

Article

Integrating Unsupervised Learning for the Factual Consistency of Generative Models

Sindhu Nair ^{1,*}  and Y. S. Rao ² ¹ Department of Computer Engineering, Dwarkadas J. Sanghvi College of Engineering, Mumbai 400056, India² Department of Electronics & Telecommunication Engineering, Sardar Patel Institute of Technology, Mumbai 400056, India; ysrao@spit.ac.in

* Correspondence: sindhu.nair@djsce.ac.in

Abstract

Text summarization involves analyzing large amounts of text, selecting the salient text features, and arranging them coherently. The graph-based TextRank and statistical topic modeling are unsupervised approaches for generating an extractive synopsis. Deep learning models are supervised, data-driven, and pre-trained on a huge corpus of data, making a significant contribution to automatic text summarization systems. Despite grammatical correctness and coherence, deep learning-based summarization systems are prone to factual inconsistency. This has hindered the applicability of transformer-based summarizers, particularly in critical domains where misleading summarization systems can lead to severe consequences due to their significant social impact. This work proposes an ingenious hybrid hierarchical approach that combines unsupervised approaches, such as the TextRank algorithm and Latent Dirichlet Allocation (LDA)-based summaries, with contemporary transformer-based language models. When validated on three benchmark summarization datasets, empirical results prove that our hybrid hierarchical transformer-based approach mitigates the factual inconsistency problem inherent in abstractive summarization. The improved summary consistency score of the abstractive summaries generated with our multilevel hybrid approach, in comparison to the fine-tuned baseline transformer-based language models, increases trust in transformer-based summarizers.

Keywords: Latent Dirichlet Allocation; summary consistency score; supervised learning; TextRank; transformer; unsupervised learning



Academic Editors: Sardar Jaf, Basel Barakat and Yongqiang Cheng

Received: 20 March 2026

Revised: 18 April 2026

Accepted: 21 April 2026

Published: 27 April 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Extracting valuable information from a large corpus of text is text mining. Text and web search, mining themes and important topics, summarization, and document understanding are tasks that are unique to text and web data. With exponentially increasing web data, there is an increasing need for categorization and summarization of text data. A generated synopsis presents the key points from the entire block of data in a crisp and precise form to make it comprehensible. Ambiguous text, long and complex documents, and text with multiple themes, creative writing styles, and highly technical content pose a challenge for the task of summarization. Text mining [1] encompasses multiple domains, including information retrieval, data mining, machine learning, natural language processing, and deep learning. The main purpose of text mining is to analyse unstructured data and find hidden patterns to facilitate decision-making. Topic modeling is a probabilistic technique for semantic analysis [1] that requires no training data and performs semantic analysis by

extracting latent topics in an unsupervised manner. Each document is categorised into various topics or themes. Each theme can be considered as a soft cluster of words. Each topic consists of a probabilistic distribution of words. Latent semantic analysis [1] and Latent Dirichlet Allocation [1] are two popular techniques of topic modeling under the umbrella of text mining. Information retrieved from large text can aid in crucial decision-making, which has a substantial social impact. The knowledge extracted from large text using text mining techniques can be applied in critical domains such as healthcare, governance, and the media, impacting our daily lives. The summarization of large texts available online can provide key insights for decision-making, where any misinterpretation of facts can have disastrous effects. In recent times, the widespread usability of deep learning-based summarization systems has been hampered due to their factual inconsistency. Any false narrative in the digital news media can affect our daily lives alarmingly. Hence, the veracity of the synopsis must be safeguarded to avoid spreading any misleading information.

Although previous work has explored the combination of LDA with the BART model, it did not address the factual consistency problem [2–6]. Cristian et al. [7] use topic labelling in which, for a corpus of documents, topic clustering is performed using the LDA model. The topic labels are generated and input into the BART model with the corpus of documents. Thus, the BART model is trained for topic labelling. Large pre-trained transformer models are fine-tuned in a weakly supervised manner using an unsupervised LDA model to obtain meaningful labels. Prior work focuses on topic labelling using the combination of the LDA model with BART, while our work provides a solution to the factual consistency problem. Jijo et al. [8] review existing text summarisation methods, including the Textrank algorithm, fuzzy logic, Latent Semantic Analysis, and deep learning techniques. However, the paper does not fuse any of the methods, nor does it address the factual consistency problem. In Chen et al. [9], the results of TextRank and the BART model are combined to obtain more thematic key sentences, which are then fed back into the BART model to produce the final summarisation. Our approach is different, as the extractive TextRank summary is input to the BART model without any splicing with the abstractive output of the BART model. Also, the prior work assesses the summaries using general quality metrics of summarisation, such as ROUGE [10], based on token overlap, and hence did not capture the factual correctness [11,12].

This work provides a viable solution to the factual inconsistency problem in transformer-based deep sequence-to-sequence models. Our work addresses factual inconsistency introduced in the synopsis due to wrong entities, incorrect factual information, swapping of entities and misleading semantics wherein the synopsis is not in concordance with the source text. Addressing factual inconsistency in deep learning-based summarisation is a timely and critical research problem, especially for high-stakes domains such as media, medical, law, and governance. Current state-of-the-art large language models provide coherent human-like summaries, but factual inconsistency or hallucination is still a challenge. In critical applications such as summarisation of medical reports, the essential quality parameter is factual consistency or correctness. Similarly, false news from the media can be misleading, affecting our daily lives. This work makes the following significant contributions: (1) Provides a solution to the factual inconsistency problem in language models, which hampers the usage of these models in critical and sensitive domains where factual accuracy is the main parameter for the assessment of the quality of the summary. (2) Introduces a hybrid hierarchical deep learning approach that combines unsupervised learning of text with transformer-based encoder–decoder models to boost the factual consistency of generated abstractive summaries. (3) The proposed hybrid approach demonstrates a significant improvement in the factual consistency score in generated summaries over the fine-tuned baseline transformer-based language models.

2. Related Work

Cao et al. [13] leverage popular OpenIE and dependency parsing tools to extract triples in the source sentence to reduce fake summaries to handle the Factual Inconsistency Problem. The dual-attention sequence-to-sequence framework is then proposed to force the generation conditioned on both the source text and the extracted fact descriptions. Dou et al. [14] propose a framework that is generalized and takes different input types as guidance signals. The paper investigates highlighted sentences, keywords, relational tuples (subject, relation, object), and retrieved summaries as guidance signals. Experiments show that using highlighted sentences as guidance achieves the best performance. The framework builds upon a modification of BART, which consists of two encoders that respectively encode the input source document and guidance signals. The decoder attends to both the source document and guidance signal. Shin et al. [15] introduce a factuality-guided module combined with a diffusion-based abstractive summarizer to generate summaries that are both diverse and factually consistent. The proposed method continually guides factuality into the intermediate noise at each denoising step, thereby generating summaries that are not only consistent with the source text but also high in diversity. Li et al. [16] propose an entity-augmented method (EAM) for factual consistency that uses cross-attention to focus on entities during the decoding process. The EAM method aims to improve the faithfulness of the generated summaries by explicitly incorporating and attending to entity boundary information within the Transformer model. The core of the proposed method involves adding two key modules to the Transformer: (1) Sparse Entity Matrix which is placed after the encoder and helps to encode boundary information about entities present in the input text. (2) Entity Cross-Attention Layer which is added to the decoder and facilitates the model's ability to attend to entity representations during the generation process. Shin et al. [17] have proposed a token-level fact-correction model in abstractive summarization that uses joint training for generic token-level fact correction. The proposed fact corrector consists of three modules: the summary fact checker, the token fact checker, and the fact rectifier. These modules are applied sequentially to a given inconsistent summary to generate a consistent summary.

