosting comprehension. The involvement of transformer architecture further nurtured the evolution of more expansive and refined LLMs, this progress is epitomised by the creation of Open-AIs GPT-3 models which has seamlessly assumed a foundational role in transformative applications like ChatGPT.

Recent years have seen notable contributions to LLM advancement with initiatives like hugging face and bard creating user-friendly frameworks and tools. These platforms empower researchers and developers to construct their LLMs further fostering the evolution and the easy accessibility of this transformative technology. Ongoing research and development continue to refine LLMs exploring novel architectures optimization techniques and addressing the computational challenges associated with training and deploying such colossal models however ethical concerns regarding bias and explain ability persist nevertheless, LLMs hold immense of abilities to collaborate with humans and democratise AI [16].

➢ *Different Types of LLM*

- **Foundation Models**, recently AI is very popularised at stanford. The LLMs are built on large amounts of raw data; it serves the things of more specialised models and the uses are notable illustration is Open-AI's GPT-3 model which epitomises the role of a foundation model.
- **Pre-training models,** like GPT-3, GPT-3.5, T5, and XLNet. These are built on an extreme amount of datasets. With the help of the models, we can easily identify the hidden parameters and structures. These are developed in an efficient way that can produce appropriate and grammatically accurate text across a spectrum of topics, these serve as initial foundations for reliable training and fine-tunings for specific applications [14, 25].

- **Fine-Tuned or Domain-Specific Models,** Fine tuning is like we can train or with the help of hyper parameters we can change our model as per our requirements suppose model training in the sense train the model with our own datasets which helps to the model to learn the required patterns and the information which is present in our data it will helps the model to focus on the specific domains such as medical or technical tasks [14, 36, 25].
- **Autoregressive Language Models**, typified by GPT, construct sentences word-by-word, proceeding from left to right. These models predict the succeeding word in a sequence by considering all preceding words.
- **Bidirectional Language Models**, BERT models exemplified by BERT itself stand out in sentence comprehension. As they adeptly analyse both preceding and succeeding words around a given word. Elevating overall understanding this dual-context consideration significantly enhances its effectiveness in identifying the nuanced meanings within a sentence.
- **Zero-Shot Learning Models**, The Open-AI's GPT models give better insights without giving a proper training for a given task. They rely on broad general-purpose knowledge acquired during initial training [14, 36, 1].

➢ *Applications of LLM*

The LLMs are revolutionised in various aspects of AI. In the case of conversational AI. where, these exhibit impressive abilities to give the output as contexts. which are similar to the human-like responses and widespread uses of bots other virtual assistants underscores their improving influence. A notable example is Google's meena, a formidable LLM that outperformed other dialogue agents in sensibleness and response quality beyond dialogue. LLMs enhance textual data creation showcasing prowess in generating articles about news, product descriptions, and creative writing, they excel on sentiment, handling customer feedback, brand monitoring and pandemic-related sentiment tracking applications, as explained by GPT and other models. LLMs exemplified by Google translate successful use of NMT since 2016. Break down the difficulties in the languages, fostering global communication innovations, like ChatGPT code interpreter plugins simplify communication between humans and machines making application development effortless for all LLMs. Their versatility continues to significantly impact various domains in artificial intelligence [16, 17].

➢ *What are all the Models that we Experimented With?*

We have mainly experimented with 5 different models Pegasus, T5, BERT, BART, RoBERTA. Observed the summaries generated using these models by varying their hyper parameters. [Table.1]

➢ *What are all the Hyperparameters?*

We have fine-tuned the model by varying its hyperparameters.

➢ *Parameters that Control the Length of the Output*

- **max_length** — how many lines of summary has to be generated by LLM
- **max_new_tokens** — maximum number of words must be present in the generated summary.
- **min_length** — minimum length of the summary that has to be generated irrespective of input size.
- **min_new_tokens** —minimum number of words that has to be there in the generated summary
- **early_stopping** — Early stopping prevents a model from overfitting. If we allow the model to be trained until completion there is a risk of it remembering the hidden patterns present in data which leads to overfitting potentially it gives 100% accuracy this may not generalise well to new data.

