JOURNAL OF SOCIAL COMPUTING
ISSN 2688-5255 03/06 pp132−144
Volume 5, Number 2, June 2024
DOI: 10.23919/JSC.2024.0010

Segmented Summarization and Refinement: A Pipeline for Long-Document Analysis on Social Media

Guanghua Wang*, Priyanshi Garg, and Weili Wu

Abstract: Social media's explosive growth has resulted in a massive influx of electronic documents influencing various facets of daily life. However, the enormous and complex nature of this content makes extracting valuable insights challenging. Long document summarization emerges as a pivotal technique in this context, serving to distill extensive texts into concise and comprehensible summaries. This paper presents a novel three-stage pipeline for effective long document summarization. The proposed approach combines unsupervised and supervised learning techniques, efficiently handling large document sets while requiring minimal computational resources. Our methodology introduces a unique process for forming semantic chunks through spectral dynamic segmentation, effectively reducing redundancy and repetitiveness in the summarization process. Contrary to previous methods, our approach aligns each semantic chunk with the entire summary paragraph, allowing the abstractive summarization model to process documents without truncation and enabling the summarization model to deduce missing information from other chunks. To enhance the summary generation, we utilize a sophisticated rewrite model based on Bidirectional and Auto-Regressive Transformers (BART), rearranging and reformulating summary constructs to improve their fluidity and coherence. Empirical studies conducted on the long documents from the Webis-TLDR-17 dataset demonstrate that our approach significantly enhances the efficiency of abstractive summarization transformers. The contributions of this paper thus offer significant advancements in the field of long document summarization, providing a novel and effective methodology for summarizing extensive texts in the context of social media.

Key words: long document summarization; abstractive summarization; text segmentation; text alignment; rewrite model; spectral embedding

1 Introduction

The proliferation of social media has seamlessly integrated electronic documents into our daily routines, influencing diverse aspects of life, from dining preferences to vacation selections. As millions of new documents incessantly populate the online sphere every month, social media platforms have evolved into an unparalleled hub of human communication. Nevertheless, the vastness of this digital repository poses considerable navigational challenges[1]. A substantial chunk of content circulating online is characterized by verbosity, complexity, and repetitiveness, making the extraction of valuable insights an intricate endeavor.

In this scenario, the crucial role of long document summarization comes into play. This technique is pivotal in the realm of social media, where it is tasked with distilling extensive texts—ranging from articles and blog posts to exhaustive threads—into concise and comprehensible summaries. Such summarization is not

• Guanghua Wang, Priyanshi Garg, and Weili Wu are with the Department of Computer Science, The University of Texas at Dallas, Richardson, TX 75080, USA. E-mail: guanghua.wang@utdallas.edu; priyanshi.garg@utdallas.edu; weiliwu@utdallas.edu.

* To whom correspondence should be addressed.
Manuscript received: 2023-11-27 ; revised: 2024-05-23; accepted: 2024-05-28

only imperative for facilitating rapid understanding and engagement among users but is also a non-trivial task that necessitates the careful extraction of core themes and critical information from a deluge of data. Crafting summaries that are both succinct and informative requires the deployment of advanced methodologies capable of handling extensive texts while ensuring clarity and spotlighting the main points effectively.

Long document summarization has been the focal point of a majority of research endeavors in the document summarization field. Researchers have employed a plethora of techniques for this purpose, which can be broadly categorized into five groups: Transformer-based models, Extractive-Abstractive approaches, Divide-and-Conquer strategies, Hierarchical Neural Networks, and Large Language Models (LLMs). Transformer-based models[2, 3], armed with self-attention mechanisms, are adept at grasping contextual nuances within and across documents, facilitating the generation of summaries that encapsulate the primary ideas and crucial details accurately. However, their application is often hampered by their demanding computational and resource requirements, especially for longer documents. Extractive-Abstractive techniques[4, 5], on the other hand, select specific sentences or phrases from the source documents, thereby maintaining the accuracy and integrity of the content in the resulting summaries. While Extractive-Abstractive approaches safeguard the crucial information from the original texts, there is an inherent risk of omitting important details if significant sentences are overlooked, which can subsequently diminish the completeness of the summaries. Divide-and-conquer strategies[6, 7] tackle this by breaking down long documents into smaller segments, paving the way for more manageable and scalable summarization while allowing for efficient resource utilization through parallel processing. However, this might result in a loss of coherence and connectivity between documents spanning different segments, leading to summaries that might seem fragmented or incomplete. Hierarchical models[8, 9] are designed to effectively encapsulate the hierarchical arrangement of information within the original document, thereby yielding a summary that is coherent and well-structured. Nonetheless, these models often require considerable computational resources for training and inference, with their complexity rendering them less scalable. Lastly, LLMs[10, 11], which are essentially Transformer-based models trained on vast datasets, excel in understanding and generating coherent and contextually appropriate text. However, their efficacy comes at the cost of substantial data, time, and computational resources, making them a less scalable option due to the extensive resources and complexity involved in processing lengthy documents.

Each of these approaches presents a unique set of advantages and trade-offs. While Transformer-based models are proficient in understanding context, they are resource-intensive. Extractive-Abstractive approaches provide a balance between content preservation and summarization but run the risk of losing valuable information. Divide-and-conquer methods offer scalability at the potential expense of coherence, and while hierarchical models capture structural nuances effectively, they are also resource-intensive. LLMs, despite their superior performance, necessitate extensive resources and extended training periods. Ultimately, the selection of the most apt approach hinges on the specific needs and considerations pertinent to the long document summarization task in question.

In this paper, we propose a unique three-stage pipeline for long document summarization. As illustrated in Fig. 1, our methodology encompasses spectral dynamic segmentation, the utilization of a compact Transformer-based model, and the application of a rewrite model. Unlike prior research[4–7], our approach melds both unsupervised and supervised learning techniques. This innovative blend not only caters to handling extensive document sets efficiently but also permits training with minimal computational overhead. Further enhancing our method, the generated summaries undergo a refining process via our rewrite model, elevating the overall quality of the output. We delineate the following contributions:

• This study marks the first academic exploration of long document summarization within the unique context of social media. By addressing specific challenges posed by these platforms, our research advances the field of Natural Language Processing (NLP), enhancing digital information representation and establishing a foundation for future developments.