Zhu et al. [18] extract the factual information from the article in the form of a relational knowledge graph and integrate its information into an end-to-end summary generation process. The knowledge graph is obtained from information extraction results and it participates in the decoder's attention. Jin et al. [19] explore the incorporation of semantic dependency graphs into abstractive summarization models. The model consists of a sentence encoder, a graph encoder, and a summary decoder. Firstly, the sentence encoder reads the input sentence and builds its contextual-level representation. Then, the graph encoder captures the semantic relationships according to the contextual-level representation and the semantic dependency graph to produce a semantic-aware sentence representation. Lastly, the decoder produces the output summary with attention to the semantic-aware representation. Feng et al. [20] filter the training data to reduce false entities in the synopsis, in addition to proposing a new metric for entity-level factual consistency. Feng et al. [21] propose a new training algorithm that maximises the proposed factual consistency metric during model training. Mao et al. [22] use a model-specific approach that integrates multi-granular factual input during encoding. Crucial factual information, such as the related time, locations, reasons, consequences, purposes, participants and involved parties, the complex semantic relations among facts, and the corresponding factual information, is considered. Pernes et al. [23] use contemporary evaluation metrics like QuestEval and FactCC to rank the synopsis. Wan et al. [24] analyze the effect of beam search and nucleus sampling on the factual consistency of generated summaries. Zhu et al. [25] integrate a knowledge graph into summary generation. Tanay et al. [26] analyze the trade-off between

the factual consistency achieved by a contrastive learning framework and the overall quality retained in the generated summary. Meng et al. [27] use fact-corrector models to improve the factual consistency, while Yue et al. [28] use question-answering models and multimasking strategies in factual correction models.

3. Proposed Methodology

Among the various challenges of abstractive summarization, factual inconsistency and unfaithfulness of the summary generated are targeted by this proposed methodology. When information is extracted from an article and its summary generated using deep learning models, the factual correctness of the article should be maintained. Furthermore, the generated summary should not differ from the source text and should be in accordance with the source text. Although the summaries generated by the transformer-based deep learning models are fluent, they are not devoid of factual inconsistency. Also, such unfaithful summaries that deviate from the source text can misguide the reader. Many times, summaries are used to make critical decisions. Providing misinformation to the information consumer can have serious consequences. Many critical applications require accurate and faithful summaries which can aid in decision-making. Extractive and abstractive are the two approaches for text summarization.

Table 1 differentiates between extractive and abstractive summarization covering features such as coherence, cost, speed, complexity, accuracy, context, novelty, and limitations. The extractive approach involves algorithms for unsupervised learning of text, while the abstractive approach involves contemporary transformer-based language models.

Table 1. Summarization approaches.

Extractive	Abstractive
Extracts key sentences	Generates new sentences
Directly copied text	Not directly copied text
less coherent	more coherent, fluent
Limited by the original sentences	condenses effectively
may miss implicit connections	captures implicit relations
High accuracy	potential for errors
Lower cost, simpler algorithms	advanced NLP
Best for speed, accuracy	Best for Complex contextual text

Two popular approaches for unsupervised learning are applied here: the probability-based Topic Modeling approach [29], where patterns in text are formed based on their occurrences together, and the graph-based TextRank algorithm [30], which ranks the sentences based on their similarity. The proposed methodology applies unsupervised learning of text, like TextRank and Topic Modeling-based summarization, to handle the Factual Inconsistency Problem inherent in transformer-based abstractive summarization. This section further discusses the LDA-based extractive summarization, and the TextRank extractive approach before explaining in detail the proposed hybrid hierarchical deep learning approach. The last subsection discusses the datasets used as a valuable testbed for our proposed methodology.

3.1. Extractive Summarization

In case of LDA-based Extractive Summarization, along with the source article, the number of topics is given as a hyperparameter when building the LDA model. Before

building the LDA model with the inputs, the source text is first sentence tokenized. Then, each sentence is preprocessed by tokenization, removal of stopwords, and lemmatization. This is followed by vectorization.

Using the probabilistic model, the input text is clustered into topics with a coherence score. The sklearn library with the LDA Model is leveraged for extracting a summary from the LDA Model in Python 3.12.13. As illustrated in Figure 1, once the input text is clustered into topics, the most representative sentences from each cluster are picked to form the summary. The number of sentences to be picked from each topic cluster is again given as input.

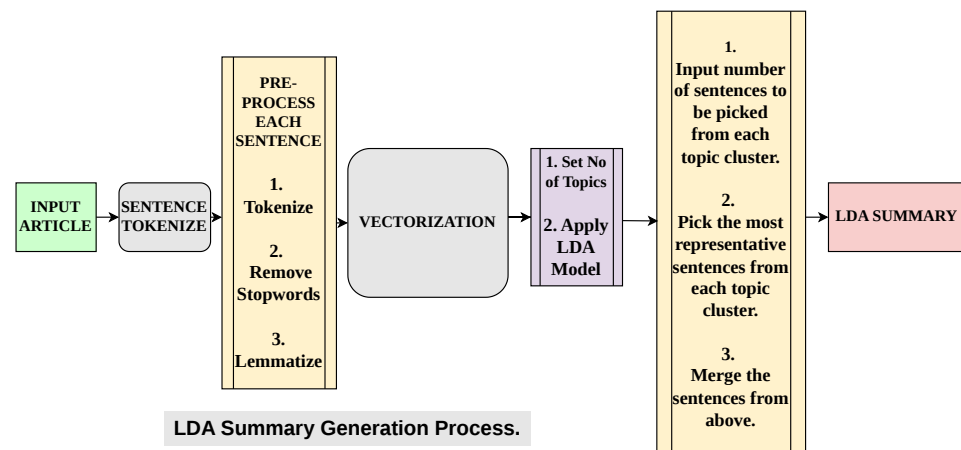


Figure 1. LDA summary generation.

TextRank is a variant of the PageRank algorithm for extractive summarization that does not require labelled data. The TextRank algorithm determines the importance of a sentence by analyzing its similarity to other sentences. A sentence is considered more important if it is similar to many other sentences. Figure 2 displays the system flow for TextRank summary generation. The input preprocessing involves removing duplicates, missing values, special characters, and irrelevant features. The text cleaning is followed by the stemming, lemmatization, and tokenization of the input. TextRank is an unsupervised approach where the text document is depicted as a graph with nodes as sentences. The edges or links between the nodes represent the similarity between the sentences. The sentences in the summary are numbered K , and this value is chosen by the user. Identical sentences tend to have the same similarity value. The cosine similarity given in Equation (1) is widely used for extractive text summarization. The sentence similarity score based on cosine similarity ranges from 0 to 1. The ranking of sentences is computed based on the score. The summary includes only the top-ranked sentences [30].

$$\text{CosineSimilarity}(A, B) = \frac{A \cdot B}{|A| \cdot |B|} \quad (1)$$

3.2. Hybrid Hierarchical Deep Learning Approach

Deep learning models have been a game-changer in the field of natural language processing. NLP tasks such as translation, summarization [31–33], and natural language generation have witnessed human-like performance with the advent of deep-learning models such as attention-based pointer networks, generative adversarial networks, and the transformer-based encoder–decoder models. Attention-based transformer models [34] have revolutionized text summarization and can be classified primarily as encoder-only models, decoder-only models, and encoder–decoder models. Encoder-only models, such as BERT [35], are applied for text classification, while decoder-only models are used to

generate text. The encoder–decoder models like BART [36], T5 [37] and Pegasus [38], which differ primarily in their pretraining objective, are apt for text-to-text transfer with varying input and output text length. Such encoder–decoder models are popularly applied for NLP tasks such as text summarization and translation. The transformer-based encoder–decoder models give good performance on summarization tasks but suffer from the Factual Inconsistency Problem. We design the hybrid hierarchical language model, which combines the abstractive approach of the deep learning models with the extractive approaches of unsupervised learning of text. The hybrid approach is multilevel with the extractive approach in the first level followed by the abstractive approach. In the first tier, which is extractive, we apply two extractive methods of summarization: TextRank and topic modeling. In the second tier, we apply three encoder–decoder models: BART, T5, and Pegasus.