➢ *Parameters that Control the Generation Strategy Used*

- **do_sample** — Usually new word prediction depends on high probability. It is also called greedy decoding. When a dosample is set to false, subsequent tokens are selected randomly from their probability distribution, sample conversely if dosample is set to true, the tokens are chosen in a different manner.
- **num_beams** —Number of beams to be chosen from which next tokens are selected when it is 1. when it is greater than 1 then the required number of beams are selected.
- **num_beam_groups** —This num_beam_groups parameter on sequence diversity in beam search is notable. Its implementation divides beams into specific groups during sequence generation fostering diverse outputs by exploring distinct areas of the search space for example when num_beam_groups is set to 2, with num_beams as 4, it establishes two groups each with two beams promoting unique and varied sequences.
- **eos_token_id** —It marks the end of a sequence at the time of decoding in some cases multiple end-of-sequence tokens may be used.

➢ *Hyper-Parameter Tuning*

Table 1: F1 Score with Manipulated Hyperparameter after a Lot of Hyper-Parameter Tuning

Model Name	temprature	top_k	top_p	length_penalty	no_reapeat_ngram_size	F1 score
T5	1.0	1	0.2	0.2	0	0.92
BART	1.0	4	0.1	0.4	2	0.84
BERT	1.0	3	0.5	0.1	4	0.67
RoBERTa	1.0	5	0.2	0.5	1	0.62
Pegasus	1.0	2	0.4	0.3	3	0.74

➢ *Why are we Choosing T5?*

We have selected the T5 model as it perfectly aligns with our core objective of summarising multiple articles extracted daily. Well Known for its ability in abstractive text summarization, the T5 model excels in capturing contextual relationships and crafting coherent summaries, making it an optimal choice for our summarization tasks [20]. Its ability to swiftly comprehend complex texts, coupled with its ease of fine-tuning for various tasks, further solidifies its position as our preferred model for news article summarization. [22]

E. Model Evaluation Metrics

Various Evaluation metrics [2] available to measure the accuracy of Summary Generated by an LLM. Comparison of all models and their accuracy metrics were shown in table[Tab.2]

- ROUGE – Recall Oriented Under Study for Gisting Evaluation [13, 1].
- BLEU – Bilingual Evaluation Understudy Used for text Translation.
- METEOR - Metric for Evaluation of Translation with Explicit Ordering
- PerPlexity
- Semantic Similarity Metrics

Even though we have a lot of accuracy metrics, we will proceed with the ROUGE Score which is used across the globe to measure the quality of generated text. The interpretation is also very intuitive, so that we will use the ROUGE score to measure the accuracy of the generated summary [12].

➢ *Accuracy Calculation:*

$$Recall = \frac{Unigram\ Matches}{Unigrams\ in\ Reference}$$

$$Precision = \frac{Unigram\ Matches}{Unigrams\ in\ Output}$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

➢ *Recall Oriented Understudy for Gisting Evaluation based on Unigram.*

Let's Consider the following example.

Original Text / Reference Text: "**It is Cold Outside"**

Generated Text: **"It is very Cold Outside"**

Total number of unigrams in the original text and generated text are 4 and 5 respectively.

Matches (Common Tokens present in both Original Text and Generated Text): There are 4 common tokens. Now let's apply this in the above formula to calculate Recall, Precision and F1 Score.

$$Recall = \frac{4}{4} = 1.00$$

$$Precision = \frac{4}{5} = 0.8$$

$$F1\ Score = 2 * \frac{0.8 * 1}{1.8} = 0.888$$

➢ *ROUGE based on Bigram.*

Let's Consider the same example.

