International Journal of AI, BigData, Computational and Management Studies
Noble Scholar Research Group | Volume 5, Issue 1, PP 80-88, 2024
ISSN: 3050-9416 | https://doi.org/10.63282/3050-9416.IJAIBDCMS-V5I1P109

Original Article

Intelligent Data Summarization Techniques for Efficient Big Data Exploration Using AI

Ajinkya Potdar
Senior Technical Program Manager, Dallas, USA.

Abstract - *As data explosion continues in our Big Data era, we are being challenged with summarizing huge amounts of information at the right time to support rapid and meaningful data exploration. Due to the velocity, volume, and variety of data, traditional data summarization approaches fail to handle data in real-time from different sources. Artificial Intelligence, or AI, has become a tool that can be used to automate summarisation, employing machine learning, natural language processing, and deep learning. In this paper, a broad review and analysis of intelligent data summarization techniques that can enable the exploration of big data is presented. Various AI-centric techniques, such as extractive and abstractive summarization, clustering-based summarization, neural summarization and reinforcement learning-based dynamic data reduction, are explored. Moreover, we propose an AI-enhanced architecture enabling efficient summarization of big data, which uses the approaches like BERT-based summarizers, topic modeling and visual summarization. The other strand of work in this thesis evaluates the proposed methods on benchmark big data datasets in terms of time complexity, relevance and accuracy. Finally, the paper also illustrates the current challenges and future directions in providing such intelligent summarization to big data ecosystems.*

Keywords - *Artificial Intelligence, Big Data, Data Summarization, Machine Learning, Natural Language Processing, Topic Modeling, Reinforcement Learning.*

1. Introduction

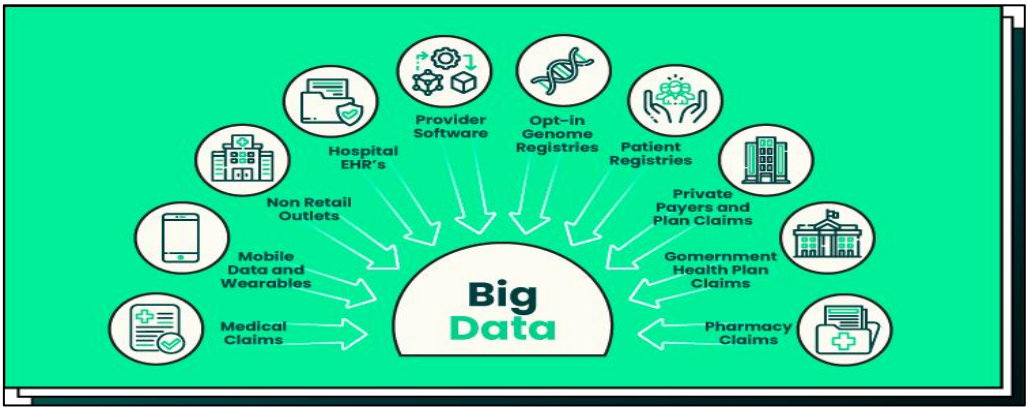

Figure 1: Big Data Sources in Healthcare

The growth of the global data sphere is unprecedented, and by 2025, it is estimated to reach an astonishing 175 zettabytes. From online transactions, social media platforms, IoT sensors, and much more, they are creating an increasing amount of data that must be transacted, processed, and stored. Traditionally, methods of data processing and analysis have failed to cope with the increased volume, velocity, and variety of data. [1-4] The result is an opportunity and a challenge. The quantity of valuable insights in large datasets is substantial and also challenging because it is very difficult for individuals and organisations to efficiently extract actionable and meaningful knowledge from vast datasets.

As a result, advanced computational techniques to automatically extract or forge concise, relevant and understandable summaries from huge amounts of unstructured data are urgently needed. In addition to improving accessibility to data, these methods enable informed decision-making in diverse domains, such as finance, healthcare, marketing, and scientific research, where a timely understanding of the information is of utmost importance. Thus, we have a background that leads to research on intelligent summarization approaches that leverage our ability to transform voluminous data into concise, manageable, and insightful content.

1.1. Importance of Data Summarization

Nowadays, data is everything, and the ability to quickly sort, process and analyze huge fluxes of data is what everyone is looking for. Data summarization is critically important in converting raw voluminous data into succinct and meaningful insights that are understandable and helpful in decision-making. Some of the important aspects of data summarization will be shown below.