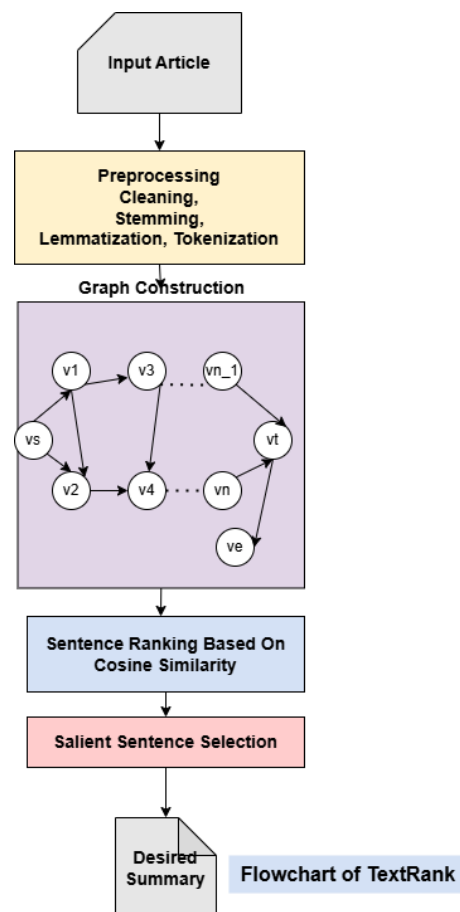


Figure 2. TextRank summary generation.

Figure 3 depicts the overall architecture of the proposed methodology, which includes the fine-tuning of the pre-trained model, the hybrid deep learning approach, including the TextRank–Transformer model and the LDA–Transformer model, the hybrid knowledge-enhanced deep learning approach, involving the LDA–TextRank–Transformer model, and the natural language inference model SummaC [39] for evaluation. For each element of the pipeline, such as LDA–Transformer, TextRank–Transformer, and LDA–TextRank–Transformer, we validate our approach across three distinct datasets (BBC, CBC, and MeetingBank) covering different text styles and lengths. The transformer-based deep learning models are finetuned on appropriate news dataset or dialogue dataset. The finetuned models can be leveraged from the hugging face framework.

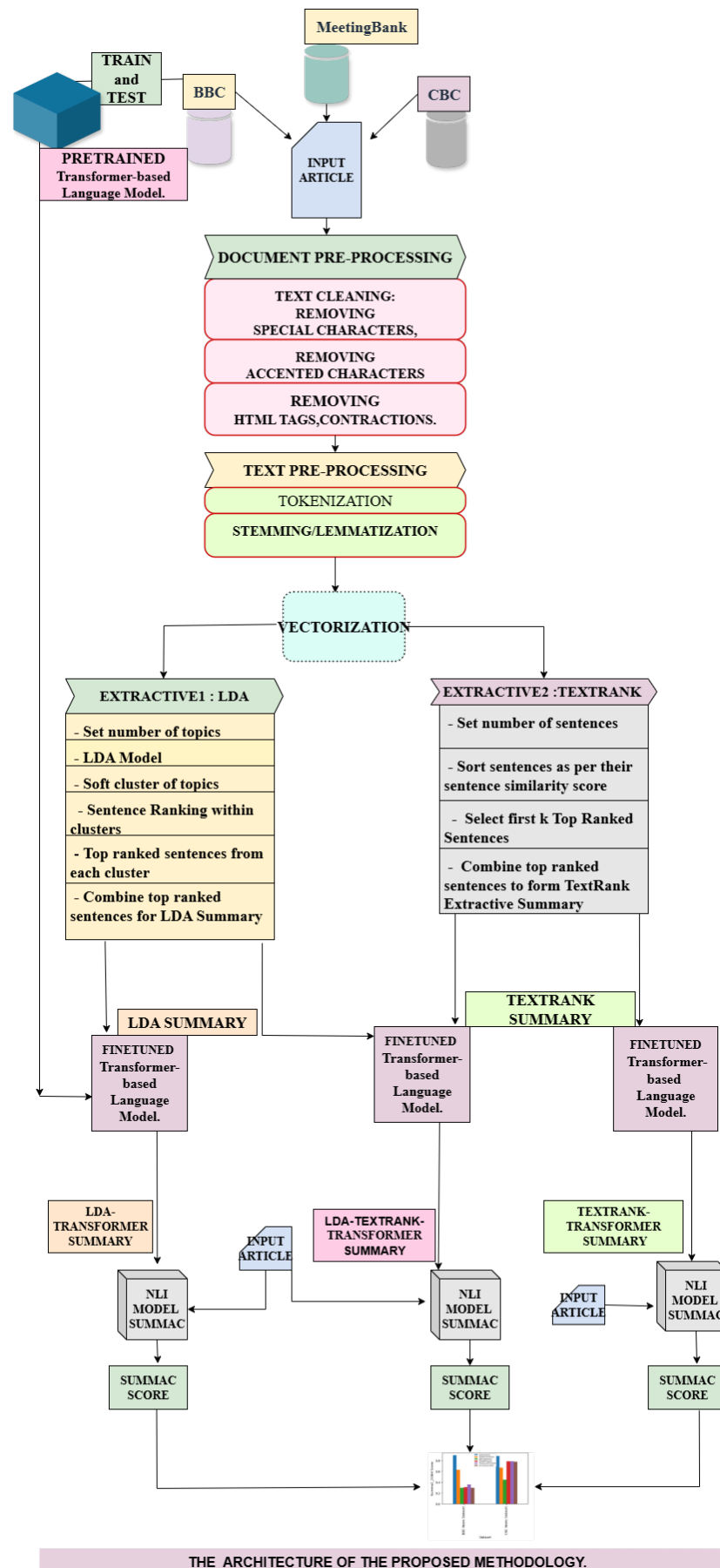


Figure 3. The architecture of the hybrid hierarchical deep learning approach.

In the TextRank–Transformer model, the graph-based extractive TextRank summary is input into the finetuned Transformer-based baseline language model. In the LDA–Transformer model, the statistical extractive LDA summary is input into the finetuned Transformer-based baseline language model. Thus, the TextRank–Transformer model and the LDA–Transformer model of the Hybrid Deep Learning Approach are hybrid and multilevel with the extractive approach followed by the abstractive approach. The rationale applied is that, with the extractive approach of summarization in the first tier, followed by the abstractive approaches, the factual errors can be curbed. The knowledge-enhanced approach addresses the problem of semantic irrelevancy by feeding additional input to the model. Examples of guidance information include keywords, length tokens, salient sentences, facts and semantics using graphs and relations, template-based reference summaries [40]. The hybrid knowledge-enhanced deep learning approach has two diverse extractive summaries of the same article input into the deep learning Transformer model. In this hybrid hierarchical architecture, one extractive summary is knowledge-enhanced/guided by the other augmenting extractive summary. The rationale is that this will ensure all the salient points from the source article is covered in addition to addressing the factual inconsistency problem. Hence, the LDA summary is augmented with the TextRank summary, and it is input to the fine-tuned Transformer model.