Matches (Common Tokens present in both Original Text and Generated Text): There are 2 common tokens. Now let's apply this in the above formula to calculate Recall, Precision and F1 Score.

$$Recall = \frac{2}{3} = 0.67$$

$$Precision = \frac{2}{4} = 0.5$$

$$F1\ Score = 2 * \frac{0.67 * 0.5}{1.17} = 0.57$$

In the same way we can keep on calculating the recall, precision and f1 score for different word grams. Instead of doing this type of calculation, we can go ahead with Longest Common Subsequence present in both generated text and original text.

➢ *Longest Common Subsequence*

Original Text / Reference Text: "**It is Cold Outside"**

Generated Text: **"It is very Cold Outside"**

In our example there are two matching longest common sequences and those are

"It is" and **"Cold Outside"**. Let's calculate the accuracy based on this now.

$$Recall = \frac{LCS(Generated, Original)}{Unigrams\ in\ Reference}$$

$$Recall = \frac{2}{4} = 0.5$$

$$Precision = \frac{LCS(Generated, Original)}{Unigrams\ in\ Output}$$

$$Recall = \frac{2}{5} = 0.4$$

$$F1\ Score = 2 * \frac{0.5 * 0.4}{0.9} = 0.44$$

➢ *Comparison of Different Models that we have Experimented*

Table 2: Comparison of Different Models Tested in this Study

Model Name	Recall	Precision	F1 Score
T5	0.94	0.92	0.929892473
BART	0.8	0.9	0.847058824
BERT	0.7	0.65	0.674074074
RoBERTa	0.67	0.59	0.627460317
Pegasus	0.8	0.7	0.746666667

F. Deployment:

➢ *Reason Behind the Selection of T5 as Final Model:*

First and foremost, T5 has showcased exemplary output across multiple evaluation metrics, including F1 score, recall and precision. With a remarkable True Positive Rate of 0.94 and an impressive precision of 0.92, the model, T5, consistently outperformed its counterparts, demonstrating its robust capacity of capturing pertinent details from the source text and succinctly articulate it in the summary. This exceptional performance is indicative of T5's inherent capability to grasp the essence of the input text and distil it into concise and informative summaries.

Furthermore, T5's versatility and adaptability make it an ideal choice for deployment in diverse text summarization tasks. In addition to its superior performance, T5 boasts impressive scalability and efficiency, allowing for rapid inference and deployment in production environments. By harnessing the capabilities of T5, we can confidently generate high-quality summaries that encapsulate the essence of the input text, thereby advancing the field of natural language processing and unlocking new opportunities for knowledge extraction and dissemination.

➢ *Streamlit Platform:*

Streamlit is super easy and one of the best platforms for machine learning models deployment. Its simplicity of usage makes it the most preferable choice to research associates and machine learning engineers. Even from the perspective of users, It is an unbeatable user friendly platform. The support system OS Streamlit is another strong reason behind the selection of this platform, which involves support like a wide range of graphs, highly customizable background theme and many more. Users were given multiple choices to choose the website from which extraction should happen. All the user has to do is to select the website, the rest of the process is taken care of by an algorithm that has been built during this study. The screenshot that depicts the model deployment has been given in the figure below[Fig.4].