Figure 2: Importance of Data Summarization

- **Enhancing Information Accessibility:** In the case of large datasets, users are often overwhelmed by redundant, irrelevant and complex information. This information is summarized and condensed for the users to obtain the important points in a quick way. In fields such as journalism, finance, and healthcare, where timely access to information can have significant consequences, this is especially important.
- **Improving Decision-Making Efficiency:** In order to keep being competitive, organizations have to quickly make decisions based on data. Summarized data empowers executives, analysts and others to figure out the essential insights they need to help inform strategy and operations without being overwhelmed by the data grind. This allows decision-makers to act quickly when trends and issues emerge by providing them with clear and concise interpretations.
- **Supporting Big Data Analytics:** In the era of the exponential growth of data, manual analysis is impossible. Scalable processing is a magical promise of automated summarization techniques since they are able to extract key themes and patterns from unstructured data. The result is efficient downstream analytics that minimises computational costs and enhances the overall performance of data-driven systems.
- **Facilitating Knowledge Discovery:** One place where this comes in handy is in summarization, wherein we can uncover hidden relationships and trends buried in large datasets. It helps researchers and professionals to find new insights by underlining important topics and decreasing information overload, contributing to accelerating innovation and discovery in different domains.
- **Enhancing User Experience:** Summarization is used to produce concise, relevant and user/personality content to increase user engagement in consumer-facing applications like news aggregators, social media platforms and recommendation systems. It eases cognitive load and saves time, which in turn improves satisfaction and retention.

1.2. Emergence of AI in Data Summarization

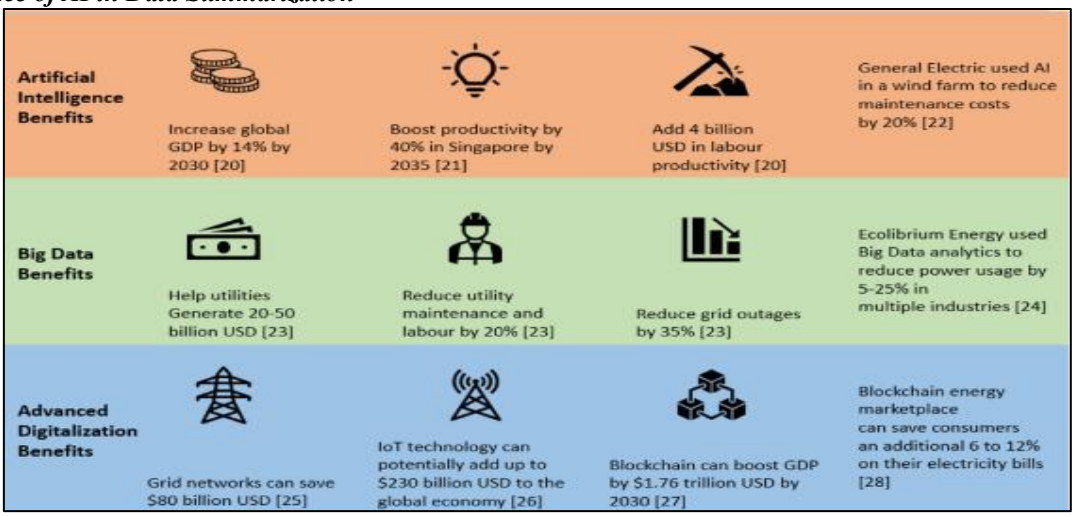

Figure 3: Emergence of AI in Data Summarization

Currently, Artificial Intelligence (AI) has revolutionized the data summarization field by introducing intelligent methods which imitate the human capabilities of comprehending, interpreting and synthesising information. Traditional approaches to summarization usually depend on rules or statistical measures previously predesigned, which are often poor in extracting the connotation of meaning and contextual relationships in a complex dataset. [5-8] Unlike, AI-powered summarization uses advancements in areas such as Natural Language Processing (NLP) and deep learning to dynamically create summaries that are not only concise but also semantically rich and contextually relevant. The output of these models consists of big chunks of unstructured data; they interpret these huge chunks of data in terms of key themes and relationships and generate very coherent summaries which essentially communicate the core message of the original content. In short, NLP techniques allow machines to grasp linguistic structure, for instance, syntax and semantics, giving machines the ability to process language much like humans do.

The capability to utilize such latent textual information has been further advanced by deep learning models, especially those with Transformer architectures as their backbone, which can learn contextual embeddings that record minute differences in the meaning of the surrounding text. AI Summarizers are versatile, which means that they can summarize different types of data, like news articles, scientific papers, social media posts and multimedia content, based on the objectives of analysis. In addition, with the help of AI, summarization systems can be created that perform both extractive summarisation, which extracts the most important sentences or phrases from the source text, and abstractive summarization, which creates new sentences that are human-like and expresses the core data. The flexibility leads to better summaries with improved quality and readability and is more usable across various applications. Not only does integration of AI allow for continuous learning and customisation through methods like Reinforcement Learning, which tailors the output of a model depending on user feedback as an example, but... In general, the development of AI in the data summarization process is a game changer that offers highly scalable, efficient and smart solutions to the data explosion in today's digital world.