• To tackle redundancy and repetitiveness in long

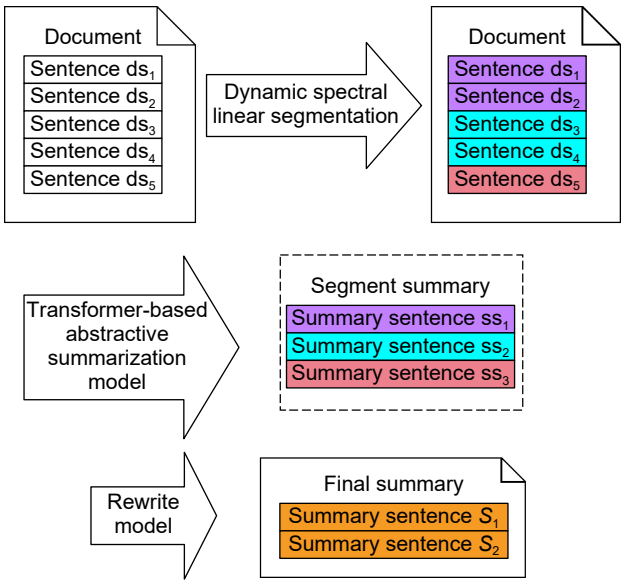

Fig. 1 Overview of the three-stage summarization pipeline. Stage 1: Spectral dynamic segmentation for semantic chunk creation. Stage 2: Alignment of chunks with summary for holistic integration. Stage 3: Application of a rewrite model to improve coherence and fluency of the final summary.

document summarization, we introduce a method for forming semantic chunks using dynamic signal segmentation[12], minimizing segment costs. Enhanced with spectral embedding[13], this technique distills each chunk into a succinct representation, maintaining essential content within a set length criterion.

• Unlike prior methods[7, 14], our approach pairs each chunk with the entire summary paragraph. This practical method allows for chunk length calibration within GPU memory limits, enabling transformers to process documents without truncation. It ensures effective summarization and reintegration of missing information from other chunks, leading to holistic summaries.

• We leverage Bidirectional and Auto-Regressive Transformers (BART)[3] to develop a rewrite model that refines and rearranges summary constructs, enhancing their syntactic fluidity and semantic cohesion. This pioneering use of an abstract summary as a training substrate for a rewrite model improves the quality of long document summarization.

• Utilizing the Webis-TLDR-17[15], Xsum dataset[16], and BART[3], our experiments on a single GPU with constrained memory demonstrate significant improvements in the efficiency of abstractive summarization transformers, surpassing the performance of larger models requiring more computational resources.

2 Background: State-of-the-Art in Long Document Summarization

This section provides a comprehensive review of the literature on long document summarization, abstractive summarization on social media, and text segmentation, highlighting advancements, challenges, and gaps. To enhance long document summarization, we developed a novel divide-and-conquer approach that integrates the strengths of text segmentation and abstractive summarization methods. Our method is designed to operate efficiently within limited computational resources, ensuring robust performance on extensive document sets. Additionally, our " abstractive then abstractive" framework produces comprehensive, contextually rich summaries, addressing the practical challenges of large-scale summarization and pushing the boundaries of output fidelity and resource efficiency.

2.1 Long document summarization

In the realm of text summarization, processing long documents poses a significant challenge. Common benchmark datasets define " long documents" as those with an average length spanning thousands of lexical tokens. The root of this challenge can be traced back to the inherent limitations of contemporary summarization systems. Specifically, many of today's cutting-edge models, even those that are pre-trained, have constraints on the number of lexical tokens they can handle, typically capped between 512 and 1024 tokens.

The crux of the matter is how to adeptly summarize these expansive documents that contain thousands of tokens. Overcoming this challenge is not just academic; it holds considerable practical implications. To navigate this challenge, researchers have embarked on a variety of strategies. Some approaches hinge on tapping into the intrinsic structure of documents[17], while others adopt a blend of compression and selection techniques[18] before delving into abstractive summarization. It is pivotal to highlight that the efficacy of these methods often stands on the shoulders of substantial training data.

Peeling back the layers of contemporary summarization methodologies reveals a significant reliance on Transformer-based models. These models,

designed for extended sequence processing, have seen iterative advancements. A notable example is the Longformer[19], which revamps the Transformer's self-attention mechanism with a dilated sliding window attention, optimizing both computational and memory efficiencies. Parallel to this, there is a surge in exploring content-based and temporal sparse attention mechanisms[20, 21]. Not to be sidelined, hierarchical attention mechanisms[22] also show promise, offering a layered representation approach that scales from individual words to complete documents.

Recently, Lim and Song[23] introduced a novel multi-stage long document summarization method that generates new relevant segment pairs and utilizes contrastive learning to enhance the training of coarse summarizers. In a similar vein, Pang et al.[24] developed a summarization approach that tackles long-distance context and quadratic complexity by assuming a hierarchical latent structure, enabling efficient updates of token representations through both bottom-up and top-down processes. Further innovations include Mixtral 8×7B (pre-train)[25], a sparse mixture of experts language model that employs a router network to selectively activate only two out of eight feedforward blocks (experts) at each layer for processing each token. This design allows for an efficient use of computational resources for long documents by activating only a subset of its total parameters per token. Additionally, Llama 3[26] incorporates a tokenizer for efficient encoding and grouped query attention, which enhances inference efficiency and is specifically designed to excel in various real-world scenarios while being resource-efficient. Collectively, these innovations represent the academic community's relentless pursuit to refine summarization models, enabling them to gracefully tackle the complexities of long documents.

2.2 Abstractive summarization on social media

Abstractive summarization plays a crucial role in enhancing user experience on social media platforms by generating concise and coherent representations of extensive content. Unlike its counterpart, extractive summarization, which merely picks out significant sentences, abstractive methods rephrase and condense the original information, delivering a more streamlined understanding. Given the voluminous influx of data on social media, it is challenging for users to quickly absorb the core message of lengthy posts or threads. This is where abstractive summarization becomes indispensable, allowing users to grasp the main ideas without delving into the full content. Such succinct summaries not only bolster user engagement and interactions but also facilitate the rapid dissemination of vital information to a wider audience.