ISSN No:-2456-2165

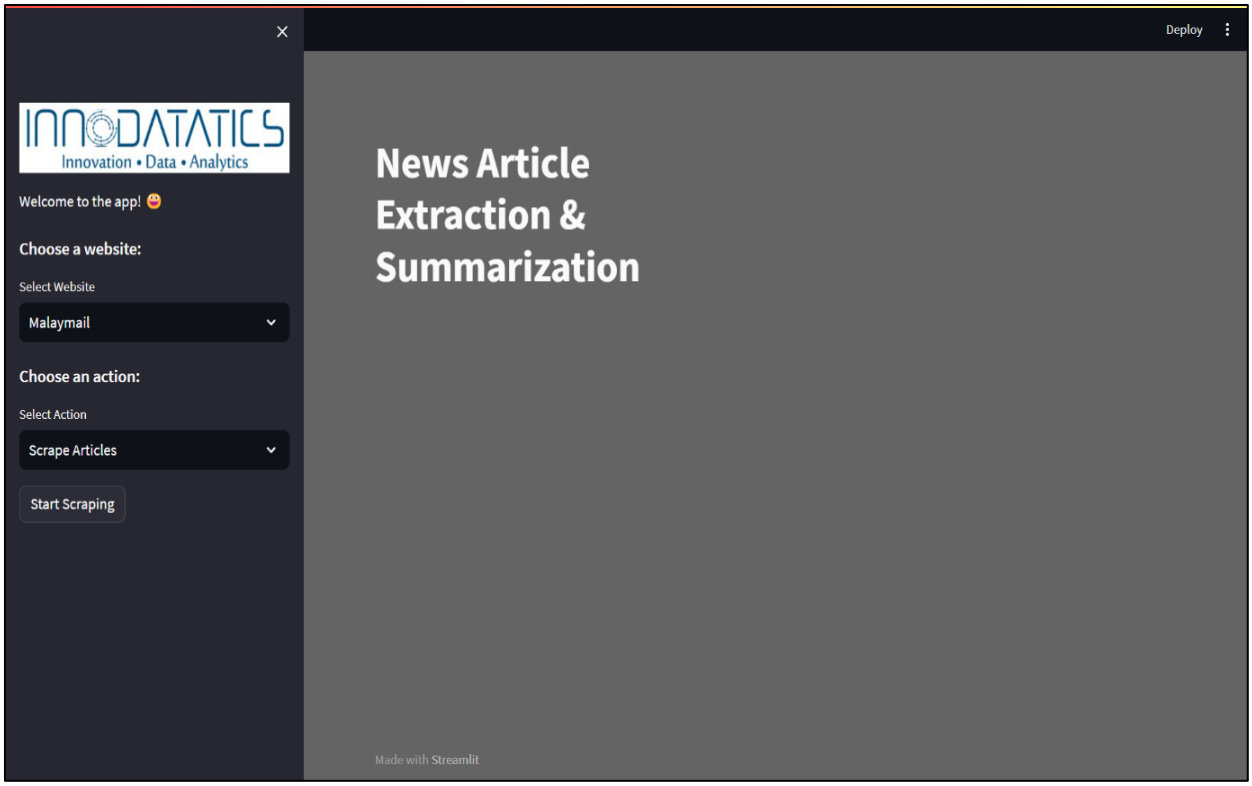

Fig. 4: Model Deployment Screenshot

In this study, the deployment was done successfully using Streamlit. The Figure [Fig.5] shows the format of the final output screen which is ready to be consumed by stakeholders and the final output is stored in AWS S3 bucket[Fig.5].