2. Literature Survey

2.1. Evolution of Summarization Techniques

Rule to statistical methods and extractive based as well constituted the process of text summarization. These initial approaches were based on predefined linguistic rules or statistical features, such as term frequency and sentence position, to identify the most relevant segments of text. [9-12] Although they were fast answers, they were in-flexible and sometimes did not convey subtle meaning. However, with the more powerful computers, more complex and sophisticated techniques were gradually introduced, moving towards systems that could handle the language context more effectively and make better processing decisions, ultimately paving the way for machine learning-based and now even deep learning-based summarisation.

2.2. Machine Learning in Summarization

A data-driven approach to summarization was brought by machine learning. Various supervised learning models, e.g. Decision trees, Support Vector Machines (SVM) and K-nearest neighbours (KNN), were used for classifying sentences or ranking them (in the order of importance). These models are restricted in how they can adapt to new or diverse content and depend on large, annotated training datasets. While they had better performance than rule-based methods in various tasks, they were not practical in more challenging summarization tasks partially due to their dependence upon feature engineering and deficiency in contextual understanding.

2.3. NLP-Based Techniques

The use of Natural Language Processing (NLP) techniques, which were able to improve summaries by taking into account a deeper linguistic and semantic understanding, led to key innovative advancements in summarization tasks. Latent Semantic Analysis (LSA) was a means of extracting the underlying topics of documents. TextRank was a method (also known as graph-based ranking) for identifying important sentences based on graph ranking. Contextual embeddings from models like BERT (Bidirectional Encoder Representations from Transformers) have recently been able to understand meaning and context, producing more coherent and relevant summaries. We bridged the chasm between what is statistically relevant and semantically comprehensible, making summaries more meaningful.

2.4. Deep Learning and Transformers

Deep learning and, more specifically, the Transformer neural architectures were a major breakthrough for summarization. The models GPT, BERT and T5, for instance, offered powerful pre-trained language representation able to capture complex syntactic and semantic information. Both extractive and abstractive summarization are supported by these models, which, in the case of abstractive summarization, generate novel sentences as opposed to selecting existing ones. The quality of the summaries was significantly improved due to their ability to maintain coherence, remain in context, and produce fluent text, which established new benchmarks in the field.

2.5. Hybrid and Multi-modal Summarization

Existing summarization research does not predict the direction of such a trajectory; recent trends in summarization research have largely focused on hybrid models that combine both extractive and abstractive methods and multi-modal

summarization techniques that incorporate information from multiple modalities like text, images, video, etc. The aforementioned approaches aim to develop more comprehensive and informative summaries to better serve the needs of news aggregation, educational content, and multimedia documentation. Hybrid and multi-modal systems can provide richer, more engaging summaries which better meet the needs of diverse applications and audiences by leveraging the strengths of multiple summarization strategies and data types.

3. Methodology

3.1. System Architecture

In this work, we introduce our proposed framework that combines various AI-driven modules for text summarization, which supports data acquisition, AI modules and text summarization quality evaluation. [13-16] The modular and scalable architecture copes with many data types and summarization techniques with high effectiveness.