Several research initiatives have been undertaken to advance this field. For instance, Syed et al.[27] documented the Webis-TLDR-17 Challenge, which emphasized the development of abstractive summarization technology for social media. Gehrmann et al.[28] explored transfer learning from pre-trained language models to produce genuinely abstractive summaries. They pointed out that conventional performance metrics might not truly capture a model's capabilities in terms of language abstraction and summarization. In a similar vein, Choi et al.[29] introduced the Unified VAE-PGN model, leveraging a Variational Autoencoder (VAE) to amplify the diversity of summaries, ensuring they are both pertinent and compact.

2.3 Text segmentation

Text segmentation, an essential mechanism in text processing, facilitates the division of text into coherent units, thereby optimizing comprehension and analytical capabilities. This technique is crucial for enhancing readability and serves as the bedrock for comprehensive textual analysis and summarization endeavors. Broadly, there are two primary methodologies: rule-based and machine learning based segmentation. Rule-based strategies deploy newline characters, indentations, and formatting indicators to demarcate paragraph boundaries, though their efficacy wanes with inconsistently formatted content. In contrast, machine learning based segmentation relies on supervised models which are primed using textual formatting combined with lexical and syntactic markers, ensuring precise paragraph identification.

Diving deeper into segmentation methodologies, linear and hierarchical techniques emerge as two prominent subfields. Linear segmentation[30], as championed by researchers such as Hearst with the TextTiling algorithm[31, 32] and Choi's probabilistic model[33], adopts a sequential approach to discern text

sections. Both methods hinge on assessing segment resemblance. Other influential works in this domain, such as those by Utiyama and Isahara[34] and Fragkou et al.[35], utilize dynamic programming to streamline segmentation. Meanwhile, hierarchical segmentation provides a more detailed view by subdividing primary sections into smaller sub-segments. The pioneering work in this area is exemplified by Eisenstein's research[36]. In contrast, Glavaš et al.[37] employed word embeddings and measures of semantic relatedness between short texts to create a semantic relatedness graph for the document. The coherent segments are then generated from the maximal cliques within this relatedness graph.

In the domain of text segmentation, recent years have marked a dramatic pivot towards end-to-end neural network models. Traditional reliance on linguistic annotations has waned, making way for the powerful extraction capabilities of pre-trained models in procuring word or sentence embeddings. A notable exemplar of this paradigm shift is the work by Li et al.[38], which employs the GloVe word embeddings for delineating Elementary Discourse Unit (EDU) segmentations. Concurrently, Wang et al.[39] integrated the prowess of Embeddings from Language Models (ELMO) pre-trained embeddings with the CRF-Bi-LSTM framework, ushering in groundbreaking achievements in the field. Furthering this momentum, Liu et al.[40] introduced an innovative segmentation-based language generation model. This model taps into powerhouse pre-trained platforms like BART or GPT to partition news articles into distinct sections and craft corresponding summaries. This methodical segmentation not only offers a structured layout but also alleviates the challenge of sifting through extensive articles.

3 Proposed Method: A Three-Stage Pipeline for Summarization

This section details the methodology employed in this study, presenting a step-by-step description of the three-stage pipeline developed for effective long document summarization. Starting with spectral dynamic segmentation for semantic chunk formation, we then explain the process of aligning summaries with segments and describe the algorithm that underpins our approach. This section culminates in a discussion of the abstractive summarization and rewrite model, showcasing the innovative techniques used to enhance summary generation.

3.1 Spectral dynamic segmentation

To effectively train transformers for summarizing extensive inputs without truncation while managing GPU memory constraints, it is imperative to incorporate two crucial components.

The first component involves the utilization of a pre-trained language model that serves as the foundation for capturing the semantic essence of sentences or paragraphs. In the course of our experimentation, we engaged the BART model, which underwent fine-tuning processes on the Webis-TLDR-17 dataset. This dataset is substantial, encompassing over four million sentence pairs. The BART model, trained under a self-supervised contrastive learning objective, functioned as an encoder for various text units, including sentences, paragraphs, and concise passages. Through this process, the model generated vectors embodying the semantic details of the provided text.

The second vital component in our approach is the employment of a spectral-embedding dynamic segmentation algorithm. This algorithm facilitates the segmentation of original documents into uniquely identifiable chunks.

Spectral embedding is rooted in the principles of spectral graph theory, where the spectrum, particularly the eigenvalues of matrices linked with a graph, most notably the graph Laplacian, holds significant importance. This technique specializes in transforming higher-dimensional data into a compact, lower-dimensional space, ensuring the preservation of the data's pairwise affinities or relationships. Essentially, spectral embedding capitalizes on the spectral decomposition of the graph Laplacian to achieve this dimensionality reduction. In our work, taking cues from dynamic change point segmentation, we employed spectral embedding to condense the dimensionality of sentence vectors effectively.

Subsequently, dynamic change point segmentation algorithms were applied to these reduced-dimension vectors to extract the text chunks. For the purpose of contrasting the vector representations of pairs of sentences, cosine similarity was utilized as the comparison metric. Under this framework, text

segmentation is effectively transformed into a task of kernel change point detection[12]. Consider $y=\{y_0,y_1,...,y_{T-1}\}$, a d-dimensional signal with T samples. Employing a cost function $c(\cdot)$ derived from a kernel $k(\cdot,\cdot)$, the cost of the sub-signal $y_{a..b}$ can be defined as

$$c(y_{a..b})=\sum_{t=a}^{b-1}G_{t,t}-\frac{1}{b-a}\sum_{a\leqslant s<b}\sum_{a\leqslant t<b}G_{s,t} \quad (1)$$

where $y_{a..b}$ represents the vectors $\{y_a,y_{a+1},...,y_{b-1}\}$ and $G_{s,t}:=k(y_s,y_t)$. Consequently, $(G_{s,t})_{s,t}$ is the Gram matrix of y. Employing a dynamic change point segmentation algorithm, we aim to minimize the sum of costs by evaluating the cost associated with all sub-sequence possibilities of a given vector, which in this context, refers to the sentence spectral embedding. This dynamic segmentation approach exhibits a complexity of $O(CKn^2)$, where K represents the number of change points to be detected, n is sample count, and C is indicative of the complexity involved in invoking the cost function on a specific sub-signal. Distinctively, our method dynamically assigns the number of chunks, ensuring that each chunk's token count aligns with the limitations imposed by the model's training size. This dynamic assignment guarantees that the token count within each segment is sufficiently small to fit within the model's training size constraints. As a result, this approach obviates the need for document truncation during the summarization phase, even when handling documents of substantial length with limited computational resources available.