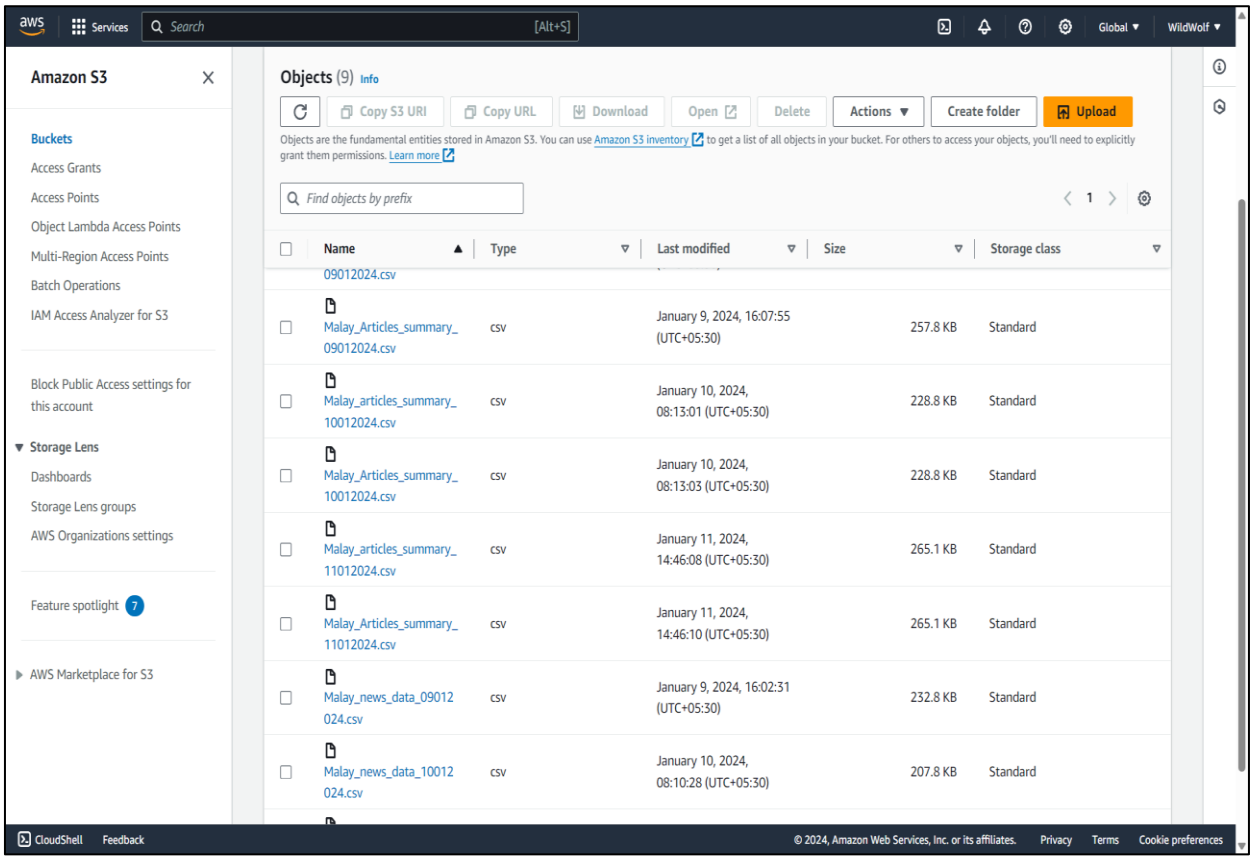

Fig. 5: Summarised Article Stored in S3 Bucket

III. RESULT AND DISCUSSION

After a prolonged study of the performance of different LLMs, it is found that practical results are quite different from that of theoretical calculations. The expected result is 100% accuracy, but in reality, what was achievable is 90% - 97% of accuracy based on hyperparameters used. It is true that a dedicated summarization can be done only through proper fine tuning of the model using PEFT (Parameter Efficient Fine Tuning), LoRA (Low Rank Adaption for LLM), and Soft Prompting. We humans understand language in different ways, whereas to make machines understand our language, a lot of statistical and mathematical calculations are used which finally results in probability distribution of next word. By applying some statistical sampling techniques on this probability distribution to choose the next word, a creative summary can be created. But, as per the study we conducted, the creative summaries contextually and intrinsically represent quite different meanings, from the content that was passed into the model. It is better to go ahead with full fine tuning methods to tune the LLM to get the most precise and coherent summaries.

IV. CONCLUSION

The idea of incorporating LLM in the field of media would result in seamless productivity increment in day-to-day work. This could also increase the newsreader base if it is properly implemented in any print or online media company. In reality, the news collection would take at least 3 - 4 hours of time if it is manually done, whereas the same task will be done 5-10 mins if it is handled by the AI. The manual intervention of humans can be reduced at least by 20% - 40% not only in terms of news collection but also in the view of summarization or creation of phrases or titles for articles. Well, this may also lead to job loss if implemented with the full potential of AI.

This powerful triumvirate not only promises to streamline the labour-intensive aspects of news extraction but also introduces a level of sophistication and accuracy previously unattainable through manual efforts alone. As the media continues to evolve, embracing these advancements is essential for ensuring that audiences worldwide receive timely, important, hot, and meticulously articulated news content tailored to their diverse tastes and preferences.