Deep learning models rely heavily on training data and are computationally intensive, while the proposed methodology fuses the statistical LDA approach and the graph-based TextRank approach with the deep learning approach, reducing the dependency on large training data. In each element of the pipeline, the input article is pre-processed, vectorised, and then input to the extractive model. The output of the extractive model is input to the transformer-based language model. In the LDA–TextRank–Transformer pipeline, the output of the extractive models is combined and then input to the transformer-based model. Finally, the output is evaluated by the NLI-based model, which gives the Summary Consistency Score. The higher the score, the more factually consistent the summary. SummaC serves as the central evaluation measure.

The other evaluation metric applied in the experiments is QAGS Score. QAGS methods involve the following three steps: question generation, question answering with the document and summary, and finding common answers in the document and summary [41]. The generated summary is input to the pre-trained RoBERTa model, which generates the questions for it. The same set of questions generated from the summary are input to the question-answering model along with the source text, to output the keyword answers. Similarly, the keyword answers are generated for the generated summary. Entity or keyword matching is performed between the two sets of answers: one generated for the source text and the other generated for the summary for the same set of questions. This evaluation gives the QAGS Score [41]. The Factsumm Python package is used for the implementation of the QAGS Score [42]. The QAGS Score [43] is calculated as shown in Equation (2).

$$\text{QAGS SCORE} = \frac{\text{Matched Entities (Summary, Source)}}{\text{Total Entities (Summary)}} \quad (2)$$

SummaC [43] is the contemporary metric for factual inconsistency based on natural language inference with two variants: (1) *SummaC_{ZS}* (zero-shot) and (2) *SummaC_{CONV}* (Convolution). From the NLI pair matrix that is generated, the technique for generating the final score varies in the two variants of SummaC. For *SummaC_{ZS}*, the final score is calculated by first taking the highest value of each column in the pair matrix and then calculating the average value. Hence, *SummaC_{ZS}* does not have trained parameters and is called zero-shot. For *SummaC_{CONV}*, the convolution layer makes it insensitive to extreme values, which can be noise or outliers. *SummaC_{CONV}* gives good performance by capturing

sentence-level inconsistencies and ignoring the noise. The probability distribution of the NLI pair matrix in the form of a histogram is mapped to a binning matrix, which when passed through the convolution layer computes the final score. *SummaC_{CONV}* gives good performance by capturing sentence-level inconsistencies and ignoring the noise. Thus, for each synopsis sentence, the convolution layer maps the entire distribution to a single score. Hence, *SummaC_{CONV}* has trained parameters since the NLI pair matrix is binned and trained using a convolution layer. Thus, the two variants of SummaC apply different aggregation techniques on the NLI pair matrix to map it to the final score.

The *SummaC_{ZS}* [43] is calculated as shown in Equation (3).

$$SummaC_{ZS} = \overline{\text{Max of Each Column(NLI Matrix)}} \quad (3)$$

The *SummaC_{CONV}* [43] is calculated as shown in Equation (4).

$$SummaC_{CONV} = f * (\text{Binning Matrix(NLI Matrix)}) \quad (4)$$

3.3. Datasets

Our proposed methodology is further tested with news articles from the BBC News dataset [44] and the CBC News dataset [45]. The Meeting Bank dataset [46] is a benchmark dataset with transcripts along with their synopsis. It provides a valuable testbed for our proposed methodology. The average length of the news articles in the BBC News dataset is less than 500 tokens. The BBC News dataset is well-curated and does not need pre-processing. The CBC News dataset consists of web-scraped pandemic news articles from the CBC News Site for 2020, which examines the sentiments of the public around the world and the narration of global events during the pandemic. The CBC News dataset consists of 6787 unique rows that reflect the count of news articles. The average length of the news articles in the CBC News dataset is approximately 1350 tokens. Hence, it consists of lengthy and complex news articles that require pre-processing. Unstructured data usually contains noise, which is filtered out by pre-processing steps before feeding it to the NLP model. Details of the CBC News dataset are depicted in Table 2.

Table 2. Unique column values in CBC news dataset.

Column	Unique Rows
ROWID	6787
AUTHORS	352
TITLE	4998
PUBLISH-DATE	4136
DESCRIPTION	4988
TEXT	5041
URL	5139

After removing the missing values from the columns, as shown in Table 3, and removing duplicate tuples, the dataset now has 5043 tuples, as depicted in Table 4. The headline of the news, the short description of the news, and the text article of the news from the dataset are used for further analysis to compute their individual length in tokens. Thus, the length of the news headline, the length of the article, and the length of the news description are appended to the CBC News dataset as new columns. Text cleaning was performed on the CBC News dataset, including removing anything not number/alphabet, emojis, large blank spaces, HTTP/URLs/HTML tags, special characters/noise, and expanding

contractions. The benefits of dataset pre-processing include faster training time as the size of the dataset eventually decreases, and it boosts performance by reducing the ambiguities caused by variations in text formatting.

Table 3. Missing values in CBC News dataset.

Column	Unique Rows
DESCRIPTION	0.0003 per missing values
URL	0.0003 per missing values

Table 4. CBC News dataset shape.

Criteria	Shape
Including duplicates and missing values	(6787, 3)
Excluding duplicates and missing values	(5043, 3)

4. Experimental Setup

Hugging Face was leveraged as the deep learning framework for abstractive summarization. The Python Sumy package in the Google Colab setup was leveraged for the implementation of the TextRank algorithm. The Sklearn Python package was leveraged to build and apply the LDA model for topic modeling. After multiple trials, the number of topics is selected based on the LDA model with the least perplexity. It is validated on both the BBC and the CBC datasets. All details of the parameters set for the implementation of extractive approaches, LDA and TextRank are presented in Table 5.

Table 5. Parameters for extractive summarization.

Model	Package	Parameters
LDA	Sklearn	NoOfTopics = 3 (for BBC News dataset),
LDA	Sklearn	NoOfTopics = 4 (for CBC News dataset),
TextRank	Sumy	NoOfSentences = (ArticleLength)/10
TextRank	Sumy	2 (for ArticleLength less than 300)
TextRank	Sumy	20 (for ArticleLength more than 2000)

The count of sentences in the synopsis is given as input to the TextRank algorithm based on the length of the source article. For very small articles with a token length of less than 300, the count of sentences in the extractive summary is set to 2. For very long articles with a token length of more than 2000, the count of sentences in the extractive summary is set to 20.

Whilst training the ‘sshleifer/distilbart-cnn-6-6’ pre-trained model, the hyperparameters include the maximum learning rate, number of epochs, per device batch size, and the maximum length. For decoding strategies, the hyperparameters set includes beam size, length penalty, max-length, and min-length.

All details of the hyperparameters using the BLURR package for the ‘sshleifer/distilbart-cnn-6-6’ model are exhibited in Table 6 above. The Blurr library, which is an integration of OpenAI with Hugging Face, is used for fine-tuning and training. For fine-tuning and model training of BART, the imported Python libraries included transformers, dataset, bert-score, blurr, and their dependencies. The pre-trained model is fine-tuned with the BBC News dataset. Tuples from each category of articles, ranging from business, entertainment,

politics, sports, and technical, are extracted from the BBC News dataset and combined to fine-tune the BART model. The hyperparameters are set and input into the training algorithm. The hyperparameters are related to both the network architecture and the training algorithm. Hyperparameters include dropout, activation function, number of epochs, learning rate, and batch size, among others. During fine-tuning, the model is trained for 10 epochs. Applying our approach to other encoder–decoder models like T5 and PEGASUS that differ in their pre-training objective. All three transformer-based models are fine-tuned on the Samsum dataset for dialogs and tested on the MeetingBank dataset, which has meeting data between interlocutors. When testing our approach on news articles, we pick the version of the encoder–decoder models from the Hugging Face hub that are finetuned on CNN news articles for T5 and Pegasus PEGASUS.

Table 6. Training hyperparameters.