In conclusion, our innovative text segmentation approach leverages a trained language model to produce semantically rich sentence embeddings. These embeddings are crucial in creating meaningful text segments, enhancing the quality and coherence of text segmentation. Drawing from spectral graph theory, our method efficiently condenses high-dimensional data into a more manageable space, preserving the intrinsic relationships within the data. Through dynamic change point segmentation applied to these compacted vectors, we can seamlessly divide the text into distinct chunks. A standout feature of our approach is its ability to dynamically determine chunk sizes based on the model's training constraints, ensuring that even long documents can be summarized without the need for truncation, regardless of computational resource limitations.

3.2 Align summary with segments

In the training phase of supervised machine learning for abstractive summarization models, a target passage is crucial. As demonstrated in Fig. 2, our approach to establishing meaningful source-target pairs involves a syntactic assignment technique that significantly deviates from previous methods.

Previously, each sentence within the target summary was paired with a corresponding passage that yielded the highest similarity score between them. In contrast, our refined approach first identifies segments that possess the highest scores. Following this, we align each of these high-scoring segments with the entire summary passage. This nuanced alignment technique fosters a model focus that extends beyond summarization; it enables the model to infer and integrate information absent in individual segments but present in the whole original document. This method enhances the model's ability to draw comparisons between segments and the comprehensive document effectively.

The drive behind adopting this innovative approach is trifold. Firstly, the use of semantic embeddings offers richer representations of sentences derived from both the original document and the summary, enhancing the depth and quality of the informational base the model interacts. Secondly, through score normalization within a consistent space, we enable the precise and reliable comparison between various pairs

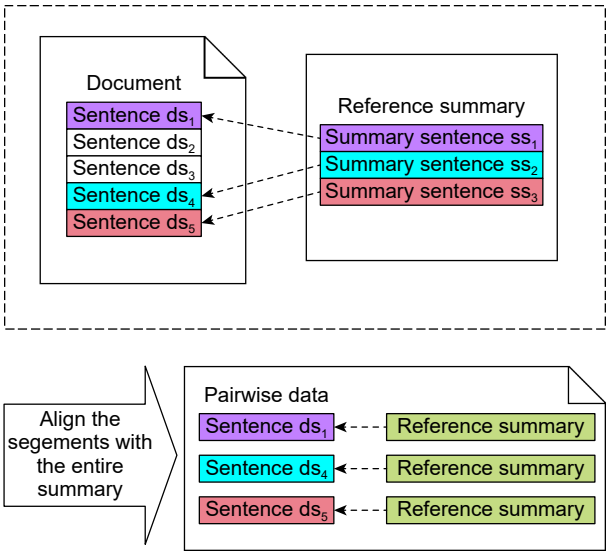

Fig. 2 Alignment strategy highlighting integration of highest-rated segments with entire summary to guarantee summarization completeness and precision.

of segments and passages. This ensures a consistent mechanism for evaluating and aligning different text segments. Lastly, by exposing the model to the entirety of the summary passage, it gains access to a comprehensive set of related information. This expanded informational view significantly augments the model's ability to perform accurate deductions and inferences, thereby improving the overall effectiveness and accuracy of the summarization process.

Through the integration of these motivating factors, our approach optimizes the alignment process between passages and summary sentences. Such optimized alignment is instrumental in effectively training abstractive summarization models, paving the way for more accurate and coherent summarized outputs.

3.3 Segmentation and summarization algorithm

For the task of long document summarization, a document d is initially split into individual sentences, yielding a set of sentences denoted as ds_m. The true summary of this document is comprised of sentences ss_n. It is imperative to define the upper size limits, represented as U_s, for a segment in the summarization process.

Our approach to creating a segment, seg_i, and associating it with the complete summary, while utilizing the highest similarity score with the corresponding ss_i, can be outlined in steps as per Algorithm 1. Below is an overview of the procedure:

• **Segment determination:** The BART model, fine-tuned prior to these operations, is employed to generate sentence embeddings. These embeddings are subsequently utilized as inputs for the spectral dynamic segmentation algorithm. Initially, the segment number, represented as SEG_{num}, is set to 1. The algorithm assesses whether any segment sizes surpass the upper limit, U_s. If they do, an additional segment, $\text{SEG}_{\text{num}} + 1$, is necessitated. If all segment sizes adhere to the upper limit, the segmentation process concludes.

• **Target alignment:** Post- segment creation, target alignment (as detailed in Algorithm 2) commences. Here, every summary sentence, ss_k, is compared against each segment. The segment maximizing the cosine similarity score, as per the fine-tuned BART model, is then assigned the complete summary. It is crucial to note that segments without assigned targets are excluded from subsequent training phases.

Algorithm 1 Spectral dynamic segmentation

Input: model ← BART;
document_sentences ← $[\text{ds}_0, \text{ds}_1, ..., \text{ds}_m]$;
summary_sentences ← $[\text{ss}_0, \text{ss}_1, ..., \text{ss}_k]$;
Parameters: U_s ← upper_size
Output: Return the segmentation-summary alignments

Let $\text{SEG}_{\text{num}} \leftarrow 1$
$[\text{enc\_ds}_0, \text{enc\_ds}_1, ..., \text{enc\_ds}_m] \leftarrow$ model.encode($[\text{ds}_0, \text{ds}_1, ..., \text{ds}_m]$)
$[\text{emb\_ds}_0, \text{emb\_ds}_1, ..., \text{emb\_ds}_m] \leftarrow$ Spectral_embedding ($[\text{enc\_ds}_0, \text{enc\_ds}_1, ..., \text{enc\_ds}_m]$)
while True **do**
  Segments←Dynamic_Segmentation($[\text{emb\_ds}_0, \text{emb\_ds}_1, ..., \text{emb\_ds}_m]$)
  for Segment in Segments **do**
    if (len(Segment) > U_s) **then**
      SEG_{num} + +
      Continue while Loop (go Line 4)
    end if
  end for
  Break
end while
summaries ← Targets alignment (Algorithm 2)
return (segments, summaries)