By automating tasks that were once reliant on manual input, media organisations can allocate resources more strategically, fostering greater productivity and innovation. Moreover, the ability of LLM models to analyse vast datasets and discern the significance of news items with unparalleled accuracy promises to elevate the quality of news reporting, fostering a more informed and engaged global audience. As we embark on this journey towards a future where news processing is both intelligent and automated, the collective effect of AI, LLM models, and Python libraries emerges as a beacon of innovation, heralding a new era of transformative possibilities in the field of media and communications.

REFERENCES

[1]. Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., & Hashimoto, T. B. (2023). Benchmarking large language models for news summarization. arXiv preprint arXiv:2301.13848. DOI: 10.48550/arXiv.2301.13848.

[2]. Wu, N., Gong, M., Shou, L., Liang, S., & Jiang, D. (2023). Large language models are diverse role-players for summarization evaluation. *arXiv preprint arXiv:2303.15078*.

[3]. Pu, X., Gao, M., & Wan, X. (2023). Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.

[4]. Myers, Daniel & McGuffee, James. (2015). Choosing Scrapy. Journal of Computing Sciences in Colleges. 31. 83-89.

[5]. Dimarco, M. H. (2023). LLM-based Comment Summarization and Topic Matching for Videos.

[6]. Sethi, Prakhar & Sonawane, Sameer & Khanwalker, Saumitra & Keskar, R.. (2017). Automatic text summarization of news articles. 23-29. 10.1109/BID.2017.8336568.

[7]. Daud, Shahzada & Ullah, Muti & Rehman, Amjad & Saba, Tanzila & Damaševičius, RoBERTas & Sattar, Abdul. (2023). Topic Classification of Online News Articles Using Optimized Machine Learning Models. Computers. 12. 16. 10.3390/computers12010016.

[8]. Hossain, Mohammad & Sarkar, Soikot & Rahman, Moqsadur. (2020). Different Machine Learning based Approaches of Baseline and Deep Learning Models for Bengali News Categorization. International Journal of Computer Applications. 176. 10-16.

[9]. Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T., & Wang, G. (2023). Text Classification via Large Language Models. *arXiv preprint arXiv:2305.08377*.

[10]. Steinberger, Josef & Jezek, Karel. (2009). Text Summarization: An Old Challenge and New Approaches. 10.1007/978-3-642-01091-0_6.

[11]. Atodiresei, C., Tanaselea, A., & Iftene, A. (2018). Identifying Fake News and Fake Users on Twitter. *International Conference on Knowledge-Based Intelligent Information & Engineering Systems*.

[12]. Ahmed, J., & Ahmed, M. (2021). ONLINE NEWS CLASSIFICATION USING MACHINE LEARNING TECHNIQUES. *IIUM Engineering Journal*, *22*(2), 210–225.

[13]. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*.

[14]. Goyal, T., Li, J. J., & Durrett, G. (2022). News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

[15]. Zhuang, B., Liu, J., Pan, Z., He, H., Weng, Y., & Shen, C. (2023). A survey on efficient training of transformers. *arXiv preprint arXiv:2302.01107*.

[16]. Abburi, H., Suesserman, M., Pudota, N., Veeramani, B., Bowen, E., & Bhattacharya, S. (2023). Generative ai text classification using ensemble LLM approaches. *arXiv preprint arXiv:2309.07755*.

[17]. Li, Z., Zhu, H., Lu, Z., & Yin, M. (2023). Synthetic data generation with large language models for text classification: Potential and limitations. *arXiv preprint arXiv:2310.07849*.

[18]. Lyu, Zhihan & Ota, Kaoru & Lloret, Jaime & Xiang, Wei & Bellavista, Paolo. (2022). Complexity Problems Handled by Advanced Computer Simulation Technology in Smart Cities 2021. Complexity. 2022. 10.1155/2022/9847249.