Training Hyperparameters	Decoding Hyperparameters
0.00003 (learning rate)	beam size = 4
10 (epoch count)	length penalty = 2
2 (batch-size)	max-length = 142
max-length = 1024	min-length = 56

5. Results and Discussion

The summaries generated by our proposed methodology are analyzed in detail and compared with the output summaries predicted by the baseline fine-tuned Transformer models. For the CBC News dataset, the average summary length in tokens of the synopsis generated by extractive probabilistic LDA and graph-based TextRank is the same, which is greater than the summaries generated by baseline state-of-the-art abstractive summarization model BART, pretrained with a denoising autoencoding objective and fine-tuned with Tesla T4 GPU. Our proposed hybrid approach generates summaries of shorter lengths than the baseline BART model. However, the hybrid knowledge-enhanced approach LDA–TextRank–BART generates summaries slightly longer than the hybrid models TextRank–BART and LDA–BART. Table 7 below depicts the length of the synopsis generated by the various models leveraged where the length of the article for the pandemic news articles is three times that of the BBC news articles.

Table 7. Comparison of summary length in tokens of BBC News dataset and CBC News dataset.

Model	Type	Mean Length		Standard Deviation	
		BBC	CBC	BBC	CBC
LDA	Extractive	127	132	28	25
TextRank	Extractive	99	134	62	101
FinetunedBART	Abstractive	91	98	31	21
TextRank–BART	Hybrid	68	66	26	22
LDA–BART	Hybrid	72	69	29	19
LDA–TextRank–BART	HybridKE	88	79	33	21

Table 8 displays the performance metrics for the BBC News dataset and the CBC News dataset for extractive summarization. For the BBC News dataset, it is observed that the informativeness and factual accuracy (given by the QAGS score) are marginally higher for the extractive approaches LDA and TextRank than for the baseline abstractive

BART model. The $SummaC_{ZS}$ Score and the $SummaC_{CONV}$ Score are much higher than the baseline finetuned BART abstractive model when validated on both benchmark datasets, the BBC News dataset and the CBC News dataset. Table 8 exhibits the fact that the factual consistency of the extractive LDA approach is greater than the graph-based TextRank approach. Thus, the probability-based LDA summarization model generates more factually consistent summaries than the graph-based TextRank as proven empirically and validated on two benchmark datasets. Hence, the extractive approaches of summarization are more factually consistent than the abstractive summaries generated by fine-tuned BART, as proven empirically.

Table 8. Performance metrics for extractive summarization.

ModelsForBBC	QAGS	$SummaC_{ZS}$	$SummaC_{CONV}$
FinetunedBART	0.83	0.17	0.30
LDA	0.87	0.60	0.90
TextRank	0.84	0.58	0.63
ModelsForCBC	QAGS	$SummaC_{ZS}$	$SummaC_{CONV}$
FinetunedBART	0.82	0.22	0.45
LDA	0.83	0.49	0.89
TextRank	0.79	0.47	0.67

Tables 9 and 10 display the $SummaC_{CONV}$ Score for the summaries generated by our proposed system for the CBC News dataset for articles with lengths ranging from 298 tokens to 2245 tokens. Our hybrid approaches have an improved factual consistency score than the baseline model as depicted in Tables 9 and 10. The results show that, in terms of factual consistency, LDA–BART, TextRank–BART, and LDA–TextRank–BART outperform fine-tuned BART summaries. Table 9 contains articles of length less than 1000 tokens where we analyze four articles from the first group of length less than 600 tokens and five articles from the group of length between 600 and 1000 tokens. It is observed that the factual consistency of the synopsis generated by our models, TextRank–BART, LDA–BART, and LDA–TextRank–BART, exceeds the baseline model, as shown by the $SummaC_{CONV}$ scores. Table 10 consists of four articles in the first bucket with lengths of less than 1500 tokens and five articles in the bucket with lengths of more than 1500 tokens. Here, again, our approaches display higher factual consistency scores, which shows the effectiveness of our methodology in nullifying factual errors. Thus, the novel innovative hybrid approaches, LDA–BART, TextRank–BART, and LDA–TextRank–BART, generate more factually consistent summaries than fine-tuned BART. The performance of our methodology in terms of factual consistency is observed to be comparable in the three LDA–BART, TextRank–BART, and LDA–TextRank–BART models, outperforming the baseline abstractive BART model. In addition, there is no clear advantage in augmenting the LDA summary with the TextRank summary before it is input into the BART model. Further, it is observed that, although LDA summaries outperformed TextRank summaries in terms of factual consistency, no such conclusion can be drawn between the hybrid models LDA–BART and TextRank–BART.

Table 11 shows the comparison of the SummaC Score for the summarization of 100 CBC news articles, which clearly displays the higher scores of factual consistency for our proposed hierarchical hybrid approach. Additionally, it is observed that the factual consistency scores of the summaries generated by hybrid models and the hybrid knowledge-enhanced approach are comparable. Table 12 displays the evaluation of the synopsis generated by our approach applied on the well-curated BBC News dataset. There is an improvement

in the factual consistency as shown by the SummaC Score. The factual consistency of the summaries generated by the baseline transformer model is competitive when applied on the well-curated BBC dataset. The informativeness and semantic similarity of the synopsis are displayed by RougeL and BERTScoreF1 [47] respectively, which is retained in our methodology. Thus, our approaches give a balanced summary in terms of coherence, informativeness, semantic similarity, and factual consistency.

Table 9. Factual consistency score $SummaC_{CONV}$ for predicted summaries by our proposed system for CBC News dataset for articles with length less than 1000 tokens.

Article-Length	298	506	514	532	629	680	819	864	928
Finetuned-BART	0.23	0.24	0.31	0.26	0.27	0.40	0.26	0.29	0.26
TextRank-BART	0.5	0.83	0.96	0.79	0.77	0.44	0.90	0.86	0.79
LDA-BART	0.60	0.83	0.96	0.90	0.85	0.86	0.88	0.81	0.64
LDA-TextRank-BART	0.42	0.83	0.60	0.80	0.86	0.81	0.88	0.82	0.70

Table 10. Factual consistency score $SummaC_{conv}$ for predicted summaries by our proposed system for CBC News dataset for articles with length more than 1000 tokens.

Article-Length	1051	1175	1187	1365	1620	1727	1764	1838	2245
Finetuned-BART	0.3	0.32	0.55	0.46	0.34	0.36	0.63	0.53	0.71
TextRank-BART	0.93	0.85	0.87	0.61	0.89	0.79	0.92	0.93	0.90
LDA-BART	0.86	0.91	0.5	0.87	0.80	0.94	0.90	1	0.91
LDA-TextRank-BART	0.77	0.84	0.78	0.87	0.85	0.88	0.88	0.92	0.74

Table 11. SUMMACscore of 100 news articles of CBC dataset.

Model	Type	$SUMMAC_{ZS}$	$SUMMAC_{CONV}$
FinetunedBART	Abstractive	0.22	0.45
LDA-BART	Hybrid	0.45	0.79
TextRank-BART	Hybrid	0.44	0.76
LDA-TextRank-BART	HybridKE	0.48	0.78

Table 12. Evaluation of 100 news articles of BBC dataset.