Algorithm 2 Targets alignment

Input: model ← BART; Segments ← $[\text{seg}_0, \text{seg}_1, ..., \text{seg}_i]$; summary_sentences ← $[\text{ss}_0, \text{ss}_1, ..., \text{ss}_k]$
Output: Return the targets of the segments

$[\text{enc\_ss}_0, \text{enc\_ss}_1, ..., \text{enc\_ss}_k] \leftarrow$ model.encode($[\text{ss}_0, \text{ss}_1, ..., \text{ss}_k]$)
$[\text{enc\_seg}_0, \text{enc\_seg}_1, ..., \text{enc\_seg}_i] \leftarrow$ model.encode($[\text{seg}_0, \text{seg}_1, ..., \text{seg}_i]$)
Let tagets←$[t_0 \leftarrow [], t_1 \leftarrow [], ..., t_i \leftarrow []]$
for enc_ss_k **do**
  Let scores←[]
  for enc_seg_m **do**
    score ← cosine_sim($\text{enc\_ss}_k, \text{enc\_seg}_i$)
    scores.append(score)
  end for
  idx ← argmax(scores)
  targets[idx] ← summary
end for
return targets

Through the outlined algorithms and steps, this approach efficiently creates segments and aligns them with the respective summary sentences, ensuring an effective training process for long document summarization tasks.

3.4 Abstractive summarization and rewrite model

In our study, we focused on evaluating the efficiency of transformer models in summarization tasks, specifically utilizing BART[3], under the constraints of limited GPU memory. We selected BART for two primary reasons: the availability of a public base version checkpoint and its unique design that emphasizes sequence masking and reconstruction. Such a design makes BART an ideal choice for generative tasks like summarization, rewriting, and deduction, as it is well-equipped to handle these operations efficiently.

The process begins with the training phase of the abstractive summarization model, during which preliminary summary sentences are generated from each segment. However, the initial summaries may exhibit issues like suboptimal sequencing, redundancy, or excessive length, depending on the segmentation algorithm used. These issues necessitate further refinement to produce concise and coherent final summaries.

To rectify the identified issues in the initial summaries, we introduced a rewrite model. This model is designed to refine and enhance the preliminary summaries generated in the initial phase. We chose BART as the underlying architecture for training this rewrite model, leveraging its strong capabilities in sequence processing tasks.

The workflow is as follows: firstly, the abstractive summarization model generates summary outputs from the training dataset. These preliminary summaries then serve as inputs for the rewrite model, which refines them into the final summaries. Through empirical evaluations, we found that the rewrite model significantly improves the coherence and quality of the final summaries, thereby confirming the efficacy of our proposed approach in enhancing summarization outcomes.

4 Evaluation: Empirical Result and Comparative Analysis

In Section 4, we present the results of our empirical studies conducted to evaluate the efficiency and effectiveness of our summarization approach. This section is structured to first introduce the dataset and experimental settings, followed by a comparison with baseline models. The results are discussed in detail, highlighting the performance of our method across various metrics and demonstrating its superiority in generating concise and coherent summaries for long documents.

4.1 Dataset

Our research leverages a comprehensive approach to evaluating long document summarization through the utilization of two distinct datasets: the Webis-TLDR-17[15] dataset and the Xsum[16] dataset. This dual-dataset methodology is designed to rigorously assess the effectiveness of our segmented summarization and refinement pipeline across a variety of document types and lengths.

We focused on documents from the Webis-TLDR-17 dataset that exceed 1024 tokens in length. Out of a total of 97 749 documents that meet this criterion, we implemented an 8:1:1 split ratio for training, validation, and testing. This resulted in 78 199 documents for training, 9775 for validation, and 9775 for testing. The Webis-TLDR-17 dataset, specifically curated for social media summarization challenges, offers a robust and diverse collection of texts for comprehensive analysis.

To complement our analysis, we incorporated the Xsum dataset, known for its concise summary challenges. Selecting documents with word counts exceeding 1024, we allocated 5057 documents for training, 632 for validation, and 633 for testing. The Xsum dataset introduces a different dimension to our research, emphasizing the creation of brief summaries from extensive texts.

During the initial phase of our study, we focused on training data from both datasets to refine our understanding and application of the spectral dynamic segmentation algorithm. This algorithm, coupled with precise sentence boundary detection, enabled us to efficiently process and summarize documents of varied lengths. In the evaluation phase, the model's proficiency was rigorously tested using the designated test datasets.

The integration of the Xsum dataset alongside Webis-TLDR-17 enriches our experimental framework, allowing us to examine our summarization approach's versatility and effectiveness across distinct types of content and summarization requirements. By leveraging these datasets, we aim to demonstrate the adaptability and robustness of our methodology in

effectively summarizing long documents in various contexts.

4.2 Experiment setting

For sentence embedding dimension reduction, we employed the spectral embedding method available through the sklearn library, subsequently applying dynamic segmentation to partition the original document into multiple chunks. Initially, the number of segments was established at one, with the upper limit set at 20; this upper limit was determined based on the maximal count of sentences in the reference summaries within the dataset.

In our approach to dynamic change point detection for text segmentation, we have configured the parameters based on the nature of our dataset. Specifically, for each document, the range for change points is set between 2 and 20. This translates to a document being divided into a minimum of 3 chunks and a maximum of 21 chunks. This particular configuration draws intuitive inspiration from the document lengths observed in the Webis-TLDR-17 dataset. Furthermore, to ensure coherent and meaningful segments, we have mandated that each segment comprises a minimum of 3 sentences, achieved by setting the minimum jump between sentences in each segment to 2. Our segmentation parameters were empirically validated by conducting experiments on both the Webis-TLDR-17 and Xsum datasets. We varied the chunk sizes and number of segments to determine the configuration that balanced segment coherence and summary length. The results demonstrated that our chosen parameters effectively optimized the trade-off between segmentation granularity and summary coherence.