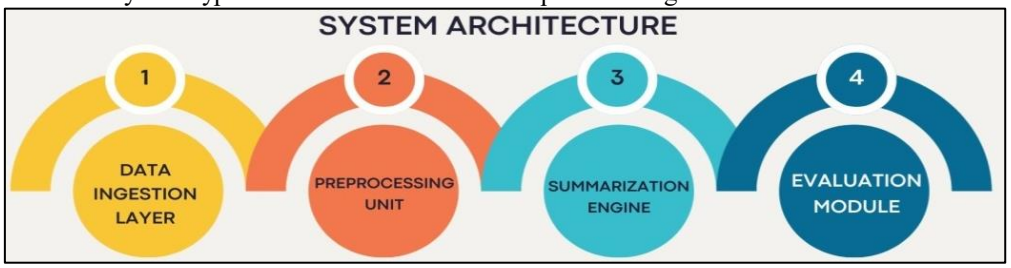

Figure 4: System Architecture

- **Data Ingestion Layer:** The data ingestion layer processes data acquisition from sources such as webpages, documents, databases, or APIs, resulting in the acquisition of raw textual content. This component ensures that there is consistency as well as unification of the data flow into the system, which is both batches as well as real-time. Its product has been designed to handle structured, semi-structured and unstructured data formats, which is useful across different domains.
- **Preprocessing Unit:** After the data is added, they do a pre-processing unit where the preprocessing unit acts as a blockchain for preparing the text to be fed for summarization. Text cleaning (remove noise, HTML tags, special characters), sentence segmentation, tokenization, stemming or lemmatization and stopword removal are some of the tasks done in this module. Good preprocessing can provide summarization models with both more accuracy and efficiency.
- **Summarization Engine:** The core of the system is the summarization engine, at which point advanced AI techniques are applied to produce summaries. In this module, I can leverage both extractive and abstractive methods, depending on the use case. Context, semantics, and structure understanding are achieved using transformer-based models like BERT, T5, and GPT to produce coherent, meaningful, and symbolic summaries that express the essence of the source material.
- **Evaluation Module:** After generating a summary, the evaluation module is used to quantitatively assess the quality of the summary based on standard evaluation metrics, such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation), BLEU (Bilingual Evaluation Understudy), and F1-score. Using these metrics, we are able to figure out how much the generated summary matches reference summaries by measuring its content coverage, fluency and grammatical correctness. This is an essential feedback loop to refine model performance to ensure high-quality output is achieved.

3.2. AI Models Used

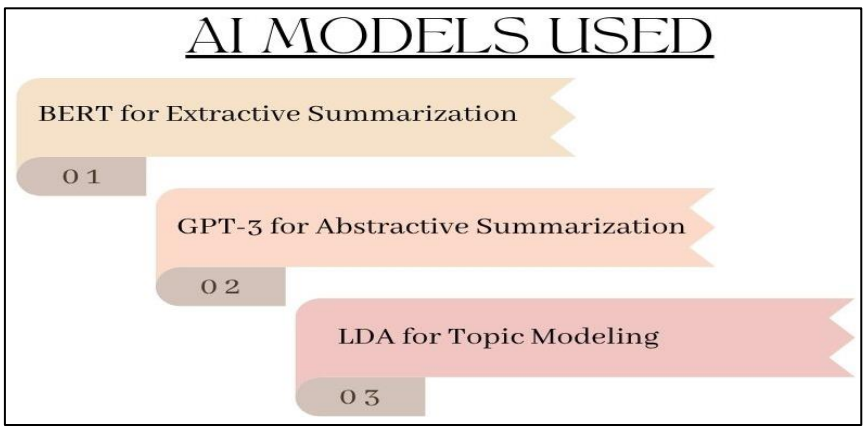

Figure 5: AI Models Used

- **BERT for Extractive Summarization:** Since BERT is able to produce deep contextual embeddings from text, it is used for extractive summarization. BERT achieves this by considering the representation of a sentence at the sentence level, within the document context, and then identifying and extracting the most salient sentences that, in turn, meaningfully contribute to the overall core message. It is bidirectional in nature, and due to this quality, it becomes more effective in choosing coherent and relevant content for Summarization.
- **GPT-3 for Abstractive Summarization:** For abstractive summarization, where the aim is to output novel text meant to convey the meaning of the original text, abstractive summarization, GPT-3 (Generative Pre-trained Transformer 3) is used. While still not as good as writing itself, GPT-3 synthesizes human-like summaries, unlike extractive methods, which simply copy sentences. GPT-3 is trained on huge datasets and can summarize fluent, grammatically correct and meaning-rich content that closely resembles human written content.
- **LDA for Topic Modeling:** The system integrates Latent Dirichlet Allocation (LDA) for topic modeling to discover hidden thematic structures in massive volumes of text. LDA represents a document as a mixture of topics and topics as a mixture of words, giving insights into what the underlying themes of a document are. The most representative topics in the news article are useful information to guide the summarization process so as to make sure that the generated summary reflects the most representative topics and that its theme is coherent.