[19]. Mridha, M. F., Keya, A. J., Hamid, M. A., Monowar, M. M., & Rahman, M. S. (2021). A comprehensive review on fake news detection with deep learning. *IEEE Access*, *9*, 156151-156170.

[20]. Malagi, S. S., & Rachana, A. D. (2020). An overview of automatic text summarization techniques. In *International journal of engineering research and technology (IJERT), Published by, www. ijert. org NCAIT—2020 Conference proceedings* (Vol. 8, No. 15).

[21]. NLP Summarization: Abstractive Neural Headline Generation Over A News Articles Corpus M. Erraki, M. Youssfi, A. Daaif and O. Bouattane, "NLP Summarization: Abstractive Neural Headline Generation Over A News Articles Corpus," *2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS)*, Fez, Morocco, 2020, pp. 1-6, doi: 10.1109/ICDS50568.2020.9268776.

[22]. Automated News Summarization Using Transformers Gupta, A., Chugh, D., Anjum, & Katarya, R. (2022). Automated news summarization using transformers. In *Sustainable Advanced Computing: Select Proceedings of ICSAC 2021* (pp. 249-259). Singapore: Springer Singapore.

[23]. Rananavare, L., & Subba Reddy, P. V. (2018). Automatic News Article Summarization. *International Journal of Computer Sciences and Engineering*, 6(2), 230-237. DOI:10.26438/ijcse/v6i2.230237

[24]. NLP based Machine Learning Approaches for Text Summarization Rahul, S. Adhikari and Monika, "NLP based Machine Learning Approaches for Text Summarization," *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2020, pp. 535-538, doi: 10.1109/ICCMC48092.2020.ICCMC-00099.

[25]. A Detail Survey on Automatic Text Summarization Sajjan, R., & Shinde, M. (2019). A detailed survey on automatic text summarization. *Int J Comput Sci Eng*, *7*, 991-998.

[26]. A Survey on Fake News and Rumour Detection Techniques Bondielli, A., & Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. *Information Sciences*, *497*, 38-55.

[27]. NLP based Intelligent News Search Engine using Information Extraction from e-Newspapers Kanakaraj, M., & Kamath, S. S. (2014, December). NLP based intelligent news search engine using information extraction from e-newspapers. In *2014 IEEE International Conference on Computational Intelligence and Computing Research* (pp. 1-5). IEEE.

[28]. NLP based Text Summarization Techniques for News Articles: Approaches and Challenges

[29]. Tarannum, S., Sonar, P., & Khairnar, A. A. K. (2021). NLP based Text Summarization Techniques for News Articles: Approaches and Challenges.

[30]. El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, *165*, 113679. https://doi.org/10.1016/j.eswa.2020.113679

[31]. Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, *1*(1), 4-20.

[32]. Gupta, A., Singh, A., & Shankhwar, A. K. A Comparative Analysis of Automatic Text Summarization (ATS) Using Scoring Technique. *ijraset*, *November-2022*.

[33]. Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, *2*(3), 160.

[34]. U Suleymanov and S Rustamov 2018 *IOP Conf. Ser.: Mater. Sci. Eng.* 459 012006 DOI 10.1088/1757-899X/459/1/012006

[35]. Balaji, N., Megha, N., Kumari, D., & Kumar, P. S. (2022). Text Summarization using NLP Technique. In *2022 International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*. DOI: 10.1109/DISCOVER55800.2022.9974823

[36]. Lloret, E. (2008). TEXT SUMMARIZATION : AN OVERVIEW ∗.

[37]. Abualigah, L., Bashabsheh, M. Q., Alabool, H., & Shehab, M. (2020). Text summarization: a brief review. *Recent Advances in NLP: the case of Arabic language*, 1-15.

[38]. Das, A., Srihari, K., Pasupuleti, S., Kumar, I., & Depuru, B. K. Maximising Operational Uptime: A Strategic Approach to Mitigate Unplanned Machine Downtime and Boost Productivity using Machine Learning Techniques. DOI: 10.5281/zenodo.10477279