In our exploration of abstractive summarization, the BART transformer architecture was implemented. The input documents were segmented into specified sizes, with size 1024 allocated for BART. The training of the BART model extended over two epochs, using parameters of a beam width of 2 and a length penalty of 2.

Regarding the rewrite model, BART was also utilized, configured at size 1024, and trained over a period of two epochs. The auxiliary parameters for this phase were kept consistent with those outlined for the abstractive summarization model.

To assess the precision and overall quality of the produced summaries, we utilized the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) score metric, first proposed by Lin[41] in 2004, along with the BERTScore, introduced by Zhang et al.[42] in 2019. These metrics facilitated a thorough evaluation and comparison of the summaries' quality generated throughout our research.

4.3 Baseline

As this study is pioneering, being the first scholarly attempt to explore the realm of long document summarization specifically in the context of social media, there is not a body of previous work or established research in this exact field to refer to or build upon. The absence of prior research in this domain renders our work groundbreaking and essential for scholars and practitioners aiming to delve into this specialization in the future. In laying the groundwork for this uncharted area of study, it became crucial to not only select a reliable and powerful baseline model to underpin our research but also to establish a comprehensive comparative analysis framework.

To enrich our analysis, we integrated a suite of widely recognized baseline models, including Lead-3, TextRank[43], C-Attention[44], PEGASUS (Large)[45], BART (Large), Mixtral 8×7B (pre-train)[25], and Llama 3 8B (pre-train)[26], each offering unique approaches to the summarization challenge. This diverse set of baselines, ranging from simple extractive methods like Lead-3 to advanced abstractive techniques as embodied by PEGASUS, BART, Mixtral, and Llama 3, allows us to evaluate our segmented summarization and refinement approach across a spectrum of strategies. The inclusion of these models serves to benchmark the performance of our proposed method, with evaluation metrics such as ROUGE scores providing a quantitative measure of summarization quality in comparison to these established techniques.

This multi-model comparative analysis enhances the validity and reliability of our pioneering research on long document summarization in the sphere of social media, setting a foundational benchmark for future explorations in this novel domain.

4.4 Result

The comprehensive analysis of experimental data presented in Tables 1 and 2 distinctly highlights the

Table 1 ROUGE scores and BERTscore comparison for models on extended documents exceeding 1024 tokens.

Model	Webis-TLDR-17 dataset					Xsum dataset				
	R-1	R-2	R-SU	R-L	BERTScore	R-1	R-2	R-SU	R-L	BERTScore
Lead-3	17.01	2.36	4.42	10.76	51.67	17.36	2.36	3.72	11.32	56.70
TextRank	17.96	2.35	4.91	11.00	51.52	16.95	2.55	3.72	11.78	56.18
C-Attention	17.49	2.02	4.46	10.61	52.03	15.59	1.96	3.25	10.50	55.41
PEGASUS (Large)	18.84	4.11	5.39	13.60	49.32	26.71	9.09	8.13	19.76	65.26
BART (Large)	21.95	5.57	7.11	15.51	56.17	29.53	10.05	8.85	21.52	66.07
Mixtral 8×7B	15.94	2.56	4.28	9.93	50.67	14.38	3.83	3.78	9.78	55.65
Llama 3 8B	12.69	2.18	3.39	7.76	48.46	10.07	2.78	2.68	7.16	52.49
Top-scored with rewrite model	21.34	5.35	6.73	15.30	55.03	29.30	10.91	9.37	21.98	66.35
Entire summary without rewrite model	19.91	5.22	6.72	11.97	53.37	20.64	7.03	6.45	14.75	57.47
Entire summary with rewrite model	**22.32**	**7.01**	**8.47**	**16.48**	**56.91**	**30.19**	**11.18**	**9.73**	**22.56**	**66.55**

Note: ROUGE-1 (R-1) refers to the overlap of single words, ROUGE-2 (R-2) refers to the overlap of two-word sequences, and so on. ROUGE-SU (R-SU) refers to the metric for Skip-Bigram plus Unigram. ROUGE-L (R-L) measures the longest common subsequence.

Table 2 Output length comparison for models on extended Webis-TLDR-17 and Xsum documents (exceeding 1024 tokens).

Model	Output length	
	Webis-TLDR-17	Xsum
Reference summary	61	24
Lead-3	51	64
TextRank	64	55
C-Attention	61	59
PEGASUS (Large)	55	43
BART (Large)	43	24
Mixtral 8×7B	148	136
Llama 3 8B	274	255
Top-scored with rewrite model	32	26
Entire summary without rewrite model	188	90
Entire summary with rewrite model	37	26

superior performance of our method over alternative summarization methods applied to extended documents in the Webis-TLDR-17 and Xsum datasets.

A critical evaluation rooted in Table 1's ROUGE scores and BERTScore vividly illustrates the unparalleled advantage of incorporating a rewrite mechanism into the summarization process. Our pipeline approach not only surpasses its counterparts in all evaluative metrics but also marks a significant advancement in summarization quality. This enhancement, primarily attributed to the sophisticated rewrite process, transforms summaries from basic extracts to rich, cohesive narratives that closely mirror the original text's depth and nuance. The comparison starkly demonstrates that the absence of a rewrite feature in other models significantly detracts from their performance, manifesting in lower coherence, accuracy, and overall summary quality.

Moreover, the exploration of summary lengths in Table 2 underscores the unique capacity of our method to reconcile the often conflicting goals of brevity and comprehensive detail. While reference summaries set a standard for succinctness, our approach adeptly achieves a near-ideal balance, aligning closely with the reference benchmarks in terms of length while ensuring the retention of critical information and readability. This model's proficiency in producing concise yet complete summaries starkly contrasts with the verbosity seen in models lacking the rewrite feature, which, despite their detail, potentially compromise reader engagement due to their extended lengths.

The empirical data conclusively reinforce the standing of our pipeline approach as a benchmark of both efficiency and effectiveness within the realm of long document summarization. It unequivocally demonstrates that the inclusion of a rewriting phase is not merely beneficial but essential for generating high-quality summaries. This model's ability to distill extensive documents into their most salient points, without sacrificing detail for brevity, underscores the rewrite process's critical role in setting new benchmarks for summarization excellence. The analysis, therefore, not only showcases the significant advantages of our approach but also establishes it as a cornerstone technique for future advancements in the field of document summarization.