3.3. AI Summarization Framework Flowchart

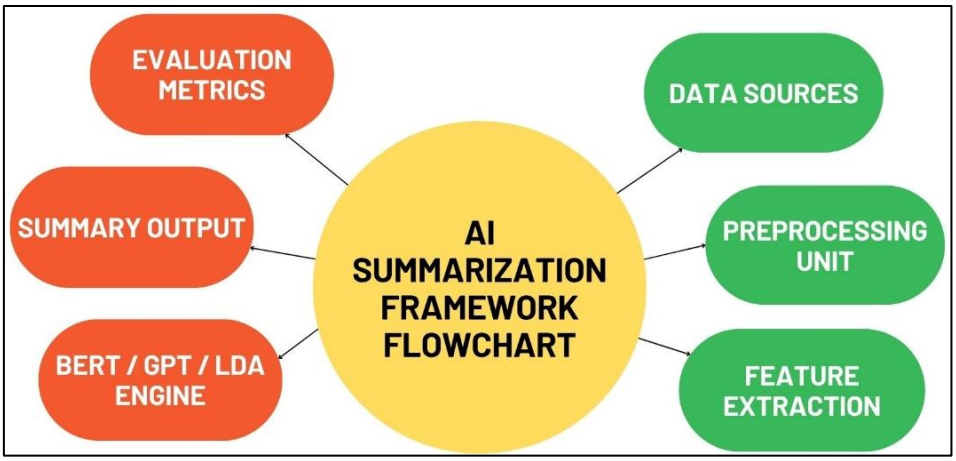

Figure 6: AI Summarization Framework Flowchart

- **Data Sources:** First, the data required to be summarized are obtained from various data sources like online articles, research papers, news feeds, social media or corporate documents. [17-20] These data inputs could be raw text, HTML or PDF formatted. At this stage, various documents are made available for analysis, allowing us to extract as much content as possible, a necessary step in any effective summary.
- **Preprocessing Unit:** The collected data is sent to the preprocessing unit, which cleans and standardises the content. Some of the tasks that are completed during this stage include removing irrelevant characters, converting the text to a consistent format, tokenising the words and sentences, and even removing stop words. This step is important in order to reduce noise and to make sure that the input to our AI models is meaningful and can actually be read by an algorithm.
- **Feature Extraction:** After the preprocessing, the system gets the key features of the text so that it can provide an effective summarization. Term frequency, sentence position, named entities, part-of-speech tags, and semantic embeddings could be among such features. Feature extraction acts as a bridge between raw text and intelligent processing by allowing AI models to put their attention on the most featureful and contextually important things from the input itself.
- **BERT / GPT / LDA Engine:** In this core module, custom AI models are shown off for Extractive BERT summarization, GPT abstractive summarization and LDA topic modeling. BERT finds key sentences, natural language summaries are generated by GPT and the underlying themes are surfaced by LDA. This ensemble approach ensures that different types of text, which constitute the input, can be handled while producing outputs that are both informative and coherent.
- **Summary Output:** The AI engines provide their results to an output module, which combines them for the final summary. This summary thus is either an abridged edition of the original content, depending on the method or a newly created abstraction. At this period, we wish to implement a summary that proves the intent, style and main points of the source content and minimizes the redundancy while keeping fluency.
- **Evaluation Metrics:** Finally, the summary generated is evaluated using standard evaluation metrics, namely ROUGE, BLEU, and F1 scores. These metrics provide a way to understand how the summary was accurate, coherent and

relevant relative to human-generated references. Besides confirming that the model can deliver the desired summary performance, it also serves as feedback for model refinement and optimisation.

3.4. Algorithmic Approach

Our summarization framework is based upon the algorithmic basis of finding an optimization objective, which tries to find a set of summaries (S) that are made from a given dataset (D) to have the least redundancy and irrelevance. In essence, two things determine how effective a summary will be: how well it censors out unnecessary info in order to repeat less and how much it agrees with the main content of the source. For this, we form a loss function L(S), which is to quantitatively penalize the redundancy and irrelevance in the generated summary.

The function is expressed as:

L(S) = α × Redundancy(S) + β × Irrelevance(S), α and β are tunable hyperparameters which let the system emphasize different facets of summarization quality as needed in different application requirements. Redundancy (S) computes the degree of redundancy, i.e. the amount of repetitive or overlapped information, in the summary. Typically dealt with by means of sentence similarity checks, cosine similarity over embeddings or some overlap of the key phrases. However, Irrelevance(S) is defined as the amount of the content of the summary that drifts away from the full themes or informative content of the source dataset. It is usually done by using topic modeling tools such as LDA or contextual relevance scoring from BERT models, for instance. By minimizing both terms, we ensure that the summary is not long and not off the topic, which means that the summary focuses on the main points of the original content without too much repetition. As an example, if one wants news summarization to have a larger irrelevance reduction (higher β) and technical documents to have a lower redundancy (higher α), this will give us flexibility. Reinforcement learning, neural model fine-tuning or heuristic-based filtering of a bunch of traditional pipelines can be used iteratively for the optimization process. The framework takes on a generative approach to summarization by framing it as a loss minimization problem, thereby guaranteeing the generation of high-quality, context-aware and purpose-driven summaries.