Model	RougeL	BertScoreF1	$SUMMAC_{CONV}$
FinetunedBART	0.33	0.88	0.30
LDA-BART	0.28	0.88	0.31
TextRank-BART	0.31	0.88	0.36
LDA-TextRank-BART	0.31	0.88	0.30

By applying the NLI-based model SummaC to evaluate the factual consistency of the synopsis text, the convolution variant $SummaC_{CONV}$ Score is more discriminating and robust than its zero-shot version, $SummaC_{ZS}$. However, there is no correlation of the increase in summary consistency score with the length of the news articles for the TextRank-BART model, the LDA-BART model and the LDA-TextRank-BART model. It is observed that the extractive approaches are lightweight with a very small processing time. This makes the proposed hybrid model lightweight and less dependent on the training data. Hence, the proposed model would be most effective when tested on previously

unknown data. This is proven empirically when the proposed novel innovative approach is tested on the CBC News dataset, which consists of web-scraped pandemic news articles from the Canadian news website. Our approach is applied on three valuable testbeds using three encoder–decoder models, and the results are displayed below. Since the hybrid knowledge-enhanced approach LDA–TextRank–Transformer did not have any advantages in the initial results, further experiments were conducted with the two hybrid approaches involving LDA and TextRank.

Table 13 shows the evaluation scores for the MeetingBank dataset. We applied our approach to three encoder–decoder models, BART, T5, and PEGASUS, which differ in their pre-training objective. The pre-training objective of BART is denoising autoencoder where original text is reconstructed from corrupted text. In case of PEGASUS, the pre-training objective is Gap Sentence Generation while T5 predicts missing spans as text. The MeetingBank dataset, consisting of meeting information between the interlocutors, is used for testing. The three transformer-based models are fine-tuned on the Samsun dataset for dialogs. We observe that the hybrid models outperform the baseline models in terms of factual consistency. Additionally, their ROUGELScore and BERTScoreF1 indicates that the enhanced factual consistency does not compromise summary quality parameters such as informativeness, coherence, and semantic similarity. The pipeline of T5 gives the best performance as its hybrid model LDA–T5 has the highest SummaC score with a 25% increase over its baseline model.

Table 13. Evaluation scores on MeetingBank dataset.

Model	RougeL	BertScoreF1	SummaC	
			ZS	Conv
BART	0.19	0.86	−0.05	0.52
LDA–BART	0.22	0.86	0.13	0.59
TEXTRANK–BART	0.21	0.86	0.04	0.59
T5	0.25	0.87	−0.02	0.57
LDA–T5	0.24	0.87	0.26	0.71
TEXTRANK–T5	0.27	0.88	0.18	0.67
PEGASUS	0.25	0.87	0.05	0.6
LDA–PEGASUS	0.23	0.87	0.19	0.66
TEXTRANK–PEGASUS	0.23	0.87	0.14	0.63

The LDA–Transformer model and the TextRank–Transformer models show improved performance when compared to the corresponding transformer baseline models, as depicted in Figure 4 below. The factual consistency of the synopsis is evaluated with the SummaC Score with its two variants *SummaC_{ZS}* and *SummaC_{CONV}*. Figure 4 compares the SummaC scores for each of the models of our proposed methodology for each benchmark dataset. To illustrate the effectiveness of the SummaC score, the two variants of the SummaC score are stacked over each other for each model. The negative *SummaC_{ZS}* Score indicates that the factual consistency is very low. In the case of the MeetingBank dataset, we observe that the best performing model in terms of factual consistency is LDA–T5. For each of the DistilBART, T5, and PEGASUS models, it is observed that the hybrid model with LDA gives more factually consistent summaries than the TextRank–Transformer model version. The hybrid models clearly outperform the baseline transformer models in terms of the factual consistency score. The same observation is true for the other two benchmark datasets BBC News dataset and CBC News dataset except for some outlier ob-

servations, such as TextRank-T5 with higher factual consistency score than LDA-T5. Also, TextRank-DistilBART generates more factually consistent summaries than LDA-DistilBART for BBC dataset.

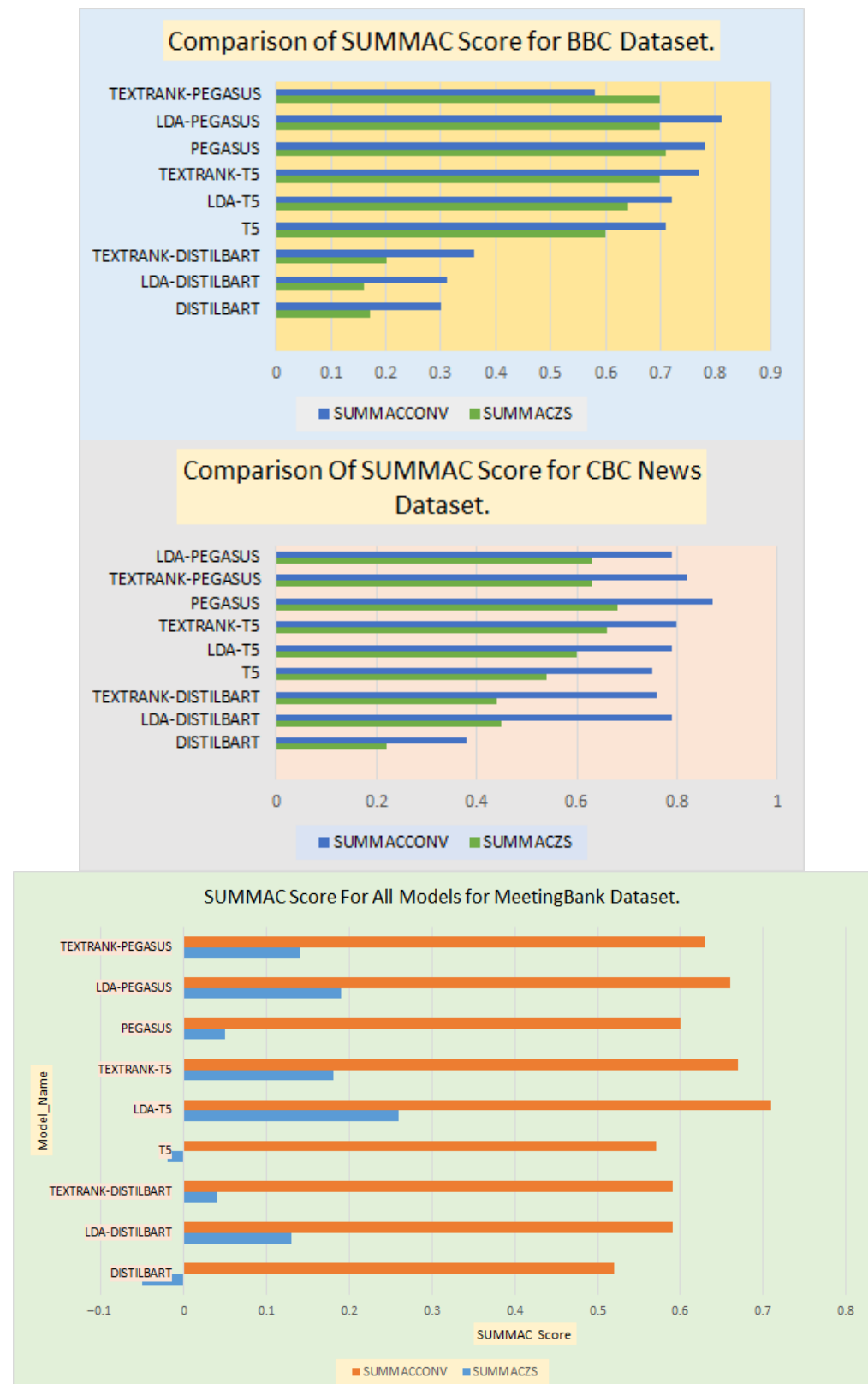


Figure 4. The comparison of the SummaC score of our proposed methodology with the baseline models.

Figure 5 compares the SummaC scores across the three encoder–decoder models T5, PEGASUS, and DistilBART for all three benchmark datasets stacked over each other. Based on the *SummaC_{ZS}* Score and the *SummaC_{CONV}* Score, the hybrid models of our proposed methodology outperform the baseline model with the LDA–Transformer models exceeding the TextRank–Transformer models in terms of factual consistency.