5 Conclusion and Future Work

This research paper presented an innovative approach to long document summarization on social media platforms, with a specific focus on documents surpassing 1024 tokens. Through the strategic employment of the segmentation and the introduction of a novel rewrite mechanism, our approach demonstrated commendable performance in generating coherent, concise, and relevant summaries. The experimental results, particularly the superior ROUGE scores achieved with the rewrite model, underscore the efficacy of our proposed methodology. Furthermore, the concise length of the summaries generated with the rewrite model suggests that our approach is adept at distilling lengthy documents into their essential elements without sacrificing the integrity of the information conveyed.

In light of the promising results garnered from this research, several avenues for future work have been identified:

• While the rewrite model has shown significant improvements in summarization performance, there is room for further refinement. Future research could develop advanced rewrite algorithms that provide more sophisticated text refinement and restructuring for social media while preserving the original meaning. This will enhance the accuracy of long document summarization and bring generated summaries closer to the ground truth.

• The current approach could be improved by incorporating deeper semantic analysis and understanding into the summarization process. This enhancement would ensure that the generated summaries are not only syntactically correct but also semantically coherent and relevant. Additionally, it would make text segmentation more meaningful and accurate, leading to higher quality summaries.

• While this research focused on long documents, the proposed approach should be optimized for documents of varied lengths and complexities. Developing an adaptive model that can dynamically adjust to the source document's length and depth would be invaluable for diverse summarization scenarios.

• Future studies should incorporate evaluations involving real users to gain qualitative insights into the readability, usefulness, and acceptability of the generated summaries. User-centric evaluations will provide a more holistic understanding of the model's performance in practical, real-world settings.

• With social media being a rich source of not only text but also images, videos, and other multimedia content, future research should develop summarization techniques that integrate and interpret multi-modal data. This presents an exciting and challenging prospect, aiming to create comprehensive summaries that encompass various data forms.

Through the exploration and development of these identified areas, we anticipate contributing to the advancement and refinement of long document summarization techniques, enhancing their applicability and effectiveness in the ever-evolving domain of social media.

Acknowledgment

This work was supported by the National Science Foundation (NSF) (Nos. 1822985 and 1907472).

References

[1] L. Dong, M. N. Satpute, W. Wu, and D. Z. Du, Two-phase multidocument summarization through content-attention-based subtopic detection, *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 6, pp. 1379–1392, 2021.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, arXiv preprint arXiv: 1706.03762, 2017.

[3] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, Virtual Event, 2020, pp. 7871–7880.

[4] N. Zmandar, A. Singh, M. El-Haj, and P. Rayson, Joint abstractive and extractive method for long financial document summarization, in *Proc. 3rd Financial Narrative Processing Workshop*, Lancaster, UK, 2021, pp. 99–105.

[5] J. Pilault, R. Li, S. Subramanian, and C. Pal, On extractive and abstractive neural document summarization with transformer language models, in *Proc. 2020 Conf. Empirical Methods in Natural Language Processing* (*EMNLP*), Virtual Event, 2020, pp. 9308–9319.

[6] S. Cho, K. Song, X. Wang, F. Liu, and D. Yu, Toward unifying text segmentation and long document summarization, in *Proc. 2022 Conf. Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, 2022, pp. 106–118.

[7] Y. Zhang, A. Ni, Z. Mao, C. H. Wu, C. Zhu, B. Deb, A. Awadallah, D. Radev, and R. Zhang, Summ^N: A multi-stage summarization framework for long input dialogues

and documents: A multi-stage summarization framework for long input dialogues and documents, in *Proc. 60th Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), Dublin, Ireland, 2022, pp. 1592–1604.

[8] S. Cao and L. Wang, HIBRIDS: Attention with hierarchical biases for structure-aware long document summarization, in *Proc. 60th Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), Dublin, Ireland, 2022, pp. 786–807.

[9] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, and N. Goharian, A discourse-aware attention model for abstractive summarization of long documents, in *Proc. 2018 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2* (*Short Papers*), New Orleans, LA, USA, 2018, pp. 615–621.

[10] OpenAI, GPT-4 technical report, Technical report, OpenAI, San Francisco, CA, USA, 2023.

[11] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., LLaMA: Open and efficient foundation language models, arXiv preprint arXiv: 2302.13971, 2023.

[12] C. Truong, L. Oudre, and N. Vayatis, Selective review of offline change point detection methods, *Signal Process.*, vol. 167, p. 107299, 2020.

[13] M. Belkin and P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, 2003.

[14] G. Moro and L. Ragazzi, Semantic self-segmentation for abstractive summarization of long documents in low-resource regimes, *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 10, pp. 11085–11093, 2022.

[15] M. Völske, M. Potthast, S. Syed, and B. Stein, TL;DR: Mining reddit to learn automatic summarization, in *Proc. Workshop on New Frontiers in Summarization*, Copenhagen, Denmark, 2017, pp. 59–63.

[16] S. Narayan, S. B. Cohen, and M. Lapata, Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization, in *Proc. 2018 Conf. Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018, pp. 1797–1807.

[17] Y. Liu, A. Ni, L. Nan, B. Deb, C. Zhu, A. H. Awadallah, and D. Radev, Leveraging locality in abstractive text summarization, in *Proc. 2022 Conf. Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, 2022, pp. 6081–6093.

[18] A. Bajaj, P. Dangati, K. Krishna, P. A. Kumar, R. Uppaal, B. Windsor, E. Brenner, D. Dotterrer, R. Das, and A. McCallum, Long document summarization in a low resource setting using pretrained language models, in *Proc. 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int. Joint Conf. Natural Language Processing: Student Research Workshop*, Virtual Event, 2021, pp. 71–80.

[19] I. Beltagy, M. E. Peters, and A. Cohan, Longformer: The long-document transformer, arXiv preprint arXiv: 2004. 05150, 2020.

[20] A. Roy, M. Saffar, A. Vaswani, and D. Grangier, Efficient content-based sparse attention with routing transformers, *Trans. Assoc. Comput. Linguist.*, vol. 9, pp. 53–68, 2021.