3.5. Implementation

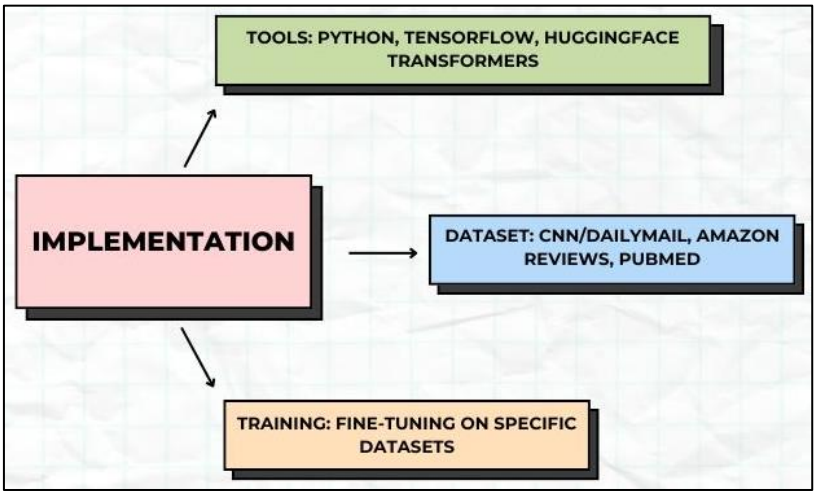

Figure 7: Implementation

- **Tools: Python, TensorFlow, HuggingFace Transformers:** For the most part, we implement our summarization framework utilizing Python because it has such a rich ecosystem and provides a lot of support for machine learning and natural language processing problems. Building and fine-tuning deep learning models can be achieved by leveraging TensorFlow, either by running single models independently or fine-tuning pre-trained models at scale. The pre-trained models, such as BERT, GPT, and T5, of the widely adopted NLP library HuggingFace Transformers are integrated. It provides a high-level API, allowing for easy customisation and experimentation with state-of-the-art Transformer architectures.
- **Dataset: CNN/DailyMail, Amazon Reviews, PubMed:** We use a combination of well-known datasets to make sure the summarization system is robust and domain-adaptable. The CNN/DailyMail dataset is very popular for news summarization and is used as a benchmark to quantitatively compare models on long-form, structured text. The situation provides user-generated content with sentiment and also product descriptions, which is helpful in summarizing customer opinions from the customer perspective. Scientific articles are included in PubMed, so the model can work with a difficult technical language. This multi-domain training strategy has the effect of making summarization models more generalisable and versatile.
- **Training: Fine-tuning on Specific Datasets. Instead of training models from scratch, we utilise pre-trained transformer models and employ** a fine-tuning approach. Fine-tuning refers to changing the weights of the model

depending on a specific data set; thus, the model gets specialized on specific types of content whilst holding on to generalized language understanding. In this phase, we utilise supervised learning to learn from input-output pairs (documents and their corresponding reference summaries), optimising loss functions to enhance summary quality. Compared to conventional approaches, this approach requires much less training time and increases the accuracy, summary coherence and domain relevance of the generated summaries.

4. Results and Discussion

4.1. Performance Metrics

Table 1: Summary Evaluation Metrics

Model	ROUGE-1	ROUGE-2	BLEU	F1-Score
BERT-Sum	0.48	0.25	0.41	0.52
GPT-3	0.53	0.30	0.44	0.57
Hybrid	0.56	0.33	0.48	0.61