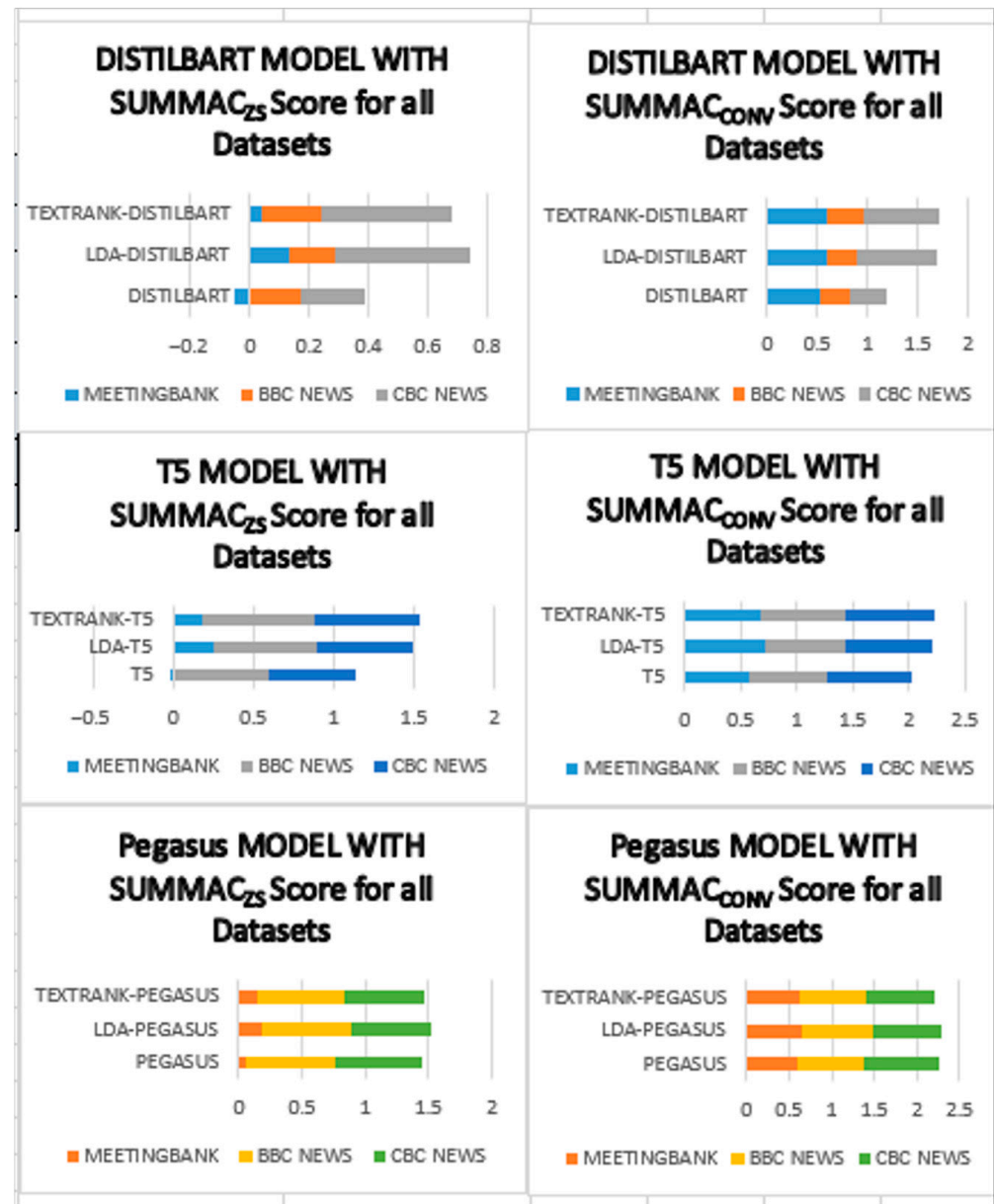


Figure 5. A comparison of SummaC scores across datasets and models.

Despite the high performance of transformer-based summarizers, surveys have shown that 30% of the summaries contain factual inconsistencies. The hybrid approach introduced in this paper addresses the factual inconsistency problem inherent in transformer-based abstractive summaries. This can positively impact the usability of transformer-based summarizers, especially in critical domains such as Medical, Media, and Governance, where factual consistency of the generated synopsis is essential. Beyond news datasets, the proposed approach can be applied to other practical domains such as biomedical, legal, or scientific text.

6. Conclusions

Transformer-based models are hampered by their reliance on static training data. Any bias inherent in the training data is magnified by the transformer-based deep learning models. The unsupervised learning of text induced in the hybrid approach reduces the dependency on the training data. The hybrid models give better results in comparison to the baseline transformer-based encoder–decoder models in terms of the factual consistency of the generated summaries. The rationale for harnessing the hybrid approach is the high factuality of extractive summarization approaches, as proven empirically. Our innovative lightweight hybrid approach is a multilevel architecture that leverages TextRank as well as LDA to solve the factual inconsistency problem in the state-of-the-art abstractive summarization model. Our approach is able to achieve a reasonable trade-off between factual consistency and other summary quality parameters like informativeness, semantics, and coherence. Empirical results prove that our hybrid approach mitigates the factual inconsistency problem inherent in abstractive summarization. Rather than relying on a single generative model, this hybrid approach combines multiple models of diverse strengths to create a more accurate and reliable final prediction. We validate our approach with three distinct datasets (BBC, CBC, and MeetingBank) with different text styles and lengths. Current large language models (LLMs) such as the GPT series, Llama, and other recent open-source LLMs achieve strong performance in abstractive summarisation but the problem of factual inconsistency or hallucinations persists. The proposed hybrid method can be applied to improve the factual consistency of modern LLMs, and generalized to other sensitive domains.

Author Contributions: Conceptualization, S.N.; Investigation, S.N.; Writing—original draft, S.N.; Writing—review & editing, Y.S.R.; Supervision, Y.S.R.; Project administration, Y.S.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Allahyari, M.; Pouriyeh, S.; Assefi, M.; Safaei, S.; Trippe, E.D.; Gutierrez, J.B.; Kochut, K. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. *arXiv* **2017**, arXiv:1707.02919. [\[CrossRef\]](#)
2. Syed, A.A.; Gaol, F.L.; Matsuo, T. A Survey of the State-of-the-Art Models in Neural Abstractive Text Summarization. *IEEE Access* **2021**, *9*, 13248–13265. [\[CrossRef\]](#)
3. Xu, G.; Meng, Y.; Chen, Z.; Qiu, X.; Wang, C.; Yao, H. Research on Topic Detection and Tracking for Online News Texts. *IEEE Access* **2019**, *7*, 58407–58418. [\[CrossRef\]](#)
4. Buenano-Fernandez, D.; Gonzalez, M.; Gil, D.; Lujan-Mora, S. Text Mining of Open-Ended Questions in Self-Assessment of University Teachers: An LDA Topic Modeling Approach. *IEEE Access* **2020**, *8*, 35318–35330. [\[CrossRef\]](#)
5. Yang, P.; Li, W.; Zhao, G. Language Model-Driven Topic Clustering and Summarization for News Articles. *IEEE Access* **2019**, *7*, 185506–185519. [\[CrossRef\]](#)
6. Biswas, P.K.; Iakubovich, A. Extractive Summarization of Call Transcripts. *IEEE Access* **2022**, *10*, 119826–119840. [\[CrossRef\]](#)
7. Popa, C.; Rebedea, T. BART-TL: Weakly-Supervised Topic Label Generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 1418–1425.
8. Jijo, S.M.; Panchal, D.; Ardesana, J.; Chaudhari, U. Text Summarization using Textrank, Lexrank and Bart model. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kamand, India; IEEE: New York, NY, USA, 2024; pp. 1–7. [\[CrossRef\]](#)