[21] M. Zaheer, G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, et al., Big bird: Transformers for longer sequences, arXiv preprint arXiv: 2007.14062, 2020.

[22] X. Zhang, F. Wei, and M. Zhou, HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization, in *Proc. 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 5059–5069.

[23] J. Lim and H. J. Song, Improving multi-stage long document summarization with enhanced coarse summarizer, in *Proc. 4th New Frontiers in Summarization Workshop*, Singapore, 2023, pp. 135–144.

[24] B. Pang, E. Nijkamp, W. Kryscinski, S. Savarese, Y. Zhou, and C. Xiong, Long document summarization with top-down and bottom-up inference, in *Proc. Findings of the Association for Computational Linguistics: EACL 2023*, Dubrovnik, Croatia, 2023, pp. 1267–1284.

[25] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, et al., Mixtral of experts, arXiv preprint arXiv: 2401.04088, 2024.

[26] Meta, Build the future of AI with Meta Llama 3, https://llama.meta.com/llama3/, 2024.

[27] S. Syed, M. Völske, N. Lipka, B. Stein, H. Schütze, and M. Potthast, Towards summarization for social media—Results of the TL;DR challenge, in *Proc. 12th Int. Conf. Natural Language Generation*, Tokyo, Japan, 2019, pp. 523–528.

[28] S. Gehrmann, Z. Ziegler, and A. Rush, Generating abstractive summaries with finetuned language models, in *Proc. 12th Int. Conf. Natural Language Generation*, Tokyo, Japan, 2019, pp. 516–522.

[29] H. Choi, L. Ravuru, T. Dryjański, S. Rye, D. Lee, H. Lee, and I. Hwang, VAE-PGN based abstractive model in multi-stage architecture for text summarization, in *Proc. 12th Int. Conf. Natural Language Generation*, Tokyo, Japan, 2019, pp. 510–515.

[30] T. Brants, F. Chen, and I. Tsochantaridis, Topic-based document segmentation with probabilistic latent semantic analysis, in *Proc. 11th Int. Conf. Information and Knowledge Management*, McLean, VA, USA, 2002, pp. 211–218.

[31] M. A. Hearst, Multi-paragraph segmentation of expository text, arXiv preprint arXiv: cmp-lg/9406037, 1994.

[32] M. A. Hearst, TextTiling: Segmenting text into multi-paragraph subtopic passages, *Comput. Linguist.*, vol. 23, no. 1, pp. 33–64, 1997.

[33] F. Y. Y. Choi, Advances in domain independent linear text segmentation, arXiv preprint arXiv: cs/0003083, 2000.

[34] M. Utiyama and H. Isahara, A statistical model for domain-independent text segmentation, in *Proc. 39th Annual Meeting on Association for Computational Linguistics*, Toulouse, France, 2001, pp. 499–506.

[35] P. Fragkou, V. Petridis, and A. Kehagias, A dynamic programming algorithm for linear text segmentation, *J. Intell. Inf. Syst*., vol. 23, no. 2, pp. 179–197, 2004.

[36] J. Eisenstein, Hierarchical text segmentation from multi-scale lexical cohesion, in *Proc. Human Language*

Technologies: The 2009 Annual Conf. North American Chapter of the Association for Computational Linguistics, Boulder, CO, USA, 2009, pp. 353–361.
[37] G. Glavaš, F. Nanni, and S. P. Ponzetto, Unsupervised text segmentation using semantic relatedness graphs, in *Proc. 5th Joint Conf. Lexical and Computational Semantics*, Berlin, Germany, 2016, pp. 125–130.
[38] J. Li, A. Sun, and S. Joty, SEGBOT: A generic neural text segmentation model with pointer network, in *Proc. 27th Int. Joint Conf. Artificial Intelligence*, Stockholm, Sweden, 2018, pp. 4166–4172.
[39] Y. Wang, S. Li, and J. Yang, Toward fast and accurate neural discourse segmentation, in *Proc. 2018 Conf. Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018, pp. 962–967.
[40] Y. Liu, C. Zhu, and M. Zeng, End-to-end segmentation-based news summarization, in *Proc. Findings of the Association for Computational Linguistics: ACL 2022*, Dublin, Ireland, 2022, pp. 544–554.
[41] C. Y. Lin, ROUGE: A package for automatic evaluation of summaries, presented at Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain, 2004.
[42] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, BERTScore: Evaluating text generation with BERT, arXiv preprint arXiv: 1904.09675, 2019.
[43] R. Mihalcea and P. Tarau, TextRank: Bringing order into text, in *Proc. 2004 Conf. Empirical Methods in Natural Language Processing* (*EMNLP*), Barcelona, Spain, 2004, pp. 404–411.
[44] P. Li, W. Lam, L. Bing, W. Guo, and H. Li, Cascaded attention based unsupervised information distillation for compressive summarization, in *Proc. 2017 Conf. Empirical Methods in Natural Language Processing* (*EMNLP*), Copenhagen, Denmark, 2017, pp. 2081–2090.
[45] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization, arXiv preprint arXiv: 1912.08777, 2019.

Guanghua Wang received the BEng degree in information security from Northeastern University, Shenyang, China in 2010, and the MEng degree in software engineering from Peking University, Beijing, China in 2014. He is currently pursuing the PhD degree in natural language processing at the Department of Computer Science, The University of Texas at Dallas, Richardson, TX, USA. His research interests include natural language processing, text summarization, and language models.

Priyanshi Garg is currently pursuing the PhD degree at the Department of Computer Science, The University of Texas at Dallas (UTD), Richardson, TX, USA. She received the BTech degree in information technology from SRM Institute of Technology, Kattankulathur, India in 2019, and the MSc degree in computer science in data science from UTD, USA. Her research interests are mainly in developing methodologies for social network analytics, machine learning and optimization, natural language processing, graph neural networks, and data science.

Weili Wu received the MS and PhD degrees from University of Minnesota, Minneapolis, MN, USA in 1998 and 2002, respectively. She is currently a full professor at the Department of Computer Science, The University of Texas at Dallas, Richardson, TX, USA. She is a senior member of IEEE. Her current research interests include the design and analysis of algorithms for optimization problems that occur in wireless networking environments and various database systems, especially data communication and data management.