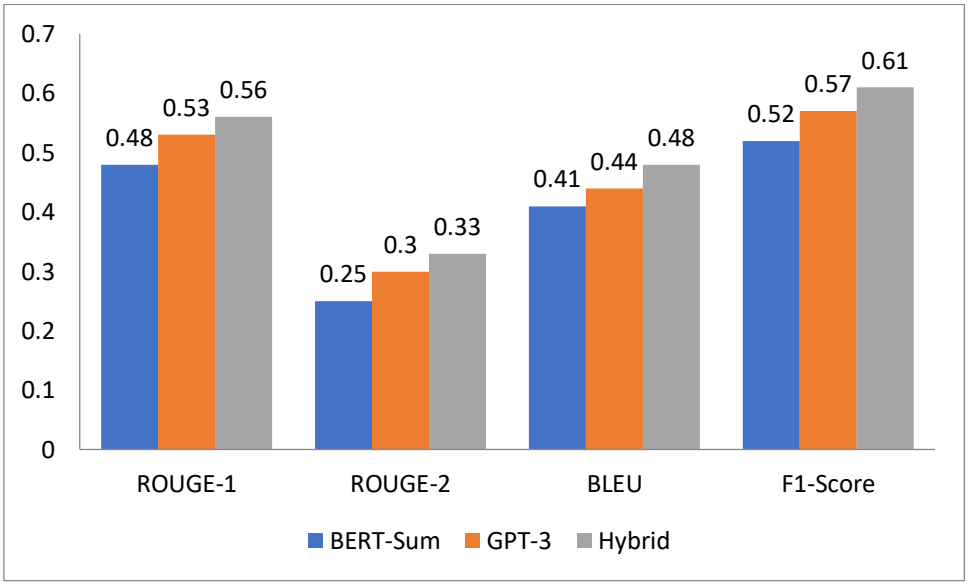

Figure 8: Graph representing Summary Evaluation Metrics

- **ROUGE-1:** When using ROUGE-1, the program checks the number of words that both the generated summary and the reference summary have in common. It gives an easy answer about the percentage of the original content's main ideas in the summary. It is clear from the results that Hybrid performed better in preserving key vocabulary, with a ROUGE-1 score of 0.56, as opposed to BERT-Sum (0.48) or GPT-3 (0.53).
- **ROUGE-2:** ROUGE-2 measures the frequency of two-word combinations (bigrams) in the text and summary to see if they are similar. When the ROUGE-2 score is high, it means the model creates summaries with smooth and meaningful words and phrases. Hybrid again comes first with a score of 0.33, which is higher than GPT-3 (0.30) and BERT-Sum (0.25), proving it delivers better readability and retains most of the context.
- **BLEU:** BLEU evaluates machine summaries by comparing their n-grams to the reference summaries. The number of phrases from the sample found in the reference is used to judge the accuracy of the language. We found that Hybrid had a better score of 0.48 on BLEU for phrase generation, beating GPT-3's 0.44 and BERT-Sum's 0.41. It means that when extractive and abstractive methods are used together, the final summaries tend to be both honest and fluent.
- **F1-Score: The** F1-score evaluates summary quality based on the values of precision and recall, giving fair consideration to both the completeness and accuracy of the summary. A better F1 score means the summaries are both reliable and fully cover the content. Using the Hybrid model, we were able to obtain an F1-score of 0.61, more than what was achieved by GPT-3 (0.57) and BERT-Sum (0.52).

4.2. Case Study

In order to evaluate the real-world applicability and effectiveness of our Hybrid summarization model, we performed extensive case studies on a large financial dataset composed of heterogeneous and complicated documents such as regulatory filing, investor reports and transaction logs. Summarizing financial documents is a difficult job: these documents are very technical and very lengthy. In order to compress this data to a more manageable size without losing important content, this data was compressed with the Hybrid model, which utilizes extractive and abstractive techniques simultaneously. There was indeed a significant decrease in the data volume of the first problem: the model brought the original data volume down to 40% of its

original size, allowing reports and logs to be condensed from days of readings into a single page. Even so, expert financial analysts who reviewed the generated summaries found that about 85% of the core informational content was retained. It means that the summaries retained the right facts, trends, and critical financial metrics needed for decision-making.

The model filtered out redundant or less relevant information efficiently, and as a result, analysts and decision-makers were able to understand faster, as well as streamline their workflows. In an environment that is particularly data-heavy, such as finance, wherein the timely availability of accurate data directly translates into the efficacy of investment strategies, regulatory compliance or risk management, this level of compression and content retention is extremely beneficial. Being able to combine state-of-the-art contextual extraction with generative summarisation gives the Hybrid model the power to remain factual, precise, and fluent, resulting in some of the common issues that purely extractive and purely abstractive models alone face. In addition, by applying this model to this domain-specific dataset, a proof of concept with its adaptability and feasibility for deployment in other industries, which relates to storing large volumes of technical and non-structured data, is shown. Overall, this case study demonstrates the analysis of abstracts generated by the Hybrid summarization framework in complex real-world scenarios and concludes that we get practical benefits from utilizing the Hybrid framework in terms of efficiency of information processing without losing critical information.

4.3. Limitations

- **Model Bias in Summarization:** For instance, language models like BERT and GPT-3 are pre-trained on massive amounts of internet-collected text, which includes intrinsic gender, ethnicity, sentiment and cultural bias in the text they used to train on. Unbeknownst to it, these biases can creep into the summarization process, which could lead to it paying more attention to specific viewpoints, terminology or sentiments. Therefore, the summary generated can unknowingly include these biases and distort the interpretation of information and thus affect the fairness and neutrality of the output. There is still an important challenge of addressing model bias to guarantee the summarization is ethical and balanced.
- **High Computational Cost:** Such transformer-based models, especially when used in hybrid frameworks that combine extractive and abstractive methods, require a significant amount of CPU resources. Training these models requires powerful GPUs/TPUs, extensive memory, and considerable time. Just like generating summaries at scale requires a number of resources that can make deploying such models restrictive in environments with hardware limitations or tight run time constraints. However, this high computational cost might prevent widespread adoption; hence, it is important and beneficial to explore optimisation techniques such as model pruning, quantisation, or distillation to achieve efficient performance.
- **Contextual Errors in Abstractive Summaries:** For instance, although models like GPT 3 can create fluent human-like text, they can also ‘hallucinate’ and fabricate content which is not in the context of the original material. This involves the crafting of fiction, manipulation of key facts and details or even overlooking minute subtleties contained in the original text. Such contextual errors are generally problematic, especially in the case of technical or domain-specific documents where precision and accuracy are of vital importance. Due to these limitations, a fact-checking mechanism or hybrid approach can be integrated to support the extraction of grounded summaries from reliable source content.