9. Chen, Y.; Song, Q. News Text Summarization Method based on BART-TextRank Model. In *IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China*; IEEE: New York, NY, USA, 2021; pp. 2005–2010. [CrossRef]
10. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. 2004; pp. 74–81. Available online: <https://www.aclweb.org/anthology/W04-1013> (accessed on 20 April 2026).
11. Wei, Y.; Ding, Y. Application of Text Rank Algorithm Fused with LDA in Information Extraction Model. *IEEE Access* **2023**, *11*, 84301–84312. [CrossRef]
12. Bongale, R.; Chafekar, T.; Chowdhary, M.; Bapecha, T. Automatic News Summarizer Using TextRank. In *IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India*; IEEE: New York, NY, USA, 2022; pp. 1–5. [CrossRef]
13. Cao, Z.; Wei, F.; Li, W.; Li, S. Faithful to the original: Factaware neural abstractive summarization. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018*; pp. 4784–4791.
14. Dou, Z.Y.; Liu, P.; Hayashi, H.; Jiang, Z.; Neubig, G. GSum: A general framework for guided neural abstractive summarization. *arXiv* **2020**, arXiv2010.080142020.
15. Shin, J.; Park, H.; Song, H.-J. Factuality Guided Diffusion-Based Abstractive Summarization. *IEEE Access* **2024**, *12*, 139302–139315. [CrossRef]
16. Li, J.; Liu, J.; Ma, J.; Yang, W.; Huang, D. Boundary-aware abstractive summarization with entity-augmented attention for enhancing faithfulness. *Acm Trans. Asian-Low-Resour. Lang. Inf. Process.* **2024**, *23*, 1–18. [CrossRef]
17. Shin, J.; Park, S.-B.; Song, H.-J. Token-level fact correction in abstractive summarization. *IEEE Access* **2023**, *11*, 1934–1943. [CrossRef]
18. Zhu, C.; Hinthorn, W.; Xu, R.; Zeng, Q.; Zeng, M.; Huang, X.; Jiang, M. Enhancing Factual Consistency of Abstractive Summarization. *arXiv* **2020**, arXiv2003.08612.
19. Jin, H.; Wang, T.; Wan, X. SemSUM: Semantic dependency guided neural abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020*; Volume 34, pp. 8026–8033. [CrossRef]
20. Nan, F.; Nallapati, R.; Wang, Z.; Santos, C.N.d.; Zhu, H.; Zhang, D.; McKeown, K.; Xiang, B. Entity-level Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 2727–2733. [CrossRef]
21. Nan, F.; Santos, C.N.d.; Zhu, H.; Ng, P.; McKeown, K.; Nallapati, R.; Zhang, D.; Wang, Z.; Arnold, A.O.; Xiang, B. Improving Factual Consistency of Abstractive Summarization via Question Answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 6881–6894. [CrossRef]
22. Mao, Q.; Li, J.; Peng, H.; He, S.; Wang, L.; Yu, P.S.; Wang, Z. Fact-Driven Abstractive Summarization by Utilizing Multi-Granular Multi-Relational Knowledge. *IEEE/ACM Trans. Audio Speech Lang. Proc.* **2022**, *30*, 1665–1678. [CrossRef]
23. Pernes, D.; Mendes, A.; Martins, A.F. Improving abstractive summarization with energy-based re-ranking. In *Proceedings of the Second Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 1–17. [CrossRef]
24. Wan, D.; Liu, M.; McKeown, K.; Dreyer, M.; Bansal, M. Faithfulness-Aware Decoding Strategies for Abstractive Summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 2864–2880. [CrossRef]
25. Zhu, C.; Hinthorn, W.; Xu, R.; Zeng, Q.; Zeng, M.; Huang, X.; Jiang, M. Enhancing Factual Consistency of Abstractive Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 718–733. [CrossRef]
26. Dixit, T.; Wang, F.; Chen, M. Improving Factuality of Abstractive Summarization without Sacrificing Summary Quality. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Toronto, ON, Canada*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 902–913. [CrossRef]
27. Cao, M.; Dong, Y.; Wu, J.; Cheung, J.C.K. Factual Error Correction for Abstractive Summarization Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 6251–6258. [CrossRef]
28. Dong, Y.; Wang, S.; Gan, Z.; Cheng, Y.; Cheung, J.C.K.; Liu, J. Multi-Fact Correction in Abstractive Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 9320–9331. [CrossRef]
29. Aguirre, A.M.G.; Robledo, S.; Zuluaga, M. Topic Modeling: Perspectives From a Literature Review. *IEEE Access* **2022**, *11*, 4066–4078. [CrossRef]
30. Mihalcea, R.; Tarau, P. TextRank: Bringing Order into Text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Barcelona, Spain, 2004; pp. 404–411.

31. Li, Z.; Peng, Z.; Tang, S.; Zhang, C.; Ma, H. Text Summarization Method Based on Double Attention Pointer Network. *IEEE Access* **2020**, *8*, 11279–11288. [CrossRef]
32. Zhuang, H.; Zhang, W. Generating Semantically Similar and Human-Readable Summaries with Generative Adversarial Networks. *IEEE Access* **2019**, *7*, 169426–169433. [CrossRef]
33. Yang, M.; Li, C.; Shen, Y.; Wu, Q.; Zhao, Z.; Chen, X. Hierarchical Human-like Deep Neural Networks for Abstractive Text Summarization. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 2744–2757. [CrossRef]
34. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
35. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. 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)*, Minneapolis, Minnesota; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186.
36. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 7871–7880.
37. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv* **2019**, arXiv:1910.10683.
38. Zhang, J.; Zhao, Y.; Saleh, M.; Liu, P.J. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. *Int. Conf. Mach. Learn* **2020**, *119*, 11328–11339.
39. Laban, P.; Schnabel, T.; Bennett, P.N.; Hearst, M.A. SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. *arXiv* **2021**, arXiv:2111.09525. [CrossRef]
40. Yu, W.; Zhu, C.; Li, Z.; Hu, Z.; Wang, Q.; Ji, H.; Jiang, M. A Survey of Knowledge-enhanced Text Generation. *ACM Comput. Surv.* **2022**, *54*, 227. [CrossRef]
41. Wang, A.; Cho, K.; Lewis, M. Asking and Answering Questions to Evaluate the Factual Consistency of Summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020. [CrossRef]
42. Heo, H. FactSumm: Factual Consistency Scorer for Abstractive Summarization. 2021. Available online: <https://github.com/Huffon/factsumm> (accessed on 20 April 2026).
43. Nair, S.; Rao, Y.S. Analysis of the evaluation metrics for the factual consistency of generative AI. In *2025 IEEE 2nd International Conference for Women in Computing (InCoWoCo), India*; IEEE: New York, NY, USA, 2025; pp. 1–6. [CrossRef]
44. Available online: <https://www.kaggle.com/pariza/bbc-news-summary> (accessed on 20 April 2026).
45. Rahman, S.; Raihan, M.; Akter, L.; Raihan, M.M.S. COVID-19 Real News Data of CBC News (1.0) [Data Set]. Zenodo. 2021. Available online: <https://zenodo.org/records/4722470> (accessed on 20 April 2026).
46. Hu, Y.; Ganter, T.; Deilamsalehy, H.; Dernoncourt, F.; Foroosh, H.; Liu, F. MeetingBank: A Benchmark Dataset for Meeting Summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023. [CrossRef]
47. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv* **2019**, arXiv:1904.09675.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.