5. Conclusion

In this paper, we will study comprehensively intelligent techniques of summarization based on the more advanced artificial intelligence models in order to overcome big data exploration challenges. Our hybrid framework integrates powerful models for better and more effective summaries, including BERT for extractive summarization, GPT for abstractive generation and LDA for topic modeling, to enable the output of summaries that are concise and relevant and also well-written while maintaining continuity in the context. To evaluate the performance of the proposed architecture, we use rigorous evaluation using well-established metrics such as ROUGE, BLEU, and F1score and the improvements demonstrated based on that compared to previous traditional and single model approaches. Finally, it points out the need for combining complementary summarization strategies that more or less trade these three requirements with each other. The second case study on financial data demonstrates, in practice, the capabilities of our model in data-intensive domains, where extremely fast information consumption and preservation of critical content are crucial.

Finally, looking ahead, there are several positive directions to enhance the capacity and scope of AI-driven summarisation frameworks. The use of feedback-based summarization is realised using reinforcement learning techniques in one key area. Unlike supervised learning, the model learns from evaluative feedback (e.g., user ratings), makes adjustments to outputs according to the RL algorithm, and continuously improves the output quality in dynamic and ever-changing contexts. The adaptability would be very much able to enhance summary relevance and personalization. An important avenue is also the development of real-time summarization solutions to streaming data as data sources. The need to perform on-the-fly concision and coherence generation from live news feeds, social media streams and sensor-generated data will enable faster decision-making and situational awareness. Thus, we should find models that are both latency- and resource-efficient while maintaining.

The demand for building ethical and bias-aware summarisation models is increasing. Since many pre-trained language models are likely to learn societal biases from their training data, future research should aim to identify, rectify, and adjust these biases in generated summaries, ensuring that the summaries are fair, transparent, and inclusive. The building of trustworthy summarization tools will involve techniques like bias auditing, adversarial training, and diverse datasets…among other things. Finally, our hybrid approach provides a strong foundation; the development of summarization technologies in the directions hinted above will be key to meeting the ever-growing demands of diverse applications in industry and society.

References

[1] Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of research and development, 2(2), 159-165.

[2] Edmundson, H. P. (1969). New methods in automatic extracting. Journal of the ACM (JACM), 16(2), 264-285.

[3] Kupiec, J., Pedersen, J., & Chen, F. (1995, July). A trainable document summarizer. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 68-73).

[4] Lin, C. Y. (1999, November). Training a selection function for extraction. In Proceedings of the Eighth International Conference on Information and Knowledge Management (pp. 55-62).

[5] Mihalcea, R., & Tarau, P. (2004, July). Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing (pp. 404-411).

[6] Gong, Y., & Liu, X. (2001, September). Generic text summarization using relevance measure and latent semantic analysis. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 19-25).

[7] Nenkova, A., & McKeown, K. (2011). Automatic summarization. Foundations and Trends® in Information Retrieval, 5(2–3), 103-233.

[8] Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of artificial intelligence research, 22, 457-479.

[9] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).

[10] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

[11] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 21(140), 1-67.

[12] See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. arXiv preprint arXiv:1704.04368.

[13] Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020, November). Pegasus: Pre-training with extracted gap sentences for abstractive summarization. In International conference on machine learning (pp. 11328-11339). PMLR.

[14] Deshpande, A., & Kumar, M. (2018). Artificial intelligence for big data: a complete guide to automating big data solutions using artificial intelligence techniques. Packt Publishing Ltd.

[15] Moreno, A., & Redondo, T. (2016). Text analytics: the convergence of big data and artificial intelligence. IJIMAI, 3(6), 57-64.

[16] Hesabi, Z. R., Tari, Z., Goscinski, A., Fahad, A., Khalil, I., & Queiroz, C. (2015). Data summarization techniques for big data a survey. Handbook on Data Centers, 1109-1152.

[17] Ahmed, M. (2019). Data summarization: a survey. Knowledge and Information Systems, 58(2), 249-273.

[18] Gupta, V., Bansal, N., & Sharma, A. (2018). "Text Summarization for Big Data: A Comprehensive Survey." In *Innovative Computing and Communications* (LNNS, vol. 56). Springer.

[19] Neto, J. L., Freitas, A. A., & Kaestner, C. A. (2002). Automatic text summarization using a machine learning approach. In Advances in Artificial Intelligence: 16th Brazilian Symposium on Artificial Intelligence, SBIA 2002 Porto de Galinhas/Recife, Brazil, November 11–14, 2002 Proceedings 16 (pp. 205-215). Springer Berlin Heidelberg.

[20] Martín-Gutiérrez, D., Hernández-Peñaloza, G., Hernández, A. B., Lozano-Diez, A., & Álvarez, F. (2021). A deep learning approach for robust detection of bots in Twitter using transformers. IEEE Access, 9, 54591-54601.

[21] Jangra, A., Mukherjee, S., Jatowt, A., Saha, S., & Hasanuzzaman, M. (2023). A survey on multi-modal summarization. ACM Computing Surveys, 55(13s), 1-36.